Measuring the Impact of Profile Signals on Online Platform Integrity and User Safety

Alejandro E. D. Cuevas V.

CMU-S3D-25-114
August 2025

Software and Societal Systems Program
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Nicolas Christin, Chair
Bogdan Vasilescu
Sauvik Das
(Rolf van Wegberg), TU Delft
(Stefan Savage), University of California San Diego

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy of Societal Computing.

This research was partially supported by the US Dept. of Homeland Security (Office of Science and Technology) and the US Air Force Research Laboratory (AFRL) under agreement number FA8750-20-1-1003; the Singapore Defence Science and Technology Agency (DSTA) under agreement CNZ2000832; CyLab's Secure Blockchain Initiative; and, a CyLab Presidential Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US Department of Homeland Security, the US Air Force Research Laboratory, the Singapore Defence Science and Technology Agency, or CyLab.

Keywords: Computer Security, Empirical Measurements, Fraud and Abuse, Cybercrime, Reputation Systems

For my mom, Maria Luisa Villalba, who deserves to have her name next to mine in all things.
Mis logros son tuyos.

Abstract

This thesis examines how profile signals—such as subscriber counts, feedback ratings, and security verification badges—can be manipulated, misinterpreted, or rendered ineffective across online platforms, ultimately undermining platform integrity and user safety.

First, we show how misleading profile signals can increase user exposure to harm. Analyzing YouTube accounts sold on an online marketplace, we find that many are repurposed: channels that built reputations under one identity are sold, rebranded, and used to disseminate content that violates community guidelines, including scams and misinformation. Despite these changes, they retain engagement metrics and visibility, misleading users and amplifying harmful content.

Second, we test which profile signals predict sales cessation and market exit—both indicators of vendor quality—on darkweb marketplaces. We first validate a widely used revenue estimation method using ground truth data from a law enforcement operation. We then apply this method to model vendor quality across eight marketplaces over a decade. While reputation ratings are assumed to be strong indicators of quality, they are outperformed by models that incorporate a broader set of profile signals, highlighting an opportunity to improve vendor assessment.

Third, we extend this modeling approach to two cryptocurrency peer-to-peer marketplaces to predict account suspension due to fraud or abuse. Using longitudinal data, we train models and conduct a prospective cohort study comparing three groups: the lowest-rated accounts, a random baseline, and accounts flagged by our model. We find that security signals promoted by the platform, such as user ratings and verification badges, are poor predictors of suspension. Instead, less visible profile signals, like trading volume and partner diversity, are significantly more predictive. This finding underscores the need to audit user interfaces, as the most prominent signals fail to reflect actual risk.

Collectively, these studies show that profile signals can both obscure and reveal user risk. This thesis offers recommendations for platform designers and operators, emphasizing the need for empirical evaluation and redesign of profile signals and the systems that rely on them. By continuously auditing these signals and adapting user interfaces, platforms can more effectively surface trustworthy users and mitigate abuse, leading to safer and more resilient online ecosystems.

Acknowledgments

Throughout this journey, I have been fortunate to receive support and encouragement from many people. Thank you, to each of you for your words, actions, and presence in my life. To my wife, Amelia, for enduring with me through the long nights, the stress, and the uncertainty. To my family, thank you for your love and support even in distance, through texts, calls, and visits. To my dear friend Asia, for your unwavering support and encouragement; thank you for always making space for celebration. To my friend Ed, for checking in and always asking me for any papers that came out. To my friends who are now in all parts of the world, for your patience and understanding when I was not present, and for the adventures when I was. To my Pittsburgh friends, for the many dinners, lunches, coffee breaks, board games, and hangouts; may we have one more Pittsburgh summer. To my advisor, Nicolas Christin, who always prioritized me finding what I truly wanted to do: "finishing a PhD is easy, finding a topic you are passionate about is hard." To my committee, Bogdan Vasilescu, Sauvik Das, Rolf van Wegberg, and Stefan Savage, for your guidance and support throughout this process. To my lab mates, past and present, for the many discussions, brainstorming sessions, and feedback on my work. To my lab at Penn State and particularly Peng Liu, for meeting with me every week even when I did not have anything to show, for your patience and understanding; I would probably not have pursued a PhD if it were not for you. To my Dutch friends, who welcomed me with open arms and never charged me a tikkie. A special thanks to Fieke, who hosted me and helped me work with difficult data. To Madeleine, who was (and continues to be) a tremendous mentor and showed me what advocating for someone looks like. To Hanan, who trusted me from the start and welcomed me to pico. To Manoel, who I got to work with and learn from in the last year of my PhD; I am looking forward to continuing our work together. To God, for giving me the perseverance to keep going, the strength to overcome challenges, and the wisdom to learn from my mistakes. To everyone else, who was part of this journey, I am grateful for your contribution to my life and my work. Thank you.

Contents

1 Introduction **1**
- 1.1 Scope, Definitions, and Goals of the Thesis . . . 4
- 1.2 Thesis Organization . . . 5

2 Theoretical Background **11**
- 2.1 Signaling Theory . . . 11
  - 2.1.1 Signaling in Online Marketplaces . . . 13
  - 2.1.2 Signaling in Social Media Platforms . . . 15
  - 2.1.3 Signaling in Other Online Platforms . . . 17
- 2.2 Challenges and Threats Involving Signals in Online Platforms . . . 18

3 Misleading Social Signals: A Study of Repurposed YouTube Channels **23**
- 3.1 Motivation and Goals . . . 23
- 3.2 Related Work . . . 25
- 3.3 Background . . . 28
  - 3.3.1 The Fameswap Market . . . 28
  - 3.3.2 YouTube Handles and IDs . . . 30
- 3.4 Methods . . . 30
  - 3.4.1 Data Collection and Sources . . . 31
- 3.5 Repurposed Channels . . . 32
  - 3.5.1 Defining Repurposed Channels . . . 33
  - 3.5.2 Repurposed Channels in the Wild . . . 35
  - 3.5.3 Prevalence Estimation . . . 36
  - 3.5.4 Repurposed Channels Before Observation . . . 36
- 3.6 Content Analysis . . . 37
  - 3.6.1 Codebook Development and Annotation . . . 37
  - 3.6.2 Presence of Problematic Content . . . 40
- 3.7 Indicators of Channel Repurposing and Prevalence . . . 43
  - 3.7.1 Regression Model . . . 43
  - 3.7.2 Regression Results . . . 45
- 3.8 Discussion and Conclusions . . . 46
  - 3.8.1 Limitations and Future Work . . . 49
- 3.9 Ethics . . . 49

4 Evaluating the Reliability of Online Anonymous Market Measurements 53
4.1 Motivation and Goals . . . 53
4.2 Measuring Marketplaces . . . 56
4.2.1 Background . . . 57
4.2.2 Literature survey . . . 58
4.3 Methodology . . . 60
4.4 Modeling Marketplaces . . . 61
4.4.1 Model Components . . . 61
4.4.2 Data Collection . . . 62
4.4.3 Data Analysis . . . 63
4.4.4 Losses . . . 63
4.5 Datasets . . . 64
4.5.1 Public View – Hansa Scrapes . . . 64
4.5.2 Admin View – Hansa Database . . . 66
4.5.3 Public View – Simulation . . . 68
4.6 Coverage and Bias . . . 70
4.6.1 Scraping Coverage . . . 71
4.6.2 Scraping Bias . . . 72
4.7 Revenue Calculations . . . 74
4.8 Simulation . . . 76
4.8.1 Coverage of One and Two-shot Scrapes . . . 76
4.8.2 Coverage and Scraping Consistency . . . 77
4.8.3 Comparison of Abundance Estimators . . . 78
4.8.4 Popularity-Driven Scraping . . . 79
4.8.5 Extrapolation . . . 80
4.9 Discussion . . . 81
4.10 Ethics . . . 83

5 Evaluating Reputation Systems in Online Anonymous Marketplaces 85
5.1 Motivation and Goals . . . 85
5.2 Background and Related Work . . . 88
5.2.1 Reputation & Feedback Systems . . . 88
5.2.2 Performance in Criminal Markets . . . 89
5.3 Methodology . . . 90
5.3.1 Data Sources . . . 90
5.3.2 Data Processing and Validation . . . 91
5.3.3 Extracting Features . . . 92
5.3.4 Ethics of Data Collection and Release . . . 93
5.4 Survivability Drivers . . . 96
5.4.1 Experimental Setup . . . 97
5.4.2 Results . . . 98
5.4.3 Reputation Slander Attack . . . 99
5.5 Predicting Success and Longevity . . . 100

5.5.1 Predicting Vendor Disappearance . . . 103
5.6 Generalizability and Feature Importance . . . 106
5.6.1 Experiment Setup . . . 106
5.6.2 Results . . . 107
5.6.3 Explaining and Improving Performance . . . 108
5.7 Discussion . . . 109
5.7.1 Interventions and Policy Takeaways . . . 110
5.7.2 Limitations and Future Work . . . 111

6 Identifying Risky Vendors with Public Signals 113
6.1 Motivation and Goals . . . 113
6.2 Background . . . 116
6.2.1 P2P Cryptocurrency Marketplaces . . . 116
6.2.2 Reputation Systems: Benefits and Challenges . . . 118
6.3 Data . . . 119
6.3.1 Collection . . . 119
6.3.2 User Suspension . . . 121
6.3.3 Ethics and Legality . . . 121
6.4 Evaluation on Existing Reputation System . . . 122
6.4.1 Feedback Signals . . . 122
6.4.2 User Collusion and Automation . . . 124
6.5 Predicting Account Suspension on Paxful . . . 125
6.5.1 User Features . . . 126
6.5.2 Machine Learning Models . . . 126
6.5.3 Evasive Measures . . . 128
6.6 Generalizing the Model Across Markets . . . 129
6.7 Prospective Cohort Study . . . 130
6.7.1 Experimental Setting . . . 131
6.7.2 Results and Implications . . . 131
6.7.3 Robustness . . . 133
6.8 Related Work . . . 134
6.9 Design Insights . . . 135
6.10 Takeaways . . . 136

7 Conclusion 139

A Chapter 3 Prompts and Qualitative Coding Materials 141
A.1 LLM Prompts . . . 141
A.2 Qualitative Coding Guide and Template . . . 141

B Chapter 4 Data Schemas and Extended Analyses 147
B.1 Object features in the Hansa database . . . 147
B.1.1 Bias analysis . . . 147
B.2 Exploratory Factor Analysis . . . 147

B.3 Abundance Estimation Algorithms 151
B.4 Scrape Dates 151

C Chapter 6 Complementary Analyses **153**
C.1 Geographical Considerations 153
C.2 Identifying Suspension 154
C.3 Suspension label validation (Paxful) 156
C.4 Platform Moderation Evaluation 156
C.5 Feedback Keyword Searches 157
C.6 Complementary Evidence of Self-promotion 158

Bibliography **161**

List of Figures

1.1 Examples of profile signals across popular online platforms. Profiles typically include traditional reputation signals, such as ratings and reviews, as well as social signals, such as the number of followers, and badges for various types of achievements for quality contributions and behavior. 4

3.1 On the left, a channel listed for sale on Fameswap about entertaining facts with 1.19M subscribers. Videos are created predominantly with AI tools. This channel would later change to a news channel discussing political issues with no trace of its previous identity (on the right). For privacy, the channels' handles and titles are blurred. 27

3.2 Data collection procedures. We collected Fameswap listings (1a) daily and a large sample of Social Blade channels (1b). We then scrape snapshots through time. On Fameswap, we scraped a channel every 3 days on average (2a). For Social Blade channels, due to the sample size, we collected a snapshot in January and later in March 2025 (2b). Finally, we obtain historical snapshots using the Wayback Machine (WM; see 3a and 3b). 28

3.3 Distribution of listing prices, subscriber counts, and price per 1,000 subscribers for Fameswap listings. Outliers (Q3 + 1.5X IQR) are hidden. Categories are self-reported at the time of listing. Categories are ordered by median price per 1,000 subscribers. 29

3.4 Subscriber growth comparison between repurposed (treatment) and non-repurposed channels (control). X-axis is relative to repurposing event. For non-repurposed accounts, the X-axis is based on the time they were first listed on Fameswap. Percentage growth is relative to the number of subscribers at t=0. 33

3.5 Time (days), since a channel is listed for sale, until it changes its handle, title, and description (cumulative). 34

3.6 Topic correlation matrix. Values represent the Pearson correlation between variables. 37

3.7 Topic presence for reused Fameswap channels. Topics on the left are self-reported at the time of listing. Topics on the right are labels assigned during our topic detection. Flows are proportional and one-to-many. That is, if a channel has several topics, flows are drawn to all destination topics proportionally. 39
3.8 Survivability curve for Fameswap channels with both potentially problematic content detected and without. "Death" event is channel removal. Bands represent 95% CI. 41
3.9 End-to-end experimental procedure, including samples, annotations, featurization, and regression. 43

4.1 Landing page of the Hansa marketplace. 55
4.2 Item page of the Alphabay marketplace. Hansa, like many other dark web marketplaces, features a similar layout and design. Given the lack of Hansa screenshots, we use Alphabay as an example. 56
4.3 Review page of the Alphabay marketplace, which we use as an example given the lack of Hansa screenshots, as described in Figure 4.2. Revenue can be estimated for each item, based on data present on the item and reviews pages. 57
4.4 Weekly counts of objects from the Hansa back-end 68
4.5 Steps involved in the generation and scraping of a simulated online marketplace. 69
4.6 Per-scrape instantaneous and cumulative coverage. 70
4.7 Calculation of Hansa's March 2017 revenue with different inputs . . . 75
4.8 Distribution of coverage for one and two-shot scrapes simulated across different request limits. 76
4.9 Scraping coverage as the number of scrapes increases, with evenly spaced scrapes and randomly spaced scrapes. The shaded area is the 95% confidence interval. 77
4.10 Average scrape coverage through a simulated market's lifetime for various popularity scraping budgets and compared to our uniformly random scraping baseline. 80

5.1 Profile page of a vendor in the Nemesis marketplace. 90
5.2 Revenue over time for all markets scaled to the same time axis. Each point is a four-week rolling window average. 92
5.3 Accuracy of the RF model using base features in predicting the end-of-market revenue quantile a vendor belongs to, across markets. Labels are balanced. Timesteps are scaled to market lifetime percentage for visualization. Decrease in accuracy is due to new vendor entrancy. "Extended" includes subreddit/forum features. 101

5.4 Accuracy of the TSF model using temporal features in predicting the end-of-market revenue quantile a vendor belongs to, across markets. Labels are balanced. Timesteps are scaled to market lifetime percentage for visualization. Decrease in accuracy is due to new vendor entrancy. "Extended" includes subreddit/forum features. 102
5.5 Average survival time in weeks for each group, across all market timesteps. Vendors who stop having sales after a given week are removed from the sample. The high/mid-risk groups across markets were assembled with the same classifier, which was trained with samples from all markets. 104
5.6 Comparison of average prediction accuracy for the revenue quantile that a vendor belongs to by the end of the market. Accuracy is averaged across market stages. We train on $n-1$ markets and test on the holdout. This plot compares training performance on the $n-1$ markets with test performance on the held-out market. 105
5.7 Comparison of test scores across categories for predicting vendor revenue quantile by market end. Category segmentation includes only vendors whose primary goods fall into the given category. The baseline line shows performance without category segmentation. 105

6.1 Example user profile on Paxful. The profile shows the user's reputation score, number of trades, and various forms of verification, among other signals. 117
6.2 Example user profile on LocalCoinSwap. The profile shows the user's reputation score, number of trades, and various forms of verification, among other signals. 118
6.3 Transaction flow in cryptocurrency P2P markets. 119
6.4 Data collection through Paxful APIs 120
6.5 Feedback interval (days) boxplots. We restrict the y-axis ranges for better visualization. 125
6.6 Suspensions on Paxful over a one-month period for three user groups (each $n = 500$): those flagged by our classifier as most likely to be suspended, those with the lowest reputation scores, and a random sample. The classifier-based group shows substantially higher suspension counts. 132
6.7 Suspension rates on Paxful when varying group size n. The figure compares the three user groups described in Figure 6.6, demonstrating that classifier-selected users consistently show higher suspension rates across group sizes. 133

A.1 Prompt used for channel repurposing annotation. 143
A.2 Topic annotation prompt. 144
A.3 Few-shot examples for topic annotation prompt. 145

A.4 Markdown document generated for each channel and provided to coders during qualitative analyses. The document provides the handle changes, a channel timeline which includes snapshots of profile information over time, and all videos observed for the channel. The channel above has since been deleted. 146

C.1 Number of users for the top 20 countries in Paxful and currencies in LCS. y-axis is in log scale. 154
C.2 Heatmap of country changes between registration and subsequent accesses (normalized by x-axis). 155
C.3 Ratio of users for which the reported country of access is different at least once from the registration country (for countries with more than 500 users). 155
C.4 Account (un)suspensions per day on Paxful, 01/08/23–03/31/23. . . . 156
C.5 CDFs of the number of days to suspension and to release. 157
C.6 Top 10 payment methods + 3 selected payments. N: non-suspended accounts, S: suspended accounts (e.g., N2S: feedback from non-suspended to suspended accounts) 158

List of Tables

3.1 Descriptive statistics for the regression features used in the Fameswap and Socialblade models. 45
3.2 Regression results for Fameswap and Socialblade models. 47

4.1 Comparisons between Hansa studies. We include counts of reviews without price information. However, we omit them when estimating revenue. 65
4.2 Marketplace objects from Hansa back-end 67
4.3 Results of the Mann-Whitney U and χ^2 tests between scraped and not-scraped listings 72
4.4 Avg. error when estimating the number of listings across scraping intervals and using either a low request limit (2 req./min) or a high request limit (20 req./min). 79

5.1 Overview of collected and processed marketplace data. 91
5.2 Cox Proportional Hazards Regression across all 8 markets, where exp(c) indicates the hazard rate increase per unit increment. The regression was stratified based on the wealth quartile the vendors belonged to at the end of the market. 96
5.3 Cox Proportional Hazards Regression on forum features extracted for Hansa and Nemesis. 99
5.4 Average classification metrics across our 4 labels (wealth tiers). The holdout is the market on which we predict while training on the others. Labels are balanced across classes. Each metric is the average score obtained across our 10 experiments. 108
5.5 Ablation study of our revenue prediction model on holdout markets. We exclude combinations of features and quantify the accuracy decrease on the model. 108

6.1 Prediction results: seven models with CI (2.5%, 97.5%). 127
6.2 Top 10 most important features for Paxful (§6.5) and LCS (§6.6) categorized by data source. Number in parentheses is the feature importance rank. 128

6.3 Performance results for two markets (Susp. = num. of suspended accounts, Acc. = Accuracy). 130
6.4 Statistical test results comparing group pairs. Entries show test statistic and p-value. 134

A.1 Coding guide for channel categorization. 142

B.1 Listing features 148
B.2 Review features 148
B.3 User features 149
B.4 Order features 149
B.5 Transaction features 149
B.6 Results of the Mann-Whitney U and χ^2 tests between scraped and not scraped reviews 150
B.7 Results of the χ^2 test between scraped and not scraped vendors 150
B.8 Results of the KMO and Bartlett's tests 150
B.9 Factor Analysis Output 150

Chapter 1

Introduction

Trust is an essential component in online platforms. Every day, millions of people decide to trust each other online when buying products, contracting services, receiving news, and even heeding advice on sensitive topics such as personal health and finances. To facilitate trust-building between users and encourage high-quality contributions and behavior, platform operators design systems that aggregate and display metrics on each user's activity, social capital, and ratings from other users in user profiles. These metrics, often composed of reputation or social signals, are designed to incentivize suppliers—of information, products, or services—to behave as good members of the platform, adhering to rules and contributing positively. In turn, these signals guide consumers to make the decisions about who to buy from or who to consume content from. In e-commerce platforms, such as the eBay marketplace, the number of positive reviews a supplier has may signal their trustworthiness, enticing buyers to purchase products from them. In social media platforms, such as Twitter/X, a verification badge may signal that an account is authentic; others may perceive this signal as an indicator that they can trust the content they are consuming. These signals should be accurate and easy to interpret. Because these signals should translate to more sales or followers, they should be attractive to suppliers to acquire, and they should be hard to get. As a result, these signals should be a reliable indicator of the quality of the user, and platforms can use them to encourage positive behavior, dissuade rule-breakers, and create a safer online environment for users.

Among the oldest and most widely adopted systems for building trust are reputation systems, which have been touted as crucial in enabling online marketplaces [1], such as e-commerce and labor marketplaces. It is no surprise that reputation systems are prevalent across platforms like Amazon, AirBnb, Uber, and Fiverr where goods and services are transacted between two parties. Reputation systems, which encompass signals such as reviews and scores assigned by other users, are prominently displayed in user profiles. In well-designed reputation systems, the accumulation of

positive reputation signals confers many advantages to vendors, such as the ability to charge higher prices [2, 3]. Accumulating these positive signals incentivizes vendors to provide better service: write accurate product descriptions, ship products in a timely manner, and provide better customer support. [1]. That is, reputation systems encourage "high-quality behavior." At the same time, reputation systems serve to penalize individuals who engage in "low-quality behavior," such as defrauding buyers, by allowing buyers to leave negative reviews and ratings [4]. These negative reviews ultimately push low-quality vendors out of the market [4]. From an economic perspective, reputation systems increase market efficiency [1]. Through a computer security lens, reputation systems are a tool to improve the security of online marketplaces by increasing adherence to community norms and terms of service, as well as deterring fraud.

Social media platforms and online communities have also adopted profile signals to build trust, create healthy communities, and dissuade misbehaving or malicious users [5]. One set of signals found in social media profiles are similar to those found in marketplace reputation systems, such as ratings and feedback. Profile signals may also include social signals, such as the number of followers, likes, and shares. In knowledge-sharing platforms, such as Stack Overflow, Wikipedia, and Reddit—to mention a few—reputation systems are used to reward users for high-quality contributions, such as writing informative answers, editing articles, and creating engaging posts [5]. In turn, these positive signals provide users with tangible recognition of their contributions, better treatment by other users, more privileges on the platform (e.g., the ability to start threads or vote on polls), among other benefits [5]. Profile social signals also play an important role in fostering connection between people online. In social media platforms, profile signals affect who users choose to form a friendship with [6, 7] or even a romantic relationship with [8]. Not only do these signals help find like-minded individuals, they can also help users filter potential accounts that may be looking to harm them through harassment and scams [9]. Similarly to signals in online markets, profile signals in social media platforms help online communities self-moderate: encouraging positive contributions, discouraging negative actions, and helping users screen other accounts that may be looking to harm them.

Profile signals are not only consumed by end-users in a platform, but they also feed into other systems, such as recommendation and moderation systems. Negatively reviewed users may be deprioritized from search results, while positively rated users may be amplified. Profile signals also influence moderation systems, whether automated or human-led. An automated moderation system may take as input the number of negative reviews a user has received to issue a penalty [10]. Or, a human moderator may consider the signals in a user's profile (e.g., reputation, account age, etc.) before deciding to take action against them [11, 12]. Thus, not only can inaccu-

rate signals mislead other humans, such as users and moderators, but their negative impact can be exacerbated by algorithms within the platform. These signals may mislead users into interacting with a seller who defrauds customers [13], mislead moderators into being more lenient with a user who regularly violates community norms—by applying lesser sanctions or foregoing penalties [11, 14]—, or even amplify users who promote scams and risky investments [15, 16]. The design of accurate signals is therefore crucial to the security of various systems within online platforms. As such, finding empirical evidence of misleading signals can inform the design of more reliable profile signals, leading to more resilient and trustworthy systems throughout the platform and ultimately better and safer user experiences.

Academic work on the design and presentation of online profile signals has traditionally involved thinking from economics, sociology, and human-computer interaction. From an economic perspective, the literature has focused mainly on the relationship between reputation, price, product quality, and support of future transactions [2, 4, 17–22]. From a sociology and human-computer interaction perspective, the literature has focused on the role of profile signals to build trust, foster personal connections, create healthy communities, encourage participation, incentivize high-quality contributions and related behaviors [5–8, 23–26].

Because profile signals often carry tangible benefits, dishonest users may attempt to manipulate these signals to gain an unfair advantage. Computer security work has primarily focused on how malicious users may manipulate profile signals by exploiting weaknesses in the design of reputation systems, creating many fake accounts, or engaging in collusion to artificially inflate their reputation[13, 27–29]. However, two areas remain understudied. First, users may leverage profile signals as a stepping stone to carry out other types of harmful activities. Yet, most work focuses on the manipulation itself: its mechanisms, how to detect it, and how to prevent it. Second, while economists and sociologists studying online communities have noted that profile signals can reduce undesirable behavior across communities, there has been less work that evaluates the effectiveness of profile signals in achieving computer security goals. That is, we lack empirical evidence of how profile signals can be used to deter misbehavior, such as fraud, toxicity, harassment, and the spread of disinformation.

This thesis fills these gaps by empirically evaluating the effectiveness of online profile signals to indicate user quality and investigating their potential shortcomings. To accomplish this, we focus on three aspects. First, we explore how the transfer of reputation signals via account sales can introduce adverse outcomes to users on online platforms. Second, we develop and validate techniques to model the impact of profile signals on financial outcomes and account suspension. Third, we use these techniques to demonstrate how they can be used to build tools that better represent the quality of users in online platforms and inform the design of more informative

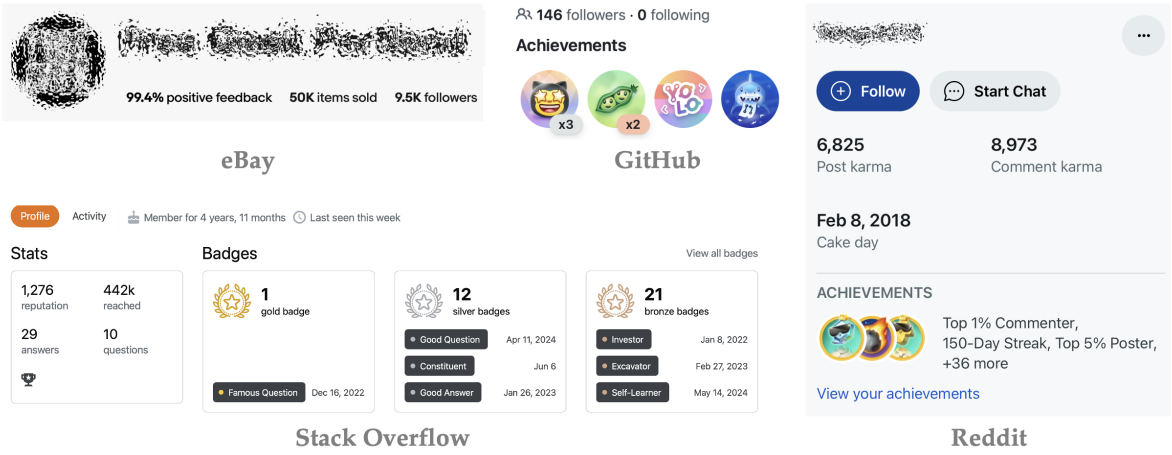


Figure 1.1: Examples of profile signals across popular online platforms. Profiles typically include traditional reputation signals, such as ratings and reviews, as well as social signals, such as the number of followers, and badges for various types of achievements for quality contributions and behavior.

user interfaces.

1.1 Scope, Definitions, and Goals of the Thesis

We consider "profile signals" as all account-level signals attributed to a user on an online platform. By "account-level", we mean that the signals are associated with an individual user account, aggregated on a user profile page, and visible in the user interface. Apart from traditional reputation signals, such as those generated by reputation systems, we also consider user-level metrics (e.g., number of posts, number of trades, etc.) and social signals (e.g., number of followers, number of likes, etc.). We present many of the signals we encounter in our daily lives in Figure 1.1.

Our focus is primarily on platforms where there are suppliers and consumers. These are platforms where there is an exchange of goods, services, or information, such as online e-commerce marketplaces and media-sharing platforms (a subset of social media platforms). Even though all users in these platforms typically have accounts and user-level signals, we focus on suppliers. In the context of e-commerce marketplaces, these users are typically sellers (also known as vendors). In media-sharing platforms, such as YouTube, these users are typically content creators. We focus on the following overarching questions:

- **RQ1**: What are potential risks that can be introduced by profile-signal manipulation on online platforms?

- **RQ2**: Do profile signals meaningfully represent user quality, as evidenced by out-of-sample prediction of sales cessation or account suspension?
- **RQ3**: Can statistical modeling audit which profile signals deserve user interface prominence by showing out-of-sample predictive validity for future guideline violations?

This dissertation posits the following thesis:

> *"The presence or absence of profile signals in user interfaces may mislead users, increasing users' exposure to harm. Using statistical modeling, we can identify the signals most correlated to various (stated) platform goals and evaluate how well the signals the platform provides align with their objectives. By continuously evaluating online platforms' signals, we can inform the design of reputation systems and interfaces that align their stated purpose with their practical use."*

1.2 Thesis Organization

This thesis is organized into three main parts. In the first part, corresponding to Chapter 3, we explore a new form of reputation manipulation in the content creation and social media platform YouTube. To do this, we collect accounts that have been advertised for sale on an online marketplace named Fameswap. In this marketplace, users can buy and sell social media accounts on platforms such as YouTube, Twitter/X, Instagram, and TikTok. In many cases, we observe that the accounts that are sold are then *repurposed*. That is, they alter their content and identity while preserving the accumulated signals (i.e., subscribers, likes, views) and existing audience. These accounts leverage the fact that YouTube provides a handle in the form of "@handle" that resolves to a given account even if the handle is changed. In the context of domain names, redirecting a URL to a different service is known as abusing the "residual trust" of the domain [30, 31]. We find that channels are routinely repurposed with millions of subscribers affected. A viewer who may have originally subscribed to a channel about cooking may later be surreptitiously served content about investment scams, high-risk cryptocurrency trading, and misinformation. Existing signals in the YouTube interface offer little indication of this change, and the repurposed channels continue to receive views, likes, and comments from the original audience. This chapter highlights two issues. First, it shows that signals in online platforms can be misleading. That is, an account that had built its reputation providing content about a given subject matter (e.g., cooking) was able to carry over that reputation to a new identity (e.g., cryptocurrency investments)—and potentially a new owner. A newcomer may then falsely believe that the channels has always been about the new subject and accumulated its metrics and signals with that content. Second, this

study shows that insufficient information in the user interface can be leveraged by miscreant users to spread content that violates community norms and may put users at risk of encountering potentially harmful content. This work is currently under submission.

In the second part, corresponding to Chapters 4 and 5, we explore the role of reputation signals in predicting financial outcomes in online marketplaces. Our goal with this estimation task is to model the relationship between reputation signals and financial success or failure. A key question we are interested in is whether reputation signals help push low-quality vendors out of the marketplace by explaining and predicting the amount of sales they receive. To do this, we focus on online anonymous marketplaces (also known as "darknet marketplaces"). These marketplaces are online platforms that facilitate the exchange of goods and services, primarily illicit, while providing users with various forms of anonymity. Online anonymous marketplaces are particularly good platforms to study reputation because they lack identity verification and thus legal consequences for sellers defrauding buyers are non-existent. Furthermore, sellers in online marketplaces have little to no reputation outside the platform that we need to control for. That is, they do not have large social media presences, celebrity endorsements, marketing campaigns, or other resources that may influence the amount of money they make. Lastly, reputation in these platforms is of utmost importance, given that there are dangerous substances being sold for which a lack of accurate information can lead to severe health consequences, and even death [32]. Because of these reasons, the signals displayed in these marketplaces' user interfaces are often all that a buyer has to determine whether a vendor is trustworthy or not.

To model the relationship between vendors' signals and financial outcomes we need a reliable way to estimate the financial success and failure of vendors. To estimate the revenue (or lack thereof) of a vendor in an online marketplace, scholars typically use reviews as a proxy for sales and the listing price as the final price of the transaction [33–36]. This estimation method, while widely used, has not been validated. There are many factors that can affect the accuracy of this estimation, such as the scraping coverage, lack of reviews for orders, or variations in the final price of the transaction due to quantities or discounts. In Chapter 4, we validate the accuracy of this estimation method by comparing the estimated revenues for a marketplace named Hansa with ground truth data obtained by law enforcement. We identify and quantify the impact of various sources which affect the accuracy of the estimation method, such as the frequency of scraping. We also use the ground truth data to design a simulation using a bootstrapped sampling method. Lastly, we test a set of population estimation techniques to correct the estimates. This study provides a toolkit for law enforcement agencies and researchers to conduct measurements on online

anonymous marketplaces. Most importantly, in the context of this thesis, this study is a stepping stone towards reliably modeling the financial outcomes of vendors in online anonymous marketplaces. This work was published in the Proceedings of the USENIX Security Symposium 2022 [37].

In Chapter 5, we use the validated revenue estimation method to model the relationship between reputation signals on financial success and failure. The work, which was published in the Proceedings of the 2024 USENIX Security Symposium [38], uses data from eight online anonymous marketplaces spanning more than a decade, we train predictive models that attempt to predict and explain the financial outcomes of vendors across each of these marketplaces. We do this using a random forest classifier. We create panel data for each vendor in the marketplace that contain the signals that they had available for display at each week (our chosen panel timestep). Using these data, we train a model at each week of the market and predict the financial outcome of the vendor at the end of the market (i.e., when the market shut down or was seized by law enforcement). In this classification task, we define the financial outcome as the revenue quartile to which a vendor belongs, ranging from the lowest quartile (bottom 25%) to the highest quartile (top 25%). We find not only that we can train a high accuracy model for each marketplace but also that we can train models that generalize across marketplaces. In addition, we also train a generalizable model that can better identify which vendors are most likely to stop selling. When exploring the features that drive the models' predictions, we find that reputation signals alone are not the main drivers of financial outcomes. In fact, when we observe the longevity of vendors with low reputation scores and compare it to a random baseline, we do not observe significant differences. However, when we train a machine learning model that employs a variety of features or signals, we can identify the vendors who are most likely to leave the market. In this context, this procedure evidences that reputation signals alone, which are among the most prominent signals in the user interfaces of online anonymous marketplaces, are not sufficient.

However, a vendor may leave the market for a variety of reasons. For example, they stopped having sales because buyers stopped trusting them or because market administrators blocked their account due to fraud. In these cases, the signals on the platform are in fact achieving the goals of deterring low-quality vendors and pushing them off the platform. However, there are many other reasons why a vendor may leave the market, such as being arrested by law enforcement or other personal circumstances. The results of this chapter do not distinguish between the specific reasons a vendor leaves the marketplace. In particular, we are interested in evaluating whether these signals are informative in predicting fraud. This limitation motivates the next chapter, where we explore the relationship between profile signals and account suspension due to misbehavior.

In the third part, Chapter 6, we apply the techniques we developed in online anonymous marketplaces and apply them to cryptocurrency peer-to-peer marketplaces. In this case, we focus on the Paxful and LocalCoinSwap marketplaces. Unlike our work in the previous chapter, we focus on the relationship between reputation signals and account suspension. Accounts are suspended for serious violations, such as fraud, platform abuse, and participation in illegal activities. When we query an account that has been suspended, the platform returns data indicating the suspension. In both of these markets, reputation is important given that users are transacting in a peer-to-peer setting and the cryptocurrency ecosystem is notorious for attracting fraud. To help users identify trustworthy counterparties, these markets provide a variety of signals in user profiles, ranging from feedback ratings, to the amount of money traded, and various forms of verification done by the platform, such as address, phone, and identity verification. Using longitudinal data from these markets, we train a classifier to predict which active accounts are most likely to be suspended in the future. We assemble three groups: the accounts our model predicts will be suspended, the lowest-rated accounts, and a random sample of accounts. We then follow these accounts for one month and count how many of them were suspended. Similar to Chapter 5, we find that the lowest rated accounts are not suspended at a significantly higher rate than the random sample. However, our model is able to identify accounts that are suspended at a significantly higher rate than the lowest rated accounts and the random sample. Notably, the features that drive the model's predictions are not the features that are most prominent in the user interface. For example, the user rating is not a significant predictor of suspension, nor the different types of verification, both of which are prominently displayed in the user interface. Instead, the model relies on features such as the amount of trades and trade partners. This chapter demonstrates that computational modeling can not only be used to identify risky vendors but also that we can use computational modeling to inform the design of user interfaces. This work was published at the 2024 Proceedings of the ACM Web Conference [39].

These findings motivate the use of statistical modeling to evaluate existing profile signals with their intended outcomes. Through this approach, we develop recommendations and tools to help platform operators, who administrate recommendation and moderation systems, identify potential risks introduced by manipulated profile signals. We also provide an empirical approach to audit user interfaces, and subsequently, inform the development of more informative profile signals. Through our approach, administrators can evaluate the effectiveness of adopting and displaying security verifications, such as badges and checkmarks. Furthermore, acknowledging that malicious users continuously evolve their tactics, this thesis advocates for continuous evaluations. By continuously evaluating profile signals, we can identify the

signals that are most correlated with various outcomes, such as predicting various forms of misbehavior. Continuous evaluation also allows us to identify when signals deteriorate in their informativeness, allowing platforms to adapt accordingly. Together, these findings provide a template to evaluate and develop profile signals that are informative, accurate, and useful for users, platform operators, and moderators, resulting in safer and more trustworthy online platforms.

Chapter 2

Theoretical Background

We provide an overview of signaling theory and its application on three types of online platforms: online marketplaces, social media platforms, and knowledge-sharing platforms. This is a non-exhaustive yet illustrative overview of the platforms within the scope of this thesis. Each of these platforms differs in their primary objectives, but they share similar goals: fostering a high-quality user base and pushing off misbehaving users. Furthermore, the design of user profiles across these platforms shares many similarities. For example, reputation systems, while originally designed for online marketplaces, have been adopted across social platforms. Social signals, such as followers, have been adopted in marketplaces, such as eBay. Lastly, each of these platforms shares similar threats and challenges, such as the manipulation of signals or the abuse of signals to carry out other types of bad activities, such as fraud.

2.1 Signaling Theory

Signaling theory is a well-established framework for understanding the relationship between signals and the underlying qualities of a subject. This framework establishes desirable properties of signals and helps explain how consumers interact with and choose the signals that. Although originally developed in the context of economics [40] and biology [41], signaling theory has been extended to online spaces and online profile signals. In this section, we provide an overview of signaling theory as a way to establish the mechanisms that explain how and why signals influence behavior, and later why dishonest users may seek to manipulate online signals, and how we can leverage these signals to the security of online platforms.

Signaling theory helps us understand how signals are used to convey information about an agent's quality and how the costs of production and consumption of signals affect their effectiveness. *Assessment signals* are signals that require the possession of a certain quality to produce the signal. For example, a FIDE chess rating is an

assessment signal, as it requires the player to have a certain level of skill to obtain a rating [41]. A subclass of assessment signals are *handicap assessment signals*, where the cost of the signal is higher for low quality agents than for high quality agents. For example, spending a large amount of money in a casino is a handicap assessment signal of wealth. That is, it would be difficult for a low-wealth individual to spend a large amount of money at a casino, whereas a high-wealth individual would not have a problem doing so. Lastly, *conventional signals* are signals where the link between quality and signal is not direct, but is established by social convention [42, 43].

Conventional signals abound on online platforms. However, keeping signals honest, especially in the face of attractive economic incentives (as we will see below), is a challenging problem with direct implications to the health and safety of online communities [44]. As Donath notes, "it is just as easy to type 24 or 62 as it is to enter one's age" [7]. Conventional signals are kept honest through social conventions and laws, where a community agrees on the meaning of a signal and the consequences of misrepresenting it. Lying about one's military service can, at the very least, lead to social ostracism, and at the very worst, to legal consequences in some countries, such as the United States [45]. In online spaces, platform administrators often establish penalties for misrepresenting online signals [46], and they spend substantial resources enforcing these policies [47]. Platforms also attempt to design interfaces and systems that allow users to ostracize dishonest users, such as product and vendor reviews in online marketplaces. Contrary to the physical world, online identities are often easily replaceable, and thus, dishonest users continuously invest in developing new techniques to bypass penalties.

As a result, another approach taken by platform administrators has been to make online identities and signals harder to replace. Signaling theory suggests that signals are more effective when they are costly to produce [40]. However, creating costly signals on online platforms has also remained a challenge because online information is typically easy to forge. It is difficult to verify the authenticity of information online, such as whether an account is owned by a real person [48] or to prove certain characteristics about oneself, such as age [49]. This problem remains unsolved in decentralized settings, with recent efforts going as far as to use retina scans as a prerequisite to create cryptocurrency wallets [50]. However, introducing a central authority does not immediately solve the problem. Instead, a central authority's credibility must be established. For example, an ELO rating in the platform Chess.com is considered a decently credible signal of a player's skill, as the platform maintains the integrity of the rating system. Most people recognize that this rating is not as credible as a FIDE rating, given that players in Chess.com may create multiple accounts, use cheating software, or otherwise manipulate their rating. Ultimately, however, there are not many financial incentives to cheat in Chess.com, as the rating in the platform

has little value outside of the platform.

However, given sufficient incentives, very few signals are impossible to fake. As suggested by Goodhart's law, "when a measure becomes a target, it ceases to be a good measure." In other words, when there are clear incentives to fake profile signals, such as increased revenue or credibility, dishonest users will find ways to manipulate these signals, leading to *signal deterioration*. There is much computer security scholarship on how attackers manipulate signals on online platforms and how to mitigate these manipulations [13]. For example, in the context of online marketplaces, fake reviews are a common occurrence, where sellers create fake reviews to inflate their reputation and presumably attract more customers [51]. In the context of social media, fake accounts are often used to amplify content. To do this, a user creates multiple accounts to like and share their own content, artificially increasing its popularity [52, 53]. Below, we provide an overview of the advantages conferred by online profile signals across three types of platforms to then explore how and why dishonest users may seek to manipulate profile signals.

2.1.1 Signaling in Online Marketplaces

Signals in online marketplaces, commonly known as *reputation signals*, are signals that are used to convey information about the quality of a seller or product. These signals have been touted as crucial in the successful establishment of online marketplaces [1]. Reputation signals facilitate trade in online settings because they introduce incentives for vendors to behave honestly, whereas without them, a rational vendor's optimal strategy is to defraud the buyer. Reputation signals are often aggregated and displayed in user profiles, product pages, and search results, as quantitative signals (e.g., numeric scores) or qualitative signals (e.g., text reviews). The mechanisms by which these signals are collected, aggregated, and displayed are known as *reputation systems*. From an algorithmic perspective, reputation systems are collaborative filtering algorithms, in which ratings for a collection of agents are determined based on the ratings given by other agents [54].

The primary goals of reputation systems are to reduce information asymmetry and moral hazard [55]. In online marketplaces, there is an information asymmetry between buyers and sellers, where buyers do not have perfect information about the good or service they are purchasing. A seller who receives money for a product from someone thousands may simply decide to keep the money and the item, if there are no consequences. Another seller may intentionally ship a faulty product—even though they advertised it as new—or a product that does not match the description they wrote. Because the buyer cannot inspect the product in person, a rational buyer would have little reason to trust the seller. Markets with information asymmetry,

where sellers have little incentive to not defraud the user, are eventually headed to failure [56].

To address information asymmetry and moral hazard, reputation systems serve two functions: sanctioning and signaling. Reputation systems are designed to mitigate moral hazard issues by acting as *sanctioning devices*. If community members avoid transacting with misbehaving sellers, then rational sellers have an incentive to cooperate as long as future gains outweigh the short-term gains from cheating [55]. As a sanctioning device, reputation systems track the quality of the seller or their effort and provide incentives (or deterrents) that stir sellers' behavior. Reputation systems also act as *signaling devices* [55]. For example, when buying a book or movie from an online retailer, reviews help users learn more about the product and whether they will enjoy it or not. In most real-life settings, reputation systems fulfill both roles simultaneously. From a computer security perspective, we are particularly interested in the sanctioning role if reputation signals, as these signals should dis-incentivize malicious actors in the platform and allow users to make informed decisions about the quality of the seller or product.

A key feature of effective signals is that they are expensive to produce [40]. In labor markets, employers want to assess the productivity of future workers, but are not able to do so before hiring. Workers signal their quality to employers by investing in education, which is costly. Education is more costly to low-ability workers than to high-ability individuals because it takes time and effort to acquire. Thus, an employer can leverage the education level of a worker as a signal of their ability [40]. The effect of signal cost is best observed in a study conducted by Mayzlin et al., where they compared the reviews for hotels between a platform that only allowed reviews from verified guests (Expedia) and a platform that allowed reviews from anyone (TripAdvisor) [57]. They found that hotels with a higher incentive to fake reviews had a greater share of positive reviews on TripAdvisor than on Expedia, since faking reviews is more costly on Expedia [57].

Reputation signals, such as reviews, are not the only signals that buyers leverage to make decisions or sellers to signal their quality. So far, we have referred to signals derived from reputation systems as reputation signals. However, there are many other quantitative and qualitative signals not directly tied to a reputation system but that may signal the reputation or underlying qualities of an individual. These signals may be about their behavior or activity on the platform. For example, in online marketplaces, the number of trades that a user has been part of can serve as a signal of their experience and trustworthiness. The number of disputes against them may indicate that a potential trade partner should be cautious. The age of an account may signal their experience in the platform. In addition, various forms of verification, such as identity verification using government-issued IDs, may highlight a user's

compliance and authenticity.

Beyond market-based signals, social signals also influence consumers' decisions in online markets. eBay displays a vendor's followers, which may be taken to indicate the popularity of a seller (a social signal that we will expand on below). In the online gaming marketplace, Steam, a user can see if a friend of theirs owns a game they are interested in purchasing. In the crowdfunding platform Kickstarter, a user can see how many backers a project has received. Even seemingly trivial information, such as a profile picture, can impact consumer decision-making. While profile pictures are intuitively very important in social media or dating sites, we would not expect a seller's profile picture to influence their sales. Yet, a missing profile image or a profile image with negative facial expressions may lead to less sales on the AirBnb platform [58]. In a landmark study, Salganik et al. find that social signals in an online "cultural market" (i.e., a music marketplace) increased the unpredictability of the success of a song, since people's choices were more detached from quality alone and increasingly influenced by the choices of others [59]. In summary, while reputation systems are a key part of economic decision making in online platforms, users also leverage a variety of other signals that inform their decisions. A wide range of profile signals can contribute to the reputation of a user despite not being part of the formal definition of a reputation system (which only considers ratings given by other users).

2.1.2 Signaling in Social Media Platforms

Self-presentation online is a key aspect of social media platforms. Unlike face-to-face interactions, where users can rely on nonverbal cues and other indicators to assess a person, online interactions are often limited to more limited textual and visual information. This limitation has led to the development of *social signals*, which are signals that users leverage to communicate information about themselves, their interests, and their activities. Peoples' online signaling behaviors, like in the animal world, are used, for example, to attract dating partners. People provide and look for signals in bios and profile images to signal traits such as attractiveness, intelligence, and socioeconomic status [8]. Similarly, when deciding to accept a friend request in a platform like Facebook, users leverage various bits of profile information and structured to make this decision. Lampe et al. found that information that helps establish common referents, such as high school, hometown, major, and classes, best predicts friendship formation [6]. Lampe's work, along with their co-authors, is also illustrative of the broader approach we use in this thesis: using statistical analysis to understand how to build informative user profiles.

Social signals are also seen as a mark of social capital, knowledge, and reliability. Social capital refers to the value of social networks, relationships, and connections.

Social capital is often reflected in the number of followers, likes, and shares a user has. On Twitter/X, the number of followers a user has is often seen as a measure of their influence and authority on a given topic. On LinkedIn, the number of connections a user has is often viewed as a reflection of their professional network, access to opportunities, and expertise in a given field. Social signals have tangible benefits. For example, users with larger follower counts on Twitter/X received more customer support on social media from airline companies [60]. Some airlines, such as Cathay Pacific, would even offer specific benefits, such as lounge access, to users with a high perceived social media capital [61]

Most importantly, social signals have become increasingly associated with influence and authority. As such, "social media influencers" have emerged as a distinct class of users who use their social signals to influence the behavior and opinions of others. More importantly, social media influencers have become a key marketing channel for brands and companies that pay influencers to promote their products and services. An influencer's social signals (such as followers, views, shares, comments, etc.) are a key element in negotiating these deals, as they are used to assess the influencer's reach. Although prices vary widely, there is typically a direct relationship between the aforementioned signals and the price per post [62]. In the U.S., the influencer market economy was estimated to be worth $9.7 billion in 2020, up from $1.7 billion in 2016 [63]. Unsurprisingly, being an influencer is a highly sought-after profession, with more than 57% GenZers saying that they would like to become an influencer if given the opportunity [64]. In that context, it is unsurprising that tools and services dedicated to growing one's online influence abound. The offerings range from cheap bulk bot-based engagement to increasingly more sophisticated products involving real human engagement [65].

Another important set of signals on social media platforms are those that attempt to verify the authenticity of a user, which are particularly relevant in the context of news generation and dissemination. Social media platforms have become a key channel for political discourse and news consumption [66]. The democratization of news coverage, commentary, and dissemination have led to new journalistic dynamics, such as the raise of "citizen journalism" [67] and news influencers[66]. According to Pew, one in five Americans say they regularly get news from news influencers on social media [66]. Not surprisingly, the veracity of the news quickly became a concern. In an effort to curve the spread of misinformation, social media platforms introduced signals such as verification checkmarks (e.g., blue checkmarks on Twitter/X and Instagram). These signals are intended to help users identify credible sources and reduce the impact of false information [68]. However, the effectiveness of these signals has been questioned, particularly as they have become commercialized [15, 69].

Social signals can be thought of as softer behavioral interventions that regulate behavior in online communities. Social signals can limit misbehavior and the damage it causes to the community. In particular in the face of cheap pseudonyms, social signals that are hard to accumulate and confer greater participatory rights or privileges can help reduce misbehavior [5]. Similarly, profile signals that summarize the history of an account's online behavior encourage good behavior discourage community guideline violations [5]. Thus, social signals are a key element in regulating misbehavior in online communities.

2.1.3 Signaling in Other Online Platforms

Signals are also useful tools in platforms dedicated to knowledge sharing, such as GitHub, Wikipedia, and Stack Overflow. In these platforms, signals are used to convey information about the quality of the content and the expertise of the contributors. Similarly to signals in marketplaces, these signals fulfill two main roles: they help users make choices and encourage high-quality contributions [23, 26, 70]. A corollary of this is that they also help users avoid low-quality content and contributors. For example, on Stack Overflow, users can upvote and downvote questions and answers, helping users find the most relevant and useful content. Thus, these signals contribute to content moderation and can help moderators and platform administrators identify low-quality content and contributors.

GitHub demonstrates how signals can be used to attract new contributors. GitHub is an online platform where users can share and collaborate on code. Many open-source projects are hosted on GitHub, allowing a wide range of users to contribute. A key challenge for many maintainers who host their project on GitHub is to attract new high-quality contributors [71]. For potential contributors, a key challenge is to identify which projects are worth contributing to. For instance, is the project well documented? Is it well-tested? Is it actively maintained? Is the code up-to-date? The experience of potential contributors is similar to what Spence describes in job markets, where employers want to assess the productivity of future workers, and workers seek to assess the quality of potential employers [40]. To bridge this gap, project maintainers employ a variety of signals to display the quality of the project [71]. These signals may include badges that indicate build status, test coverage, and up-to-dateness of dependencies [26]. These signals may convey to a prospective contributor that the project is actively maintained and has well-documented and tested code.

However, social signals also incentivize contributions. In Stack Overflow and Reddit, users can earn reputation points by asking and answering questions, as well as a variety of badges for various accomplishments. These signals have little monetary value or direct benefit. High-quality answers can take a significant amount of

time to write. Why trade time for reputation points? Despite seemingly irrational motivations, several previous studies have found that users are motivated by intrinsic feelings such as the desire to help others, a sense of accomplishment, and a desire to be recognized by peers [5]. In Stack Overflow and Reddit, reputation points can provide the user with a tangible representation of their contributions, providing a sense of accomplishment and recognition. Reputation points also serve as extrinsic motivators, as they can unlock additional features and privileges on the platform, such as the ability to comment, vote and moderate content [5, 23]. They may also affect how other people in the community perceive and respond to the user. Requests from high-status users in a community lead to more compliance than if they come from anonymous or low-status members [5].

Reputation signals in online communities can also deter low-quality behavior by serving as a sanctioning device [44] or as signals of comparative performance [5]. For example, on Reddit, posts and comments can be downvoted by other users and can even accumulate a negative score, reflecting the stance of the community towards the content. In communities with scores, such as levels and victory rates in online games or platforms with leaderboards, comparative performance feedback can improve motivation [5]. On the Slashdot platform, comments could be rated between -1 and 5 and the purported goal is to "promote quality, discourage crap" [72]. However, many social platforms have been reluctant to implement negative valence signals, fearing that they may reduce engagement, lead to echo chambers or be used as a tool for harassment [73–75], challenges we discuss below.

The examples above illustrate how social signals can be powerful tools to steer behavior in online communities. Social signals matter when the people who perceive them seem them as valuable. Successful online communities are able to create systems and interfaces that provide desirable signals. Which signals (e.g., badges) are desirable is hard to predict, but can be evaluated in retrospect to inform how to design future signals. The examples above also demonstrate that careful, intentional, and iterative design is needed. As Rob Malda, cofounder of Slashdot, notes with regards to their rating system, the perspective of the user and the moderators "is always in flux."

2.2 Challenges and Threats Involving Signals in Online Platforms

Because profile signals are influential in online platforms, dishonest users may want to manipulate their signals to gain an advantage in the platform. Traditionally, past work has focused on financial incentives to manipulate online signals, such as rat-

ings and reviews in e-commerce websites or adjacent rating platforms like Yelp, a rating site for restaurants [13, 29]. These manipulations not only reduce trust in online marketplaces but may also increase benign users' exposure to scams and fraud [76]. Profile signal manipulation is also rampant across social media [27, 28]. However, academic work linking profile signal manipulation with specific harms is scarce, despite increasing reports of users who are seemingly obtaining signals, such as verification checkmarks, to launch scam campaigns [16]. There is also increasing evidence that profile signals may affect online moderation systems by potentially biasing moderators' decision making which can weaken the platform's security. We provide an overview of computer security work involving attacks on reputation systems, motivations, and impacts.

There are a wide range of attacks on reputation systems. In an early survey of these attacks, Hoffman et al. categorize attacks into five categories [13]. Self-promotion, where attacks manipulate their own reputation and falsely increase it. Whitewashing, where an attacker abuses the system to repair their reputation. Slandering, where attackers attempt to damage the reputation of other nodes (e.g., posting fake negative reviews). Orchestrated attacks, where a group of attacks employ several of the aforementioned attacks in a coordinated manner. Lastly, denial of service, where attacks can prevent the calculation and dissemination of reputation scores, thus preventing the reputation system from functioning properly [13].

There are clear financial incentives for these attacks. For example, as described above, better reputation signals often lead to better financial outcomes in markets. A highly rated hotel can attract more customers [57] or ask for higher prices [2]. Better signals are beneficial not only in online markets, but also in social media. The increasing opportunities for monetization within and outside social media platforms have led to a proliferation of incentives to manipulate signals [77–79]. For example, users with fake signals may attract unsuspecting brands and companies to pay them for advertising [62]. Similarly, fake signals can be used to portray credibility, such as cryptocurrency projects that use artificial followers to appear as if they have a large community [80] and fake GitHub stars to appear as if they have a popular project [81]. The perception of having a large active community and a popular project can lead to more funding from investors and more buyers of their cryptocurrency [81]. Finally, platforms can exacerbate the incentives to manipulate signals by allowing users to purchase signals of credibility, such as verification badges [69].

There are also clear ideological incentives for manipulating signals or portraying engagement on a topic. State actors may be interested, for example, in manipulating the perception of a political candidate or party by creating fake accounts that amplify their content and artificially inflate their popularity [82–88]. The perception of legitimacy and popularity can attract real users to engage with the content or influ-

ence their opinion on a given topic [82]. Recent reports have also highlighted how state actors, such as North Korea, have been using fake accounts on platforms like GitHub and LinkedIn to obtain employment in US companies [89]. There is also increasing evidence of ideologically motivated actors who distribute content through the accounts of financially motivated actors. In this arrangement, financially motivated individuals may offer a variety of distribution services for a fee, such as creating artificial engagement. Among the many types of customers of these services are ideologically motivated actors who seek to amplify their content and manipulate the perception of a given topic [82]. These examples highlight not only how trust is deteriorated in online communities, but also demonstrate how larger harms can be carried out through profile signal manipulation, such as political influence operations or even industrial espionage through deceptive hires.

The underpinnings of many of these attacks are the usage of bots and fake accounts. As such, the computer security literature has primarily focused on the attack of reputation systems—usually the techniques that allow users to manipulate their reputation. In particular, prior work has put a strong emphasis on techniques to improve the detection of fake accounts (i.e., Sybil accounts) as a way to mitigate fake reviews, fake followers, fake votes, and other forms of fake and inorganic behavior [13, 84, 90–99]. However, recognizing that bad faith actors will (inevitably) adapt to detection mechanisms is crucial. As such, there are other defensive mechanisms that come into play. Two strategies in particular stand out: reducing the economic incentives to manipulate signals [24] and making it easier for humans to distinguish between real and fake signals. Reducing the economic incentives may involve increasing the costs of manipulating signals, such as requiring more sophisticated bots or more human labor to create fake engagement [27]. Making it easier for humans to distinguish fake or malicious content has taken a few different forms, ranging from better tools to support content moderators [11, 100, 101] to educating end users on how to identify fake signals, such as psychological inoculation [102] and content labeling signals [103–106]. On the latter, the key challenge is still to ensure that the signal is authentic and hard to fake by bad faith actors.

Profile signals and reputation systems can deter fraud and abuse on online platforms. However, there has been substantially less work on evaluating profile signals as tools that improve the security of online platforms, such as by incentivizing adherence to community norms or decreasing the engagement with accounts that routinely violate platforms' policies (e.g., spreading disinformation, engaging in harassment, or defrauding users). Only recently has work begun to explore the perception of user profile signals, such as verification badges [15, 69]. In line with our work, recent efforts have proposed using computational methods to improve profile signals and reduce the cognitive load on users. For example, the Seriously Rapid Source Review

(SRSR) helps journalists assess and filter the verity of sources using account-level characteristics [107]. In an effort to provide signals to steer away users from toxic accounts, Im et al. proposed "synthesized social signals" on Twitter/X. In their work, they explored signals derived from an account's posting history to derive indicators of toxicity [108]. The new signal they propose helps users more easily avoid toxic accounts [108].

On the flip side, profile signals may also facilitate platform abuse both directly and indirectly. Indirectly, profile signals can lead to more lenient sanctions for misbehaving accounts. For example, in a study we conducted on the Tradingview platform, a large financial social media platform, we found that users who had high reputation scores or were paying customers were significantly less likely to be sanctioned for the same number of violations as other users [14]. We do not know if this is because human moderators may unconsciously give the benefit of the doubt to users with higher reputation scores or whether they are incentivized to give more leniency to paying customers and members with substantial contributions. However, we do have evidence that profile signals play a role. When interviewing moderators, Kuo et al. found that moderators used profile signals in users' accounts to determine the sanction they would receive [11]. There are also ways in which profile signals can explicitly facilitate platform abuse. For example, users may leverage profile signals as a stepping stone to carry out other types of harmful activities because it confers them credibility. For example, when Twitter/X allowed users to purchase verified checkmarks, there was a surge of fraud accounts that used the checkmarks to claim legitimacy and launch scams [16].

In summary, profile signals can be useful tools in creating safer online markets and communities. However, given the right incentives, dishonest users may employ a variety of techniques to manipulate these signals. Misleading profile signals can lead to a variety of negative outcomes for the platform and its users, but academic work linking manipulated profile signals to negative outcomes remains scarce. Furthermore, while statistical modeling and machine learning techniques have been instrumental in curbing profile signal manipulation by informing the design of better detection algorithms, there has been less work towards using these approaches to inform the design of user interfaces.

Chapter 3

Misleading Social Signals: A Study of Repurposed YouTube Channels

---

This chapter is adapted from our working paper:

> [109] Alejandro Cuevas, Manoel Horta Ribeiro, and Nicolas Christin. "Chameleon Channels: Measuring YouTube Accounts Repurposed for Deception and Profit". In *arXiv preprint arXiv:2507.16045*, July, 2025. Available at: https://arxiv.org/abs/2507.16045.

3.1 Motivation and Goals

This chapter investigates an understudied vector of profile signal manipulation: the resale and repurposing of high-signal social media accounts. In contrast to traditional forms of artificial engagement——such as buying bulk followers or likes——this emerging practice involves acquiring entire accounts with pre-established audiences and reputation signals. Platforms like Fameswap have enabled this market to grow significantly, with tens of thousands of social media accounts listed for sale and an advertised value exceeding $64 million [110]. Fameswap boasts the highest number of sellers and has experienced an almost 9x growth since being first described by Chu et al. in 2022 [77].

These accounts often boast high subscriber counts and engagement metrics, making them particularly attractive to buyers seeking rapid distribution, legitimacy, or monetizable reach. A key question is whether platforms like Fameswap constitute a new paradigm for online engagement—as opposed to just a new packaging for already known artificial engagement strategies. If these ready-made accounts represent

a new type of product, what are the implications for the platforms whose accounts are commercialized? We aim to answer these questions by observing how these accounts are used after being sold. We conduct this observation on YouTube by leveraging the persistent nature of channel IDs—invariant identifiers that map to a channel, and do not change even if the channel entirely alters its identity by changing handles, title, description, and/or content—what we call *repurposing* a channel, shown in Figure 3.1.

The same property that enables us to observe channel repurposing is, ultimately, what allows (potentially dishonest) users to repurpose channels. That is, currently there is no technical mechanism that would prevent a channel from changing its identity, nor there are any visual cues that would alert users to the fact that a channel has changed its identity. There are rules against this practice. However, given the financial incentives, it is unsurprising that some users are willing to circumvent these guidelines.

We frame this chapter around the concept of channel repurposing, where an account changes identity (its handle, name, content, and description) while retaining accumulated signals such as subscribers, likes, and comments. By comparing repurposed YouTube accounts listed on Fameswap and a random sample from Social Blade, we estimate the prevalence of this phenomenon and assess its implications for user safety, content trust, and platform governance. And, by observing the uses given to repurposed channels, we can also directly link profile signal manipulation to potential user harm.

This shift toward resale and repurposing marks a step toward the commodification of online audiences, with second-hand account marketplaces evolving into liquid, assisted markets [111]. These markets lower the barriers to entry for malicious actors and may give rise to new forms of specialization, such as actors who cultivate and flip accounts for resale. We hypothesize that *repurposing* channels alleviates a critical constraint for financially and ideologically motivated adversaries: rapid access to distribution without the resource expenditure required to cultivate an audience from scratch. To evaluate this hypothesis, we make the following contributions:

- We characterize the Fameswap marketplace, documenting vendor sales, advertised prices, and channel characteristics. We observed a lower bound of over USD 1M in sales.
- We manually annotate channels for potentially problematic content, expanded from YouTube's community guidelines, and then scale this annotation with statistical guarantees using LLMs and the design-based supervised learning framework. We find that 36% of channels engage in potentially problematic content.
- We estimate the prevalence of channel repurposing by analyzing a large random sample of $\sim$1.4M YouTube channels from a leading analytics platform, Social Blade. Around 0.25% of channels in the YouTube population may have been

repurposed between January and March 2025.

- We identify behavioral and content-based indicators of channel repurposing and find that specific categories of content are more significantly more likely to be disseminated through repurposed channels, such as cryptocurrency and political content, as well as content generated with artificial intelligence.

To investigate these questions, we conducted a mixed-methods study in which we characterize the Fameswap marketplace and collect YouTube channels advertised for sale, exploring the time it takes until a channel is repurposed, the change in subscribers after the change, and the rate of repurposing. We then collect a large random sample of YouTube channels from an analytics platform named Social Blade, and identify repurposed channels in the wild. Using our Fameswap and Social Blade samples of repurposed channels, we annotated each channel based on potentially problematic content categories.

Our work provides an in-depth view of an emerging vector that currently faces few mitigations. Its implications are the same for all other platforms for which there are accounts for sale. This chapter contributes to the thesis by demonstrating a real-world case in which profile signals not only fail to dissuade harm but actively facilitate it. Rather than advocating for deplatforming account marketplaces we argue for user interface changes that surface significant account changes to users. In doing so, platforms can enhance user awareness, increase resilience to manipulation, and better align their profile signals with actual account behavior.

3.2 Related Work

There is considerable related work in areas adjacent to this study, which we will discuss next. In particular, we explain how our study differs from these previous efforts.
Abusing Reputation Systems. Online reputation refers to metrics that signal user trustworthiness. First developed for online marketplaces, reputation systems—reviews, ratings, etc.—were key to their success [1]. These systems were later adopted by social media to incentivize contributions [44], encourage participation [5, 70], and signal authenticity [112]. Today, they take the form of likes, followers, badges, and similar indicators.

These indicators are valuable targets for manipulation. Over the past two decades, researchers have documented widespread abuse across platforms [51, 57, 113–115]. Misuse has evolved from bots and compromised accounts [90, 94, 116, 117] to more human-like tactics: crowdsourced engagement [29, 118, 119], automation of real user accounts [27], collusion rings [28, 120], and click farms [65, 121].

Account hijacking has similarly shifted toward high-value targets. Rather than

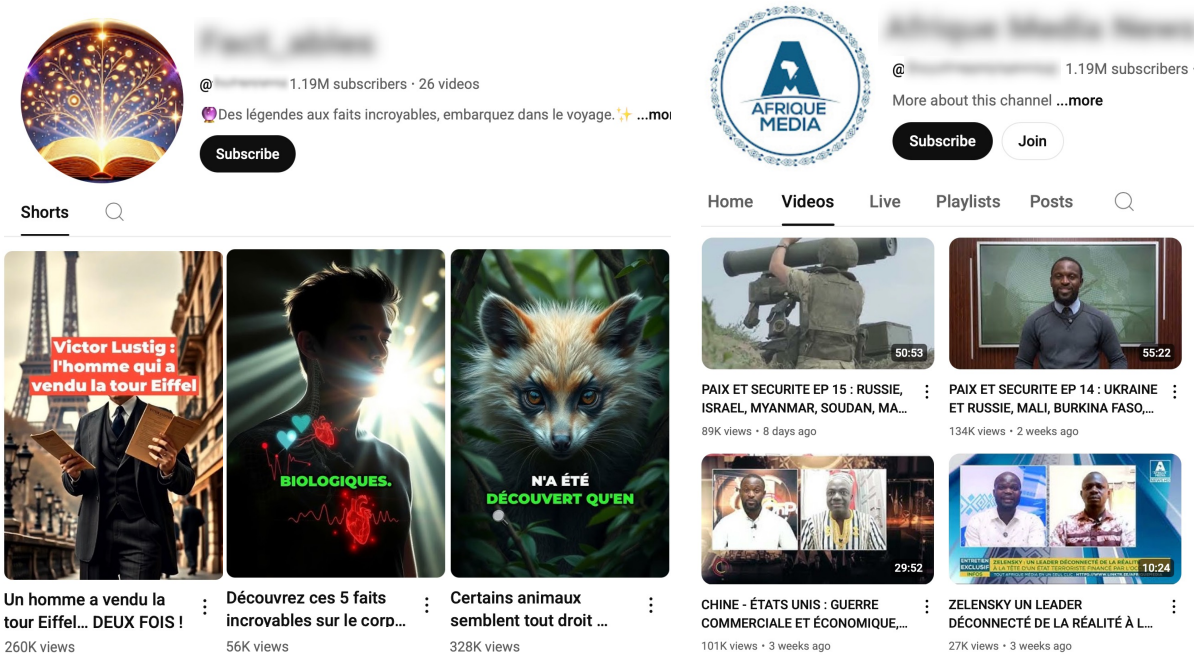
compromising many small accounts, actors now focus on fewer but influential ones [122]. This shift is tied to the ease of monetizing influence, particularly in finance and crypto spaces. Notable examples include the SEC's hijacked account used for market-moving tweets [123] and Vitalik Buterin's account used to spread phishing links [124].

Reputation abuse also occurs without hijacking. Liu et al. [125] show that scammers use legitimate-looking YouTube channels (e.g., "Ripple XRP" with 114,000 subscribers) to host ephemeral livestreams and direct viewers to off-site scams.

Licit Yet Harmful Online Influencers. Influential accounts that promote potentially harmful content but within gray areas have become an increasingly prevalent issue. For example, many reported cases of influencers promoting "meme coins" have left viewers with little recourse after the coin collapses [126, 127]. It is unclear whether YouTube's terms-of-service restrict this behavior, whose (il)legality is not even completely settled. Similarly, influential accounts have also been used to disseminate problematic narratives, such as smear campaigns, before elections. The latest case occurred just last year, when a criminal investigation linked payments from a Russian disinformation campaign to American influencers [128]. This example also illustrates the increasing interplay between financially motivated actors (in this case, influencers) and ideologically motivated actors (state sponsors).

Influence Operations. There is a substantial amount of work on state-sponsored disinformation campaigns across social media. Most of the research in this realm focuses on forensic analyses of social networks or publicly available datasets of trolls and bot accounts [84–86, 129]. Much of this work focuses on developing identification and mitigation techniques based on inauthentic behaviors, typically at scale [84, 129]. However, researchers are increasingly documenting the practices of human operatives that participate in influence operations through legitimate accounts [82, 83]. In particular, an account does not need to have a track record of posting on a political account to participate in influence operations. Memes and humor-oriented accounts are popular conduits for political messaging [82], and memetic warfare (i.e., the influence of foreign ideological spheres through memes) has long been an area of military interest [130, 131].

Disinformation and Underground Markets. Services to assist or implement disinformation campaigns have often overlapped with underground markets and communities [87, 132–134]. On the one hand, these marketplaces specialize in the dissemination of content that is frequently at odds with platform guidelines, for example, advertising restricted products (e.g., pharmaceuticals, tobacco products, unregulated gambling websites, etc.), including disinformation. On the other hand, these marketplaces also provide tools and infrastructure that can aid disinformation operators (e.g., coordinated engagement, bulletproof hosting, DDoS protection, advertising net-



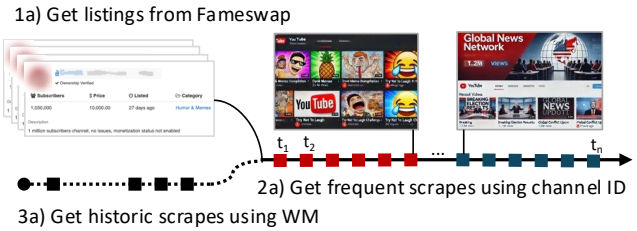
(a) A channel when listed for sale.

(b) Channel after reuse.

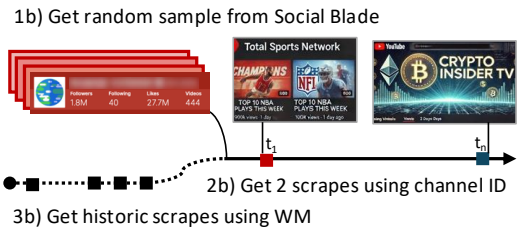
Figure 3.1: On the left, a channel listed for sale on Fameswap about entertaining facts with 1.19M subscribers. Videos are created predominantly with AI tools. This channel would later change to a news channel discussing political issues with no trace of its previous identity (on the right). For privacy, the channels' handles and titles are blurred.

works, etc.) [135]. Most importantly, off-platform sources of monetization are an important driver for channels with problematic content to thrive [77–79].

Domain and Handle Reuse. Our work shares many similarities with work on residual trust across domains and namesquatting [30, 31]. Most related to our work, Mariconti et al. study the reuse of profile names on on Twitter [136]. Their study focuses on accounts that register handles that had been de-registered to capture the residual reputation and backlinks pointing to those handles [136]. However, while some of our findings, such as prevalence of issues and the fact that reuse is often used for questionable purposes [136], overlap, our work is substantially different, given that a repurposed channel (YouTube) carries all previous followers. The market for ready-made accounts has remained largely understudied, only receiving a brief overview by Chu et al. [77], and more recently a more in-depth characterization by Beluri et al. [110]. To our knowledge, our work is the first to explore how sold accounts are repurposed and to identify channel repurposing in the wild.



(a) Collection procedure for channels listed on Fameswap.



(b) Collection procedure for reference sample.

Figure 3.2: Data collection procedures. We collected Fameswap listings (1a) daily and a large sample of Social Blade channels (1b). We then scrape snapshots through time. On Fameswap, we scraped a channel every 3 days on average (2a). For Social Blade channels, due to the sample size, we collected a snapshot in January and later in March 2025 (2b). Finally, we obtain historical snapshots using the Wayback Machine (WM; see 3a and 3b).

3.3 Background

We provide background on the market we study and relate it to some of the earlier findings in the literature.

3.3.1 The Fameswap Market

Fameswap is an online marketplace for buying and selling accounts on YouTube, TikTok, Instagram, Twitter/X, and websites. Recent measurements identify it as the largest such market by number of sellers [110]. The site resembles a standard e-commerce platform: each account has a dedicated listing with a description, price, seller reviews, and a content category (e.g., humor, sports). Buyers can bid, and the platform offers escrow, dispute resolution, and account verification—features similar to marketplaces like eBay.

Fameswap earns revenue from escrow fees (3% or a USD 50 minimum) and premium accounts, which offer advanced search, additional metrics, lower fees, and more. Payments are accepted via PayPal, wire transfers, and cryptocurrencies (e.g., BTC, ETH, USDC).

As of writing, 25,404 listings across all platforms are advertised for a combined total of USD 366.7M, claiming over 2.6B followers. For YouTube, we filtered for unique channel IDs, identifying 4,641 listings with 823M subscribers (confirmed via the YouTube API by March 31, 2025) and a total listed price of USD 160.4M. Chu et al. found 3,112 listings between 2019–2021, suggesting a nearly 9x growth [77]. The average listing price is USD 5,400 with 105,751 subscribers—comparable to Chu et

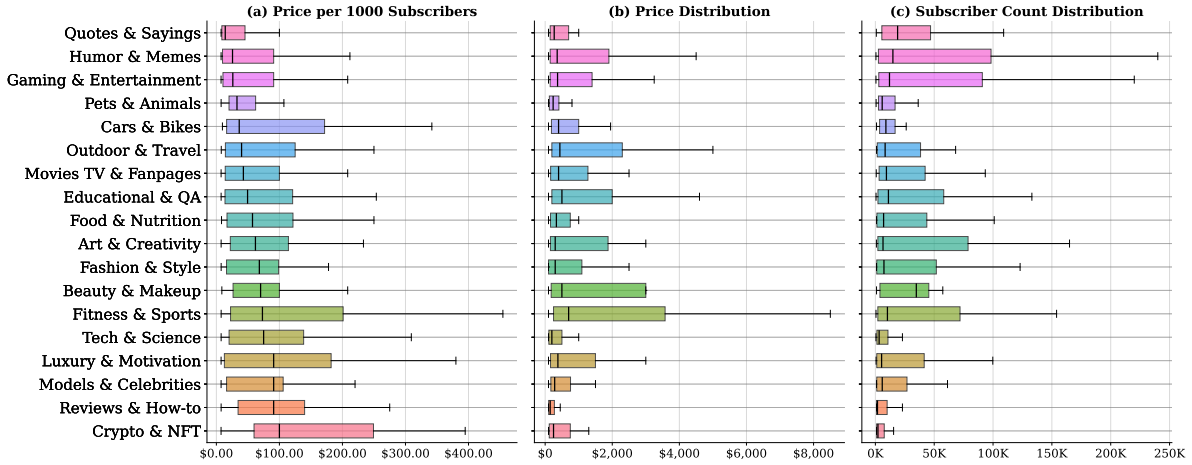


Figure 3.3: Distribution of listing prices, subscriber counts, and price per 1,000 subscribers for Fameswap listings. Outliers (Q3 + 1.5X IQR) are hidden. Categories are self-reported at the time of listing. Categories are ordered by median price per 1,000 subscribers.

al.'s findings. Like Beluri et al., we observed inflated pricing: 41 listings exceeded USD 100K and 16 exceeded USD 1M.

To contextualize these prices, we computed the cost per 1,000 subscribers—a common metric on engagement forums [115]. As shown in Figure 3.3, this varies widely by category. The median cost on Fameswap is substantially higher than the USD 16.52 per 1,000 subscribers found by Nevado-Catalán et al. [115]

Verifying sales remains difficult, as with other scraping-based estimates [37, 137]. During our study, we scraped 16,110 seller profiles——7,000 more than Beluri et al.——identifying 2,930 escrow transactions and 684 reviews[110]. Only 1,590 transactions disclosed sale amounts, totaling USD 1.16M. Though modest, this likely represents only a fraction of actual sales, which also occur via forums, Discord, Telegram, and private deals on Fameswap. [1]

Account Ownership Verification

Fameswap verifies account ownership by providing sellers with a unique randomized string that should be placed in the advertised social media profile. For example, a seller receives the string `aSgc5H3s` and places it temporarily in their YouTube channel description, allowing Fameswap to verify the ownership. This process is similar to using TXT records to verify domains.

[1] We suspect public sales skew toward smaller, cheaper channels, as the average sale price (USD 731) is far below the average listing price (USD 5,400).

Other Social Media Account Markets

Chu et al.[77] identified five marketplaces for social media accounts: SWAPD[138], Accs-Market.com [139], Trustiu (now inactive), ViralAccounts [140], and Fameswap [141]. We compare primarily with Chu et al.'s work, as Beluri et al.'s study became public after our data collection began and is referenced retrospectively. We also identified additional venues, including OGUser (similar to SWAPD) and Telegram channels.

Of these, only Fameswap and Accs-Market offer e-commerce-style interfaces, with listings, escrow, and reputation systems. ViralAccounts operates as a broker, and SWAPD functions as a private forum, limiting large-scale analysis. At study onset (October 2024), Accs-Market was inaccessible (HTTP 500), and archived pages returned HTTP 301 errors, preventing further analysis. We therefore focused on Fameswap, given its accessibility, scale, and apparent prominence.

3.3.2 YouTube Handles and IDs

In October, 2022, YouTube introduced handles as a way to uniquely identify accounts [142]. Different from channel titles, handles ("`@SomeChannel`") are unique and can be accessed through "`www.youtube.com/@SomeChannel`." Handles were rolled out gradually, but by the end of 2023 most accounts would have chosen or received a unique handle for their account. Handles resolve to a unique channel ID, allowing a channel to change its handle but keep pointing to the same channel. YouTube's channel ID is a string that begins with "UC", followed by 22 characters (letters, numbers, dashes, and underscores). Channel IDs can also be used to access a channel in the form: "www.youtube.com/channel/UC...". Importantly, while an account can change its handle, its channel ID cannot change, which allows us to monitor accounts over time.

3.4 Methods

We measure the repurposing of YouTube accounts and their participation in disseminating problematic content through the lens of sold accounts extracted from a social media marketplace (Fameswap) and by observing handle changes and channel repurposing in the wild (from Social Blade). Specifically, we structure our study in three parts. First, we define, characterize, and estimate the prevalence of channel repurposing (Section 3.5). Second, we qualitatively annotate channels to identify potentially problematic content based on YouTube's guidelines and prior work (Section 3.6). Third, we describe our framework for deriving statistically valid estimates

using textual annotations and quantitatively test what features are indicative of repurposing (Section 3.7). An end-to-end view is later provided in Figure 3.9.

This section discusses our data collection practices and each sample used in the study. The following sections describe each experiment in turn.

3.4.1 Data Collection and Sources

YouTube Channels For Sale on Fameswap

As illustrated in Figure 3.2a, we conducted daily scrapes of Fameswap from October 26, 2024, to March 31, 2025. The Fameswap interface provides a paginated list of all account listings. We scraped all historical listings. This set includes all listings not deleted or hidden from the website. On October 2024, Fameswap began displaying channel IDs with YouTube listings (whereas before they only disclosed the channel title). Together with our Fameswap scrapes, we began conducting regular scrapes of all YouTube channels advertised across Fameswap listings. We scraped channels every 3 days on average. As of March 31, 2025, we collected 4,641 YouTube channel IDs advertised for sale, each with about eight observations per month. For each channel we collected from Fameswap, we scraped their YouTube channel using the YouTube Data API v3.

Collecting a Large Sample of YouTube through Social Blade

To estimate the prevalence of channel repurposing and identify repurposing indicators beyond marketplaces, we aimed to observe this phenomenon "in the wild." Randomly sampling YouTube is difficult due to the lack of a centralized channel directory and method to enumerate channel IDs [143]. Horta Ribeiro and West address this by sampling from Social Blade [144], an analytics site that has indexed over 68M channels in the past 17 years [145]. Social Blade is a widely used reference for creators [146] and has informed multiple YouTube studies [147, 148]. Building on this method, we sampled a large set of channels from Social Blade using a sample from Horta Ribeiro and West. While Social Blade's crawling process is not public, it likely favors more popular channels. This, however, is acceptable given that our focus is content creators with larger audiences.

We sampled 1.4M channels and conducted an initial scrape from December 23, 2024, to January 21, 2025 (Figure 3.2b). We collected metadata from 1,397,586 channels and 139M corresponding videos. A second scrape, from March 21–31, 2025, captured updated metadata and 20M new videos. Only 1,351,912 channels returned data, indicating a 3.3% deletion rate. We used `yt-dlp`, a common tool for archiving YouTube content [149–151], discussed further in Section 3.9.

Samples Considered in this Study

- **Set of channels for sale on Fameswap**: All channels collected from Fameswap, n=4,461 from October 21 2024, to March 31 2025 (see Section 3.4.1).
- **Set of repurposed channels from Fameswap**: Subset of repurposed Fameswap channels, n=1,084 (see Section 3.5.2).
- **Large random sample from Social Blade**: All Social Blade channels, appearing both in January and March 2025, n=1,351,912 (see Section 3.4.1).
- **Set of repurposed channels from Social Blade (>1,000 subscribers)**: Subset of Social Blade channels that we classified as repurposed, n=1,040 (see Section 3.5.2).
- **Random subsample from Social Blade (Baseline)**: Random subset of channels drawn from Social Blade with more than >1,000 subscribers, n=3000.

Historical View of YouTube Channels

To obtain a historical view of changes, we leveraged snapshots from the Wayback Machine [152]. The Wayback Machine is a digital archive initiative by the Internet Archive; it allows users to go "back in time" to see how websites looked in the past. The Wayback Machine allows users to capture pages for archival purposes [153]. The frequency of snapshots varies per website. More popular websites (e.g., higher-ranked by Alexa, higher number of inbound links, etc.) will be crawled more often [153]. Given a sample of YouTube channels, we attempt to obtain monthly snapshots from October 2022 to March 2025.

3.5 Repurposed Channels

Our first goal is to detect channels that have been repurposed, such as the example shown in Figure 3.1. In this figure, we observe how a channel that was posting primarily entertainment facts then becomes a channel that shares political news.

We first apply a qualitative approach to define what constitutes a channel being repurposed. We use these qualitative insights to develop an LLM prompt which we can then use to scale our annotations. By relying on an expertly coded subset, we can estimate the prevalence and error rate. Lastly, we explore the time it takes for a channel to be repurposed and a channel's subscriber count after a channel has been repurposed.

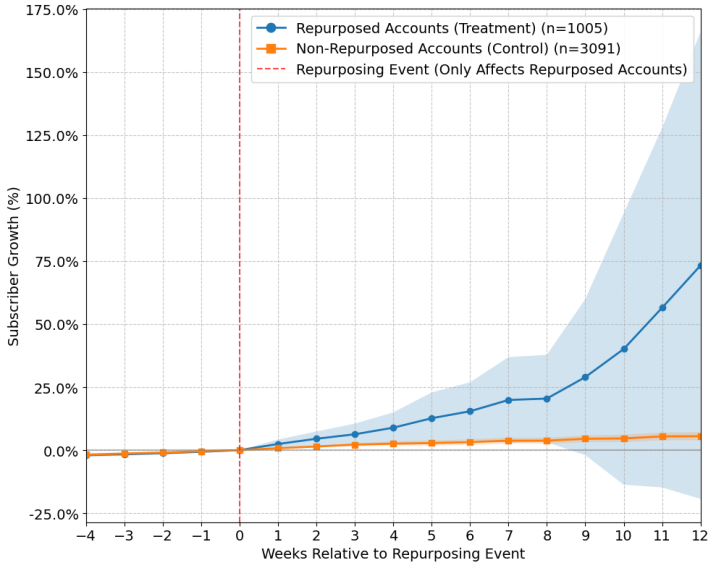


Figure 3.4: Subscriber growth comparison between repurposed (treatment) and non-repurposed channels (control). X-axis is relative to repurposing event. For non-repurposed accounts, the X-axis is based on the time they were first listed on Fameswap. Percentage growth is relative to the number of subscribers at t=0.

3.5.1 Defining Repurposed Channels

To define what channel repurposing is, we followed a three-stage process. First, the lead author manually monitored a random sample of 104 channels (10%) that changed handles after being listed for sale. Each channel was visited weekly over a month to observe changes in content, metadata, and activity. We created standardized weekly snapshots capturing each channel's videos, thumbnails, titles, timestamps, and descriptions (Appendix A.4). Second, using open coding, two researchers independently analyzed a new random sample of 54 channels (5%) using these documents. Each was prompted: "Was the channel repurposed?" Coders discussed their annotations to identify edge cases and refine definitions. Disagreements arose around dormant channels, subtle content shifts, and stylistic similarities with different themes (e.g., AI-generated videos with varying topics). This coding was exploratory; we did not compute agreement scores. Coders prioritized changes in handles and titles, using descriptions and video content as secondary evidence. Minor changes (e.g., `@HealthReporter` to `@TheHealthReporter`) were not considered repurposed. Significant discrepancies (e.g., `@HealthReporter` to `@247News`) prompted checks for overlap in topics, URLs, or referenced entities. If no continuity was found, the channel was flagged as repurposed. A substantial shift in video content (e.g., religious to crypto content in another language) served as a confirmatory signal but was

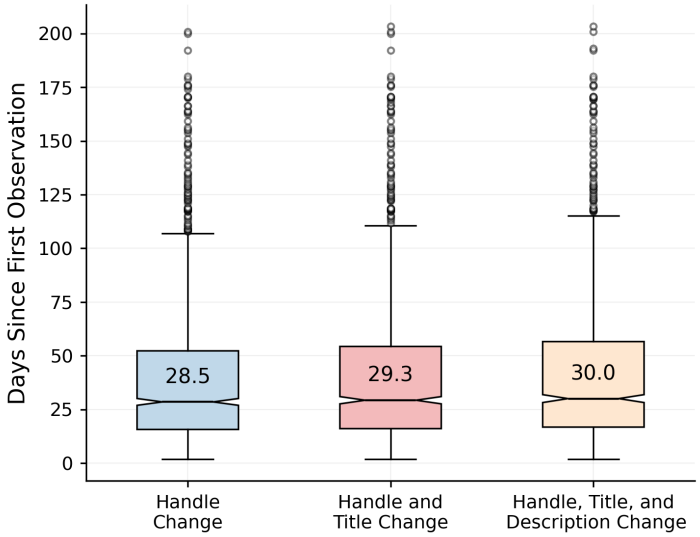


Figure 3.5: Time (days), since a channel is listed for sale, until it changes its handle, title, and description (cumulative).

not sufficient on its own. In the third step, the lead author and an external coder with a large social media following reviewed another 54 channels (5%) as a sanity check to refine the working definition.

Definition: Repurposed Channel

We consider a channel to have been repurposed from time t to $t+1$ (two observations in time) if there is no perceivable association between the channel's new handle and new title from its prior handle and title. Further, the channel description must not contain any references to its previous identity (e.g., "this channel was previously named X"), nor any overlapping text, nor any pointers to the same resources (e.g., a URL pointing to a social media account mentioned at time t).

We created an LLM prompt based on our final definition (Appendix A.2). Following, best practices suggested by Ziems et al. [154]. Using a prompt based on this codebook, we annotated Fameswap channels that had a changed handle using `gpt-4o-2024-11-20` with temperature zero and top p of one. Although there is still a lack of consensus on the optimal parameter choice for LLMs, recent work recognizes these parameters as the current standard [155]. We describe the validation in Section 3.7, alongside the framework we employ to obtain statistically valid estimates using text annotations.

Time Until Repurpose

By annotating Fameswap channels, we identified 1,084 channels (23%) (boasting 220M+ subscribers) as repurposed. For each channel, we computed how long it took them to change their handle, title, and description. As seen in Figure 3.5, the median time for a handle change was 28.5 days, followed by 29.3 days for a change in the handle and title, and 30 days for a change in handle, title, and description. As described in Section 3.4, we have an average of one observation every three days. Interestingly, handle, title, and description changes do not necessarily occur at the same time, but instead over a period of 36 hours (1.5 days), a fact that could inform an anomaly detection system. These results can be interpreted as a proxy for the median time of a sale.

Impact of Repurposing on Subscribers

To assess the impact of repurposing on subscriber count, we aligned all time series so that $t = 0$ marks the week of the channel's change (Figure 3.4). We tracked subscriber change (%) over the four weeks before and 12 weeks after this point. We compared the subscriber growth against a control group composed by channels from Fameswap that were not repurposed. For these channels, we defined t=-4 as the first week that we observed them on Fameswap. That is, their subscriber growth represents 16 weeks of growth since their first observation.

Most channels show slight but steady growth prior to the change, likely due to residual recommendation rather than active content updates. After repurposing, subscriber counts increase on average—often substantially. This is in contrast with the channels that continue to just grow slightly (channels that were not repurposed). While some unsubscribes may be masked by new subscribers, the consistent growth suggests most users remain unaware of the change and stay subscribed.

3.5.2 Repurposed Channels in the Wild

Standard NLP approaches (e.g., edit distances, n-grams) struggle to detect meaningful change in handles and titles because they lack semantic understanding [156]. A significant transformation may be semantically stable (e.g., `@TheLevenshtein` to `@EditDistance`). LLMs are well-suited for this classification task (known as entity matching) given their semantic understanding of text, but also because of their memory [157]. Because they have been trained on large text corpora, LLMs can identify connections between entities a human annotator may miss, such as in the example above [157].

LLM and Expert Annotations of Channel Repurposing

From our sample of 1.4M YouTube channels, we detected 10,200 (0.73%) channels that changed their handle between January and March 2025. These handle changes do not necessarily mean a channel was repurposed. To identify channels that meet our repurpose definition (see Section 3.5.1), we use an LLM (prompt in Appendix A.1) to classify each paired observation of channels. To validate the LLM's annotation, we randomly sampled 10% of the channels (n=1,020) and manually annotated each pair of observations according to our definition. The false positive and negative rates are 4.9% and 3.8%, respectively.

3.5.3 Prevalence Estimation

The LLM repurpose classifier achieved an accuracy between 94.2% and 96.6% (95% CI), resulting in an estimated 3,384–3,456 repurposed accounts. These results indicate that, from January 2025 to March 2025, approximately 0.24-0.25% of channels in the Social Blade population (68M+ channels) were repurposed. Of these repurposed channels, 1,074 had more than 1,000 followers. In total, these repurposed accounts have a total audience of 43,975,420 subscribers, a meaningful audience on YouTube during this period. We summarize the descriptive statistics in Table 3.1.

3.5.4 Repurposed Channels Before Observation

As described in Section 3.4.1, we queried the archive for snapshots taken after YouTube introduced handles (October 2022) for all Fameswap-listed channels (n=4,461), including those not yet repurposed, and all repurposed Social Blade channels (n=1,040).

For Fameswap, we found snapshots for 819 channels (17.6%). Of these, 332 (40.5%) had a different handle than the one listed at the time of sale, and 71 (9.5%) showed more than one handle change across archived versions. Among the repurposed Social Blade channels, 289 (27.8%) had snapshots; 116 (40.1%) had previously used different handles, and 29 (10.0%) had changed more than once. Due to the absence of additional metadata (e.g., channel descriptions), we could not assess whether these changes were substantial or cosmetic.

The frequency of handle changes may be meaningful. Prior work by Mariconti et al. [136] found that Twitter accounts with multiple profile name changes were more likely to engage in misbehavior—a pattern that may hold for repurposed YouTube channels as well.

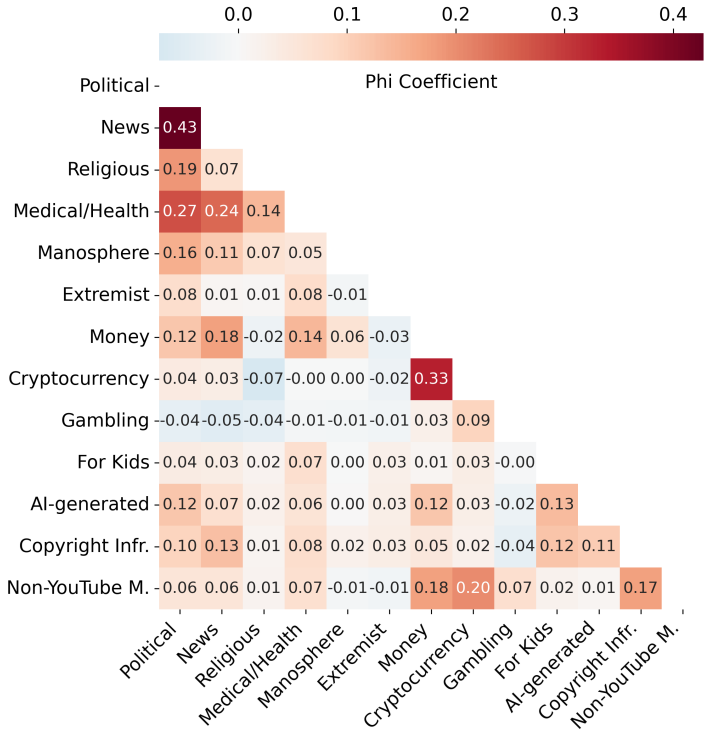


Figure 3.6: Topic correlation matrix. Values represent the Pearson correlation between variables.

3.6 Content Analysis

Next, we explore the content disseminated through repurposed channels. Reusing a channel allows a user to distribute content to an existing audience or potentially increase the chances that the recommendation algorithm amplifies new content [158, 159]. Our goal is to better characterize the actors' motivations through the lens of the content they produce after repurposing a channel. To this end, we begin by identifying relevant types of potentially problematic content, based on YouTube's community guidelines, and define a set of codes. Using these codes, we annotate the videos on repurposed YouTube channels. Lastly, we explore what types of content repurposed channels were creating prior to changing and estimate their survivability.

3.6.1 Codebook Development and Annotation

We use an LLM to extract potentially problematic content in a channel. To narrow down the types of content for which we should annotate, we first conducted a qualitative analyzed channels and their contents.

Content Categories

To develop our category annotation codebook, we began with a deductive approach, selecting a set of restricted themes (or categories) from YouTube's community guidelines [46]. However, YouTube's content restrictions are often broad, cover a wide range of topics, and definitions often lack specificity [160]. To overcome this, we opted to empirically choose the topics we annotate and based on their presence in a smaller random subset. This approach lets us develop an intuition for the types of content disseminated and how they are presented, which will later be necessary to produce few-shot examples for LLM annotation. To begin this analysis, the lead author reviewed the videos of 104 repurposed Fameswap channels (10%), flagging videos that had content covered by community guidelines.

Annotating for *kids* content was straightforward based on YouTube's definition [161], as well as *content that may infringe copyright* [162]. Similarly, based on the misinformation guidelines [163–165], we created three codes covering *political*, *medical*, and *news* content. Although news and political content overlap (i.e., politcal news), non-news political content, and non-political news content also exist and are important to distinguish [166]. We do not attempt to verify whether videos' content constitute misinformation or disinformation—the latter define by its intentionality. We address this limitation in Section 3.8.1. In addition to political and news content, we noticed a substantial number of *religious* videos. Although religion is not a topic captured in community guidelines, we opted to include it as a code given its frequent co-occurrence with geopolitical content and frequently observed content involving religious leaders with political appointments. Past work has also found religious videos to frequently co-occur with extremist and hateful content [167, 168].

On the financial side, we created a code for *Gambling* content, which is directly covered by the guidelines [169]. However, we note a substantial amount of content related to making money online through a variety of ways: pay-per-click websites, dropshipping, trading stocks, cryptocurrencies, get-rich-quick schemes, and ironically, courses to grow and monetize social media accounts. These topics are only partially covered by YouTube's guidelines (e.g., get-rich-quick schemes, fraud, broadly "spam", and broadly "scams") [170]. Cryptocurrency-related content was particularly prevalent. Informed by research on cryptocurrency scams and their potential to harm users financially [125, 171], we create a *cryptocurrency* code. We capture the rest of the aforementioned content under *money-making content*.

Using this codebook, two authors with extensive experience with the YouTube ecosystem independently coded 54 repurposed Fameswap channels (5% of reused channels). Each researcher was given a summary text document containing all of the channel's observation snapshots. We provide an example of this document in Appendix A.4. After the first round of coding, the coders discussed the procedure

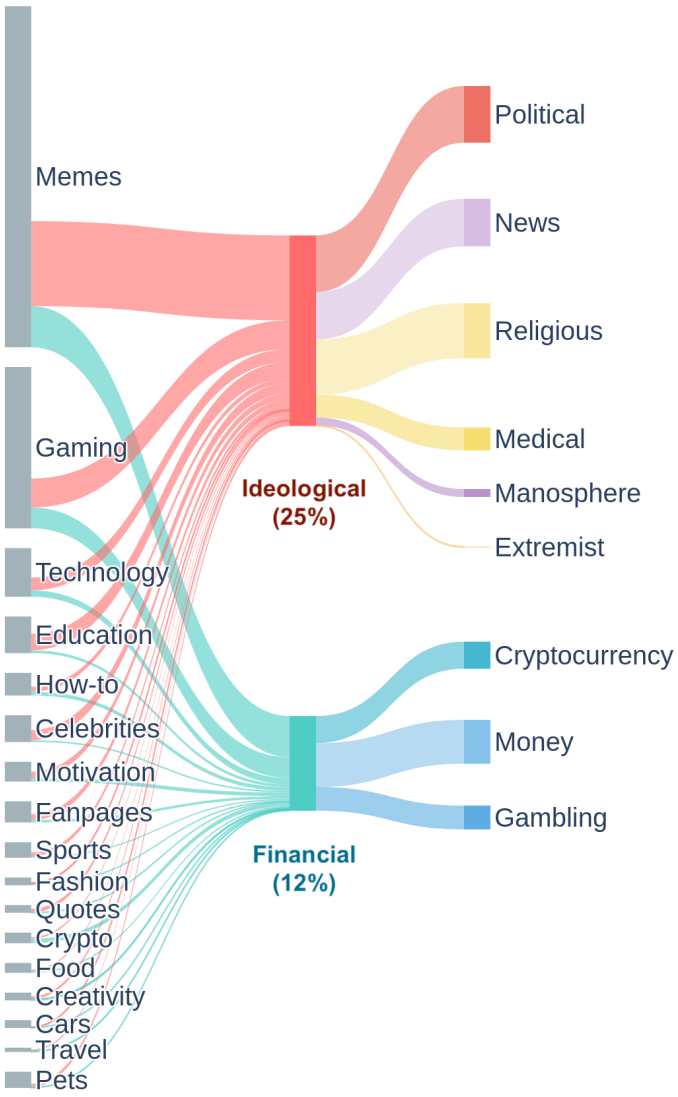


Figure 3.7: Topic presence for reused Fameswap channels. Topics on the left are self-reported at the time of listing. Topics on the right are labels assigned during our topic detection. Flows are proportional and one-to-many. That is, if a channel has several topics, flows are drawn to all destination topics proportionally.

and tweaked the codebook to address the shortcomings. We also extracted additional codes from the open-ended field if they matched community guidelines violations or were potentially harmful and not yet captured by existing codes. In this additional round, we detected and included a category for *extremist* and *manosphere* content [144] through this procedure.

We repeated an additional coding round between a member of the research team and a member external to the research team with extensive knowledge about the

creation of social media content. They coded a new randomly sampled but non-overlapping 5% of the dataset. The goal was to identify ambiguity in the existing definitions. Together with this external member, we derived our final codebook. We then converted our final codebook to a prompt for annotation. These artifacts can be found in Appendix A.1 and A.2, respectively. Note, we did not compute agreement during these steps since the goal of these steps was to create a codebook, not to use the codes for downstream analyses. We address the statisical soundness of our annotations in Section 3.7.

LLM Annotations

Using a prompt based on this codebook, we annotated three sets of channels using `gpt-4o-2024-11-20` with temperature zero and top p of one. We annotated: 1) the set of Fameswap channels that changed handles (n=1,084); 2) the set of repurposed channels we identified in our reference sample with over 1,000 followers (n=1,074), and; 3) a random set of channels without handle changes drawn from the reference sample as a baseline (n =3,000). All sets are described in Section 3.4.1.

We feed the LLM up to 50K tokens per channel, combining the channel description with video titles and descriptions. Most channels fit this limit. If a channel exceeds it, we (i) de-duplicate strings with ¿ 0.9 edit similarity to capture more diverse content, then (ii) randomly sample the remaining text until the total is $\leq$50K tokens. We do this to capture a greater diversity of information, given that some channels post many videos with the same or similar titles. We cap input at 50K (well below GPT-4o's 128 K window) because longer contexts seem to degrade classification accuracy [172].

3.6.2 Presence of Problematic Content

To better understand how problematic content manifests in repurposed channels, we examine three aspects. First, we analyze how different content types co-occur by computing correlations between annotated topics. This helps reveal common content pairings, such as between financial and ideological themes. Second, we trace how channels transition from their original categories to new potentially problematic categories after being repurposed. Using self-reported Fameswap categories and annotations after a channel has been repurposed, we highlight shifts from innocuous to concerning content types. Third, we assess whether such transitions are met with platform-level enforcement by measuring channel suspension rates over time. These analyses shed light on the nature, evolution, and consequences of problematic content in repurposed channels.

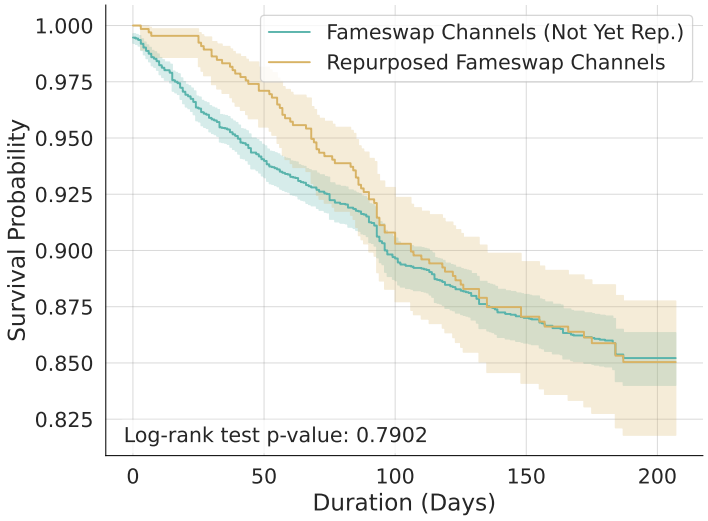


Figure 3.8: Survivability curve for Fameswap channels with both potentially problematic content detected and without. "Death" event is channel removal. Bands represent 95% CI.

Topic Correlation

We expected specific topics to be correlated, such as channels containing content about investing in stocks ("money" topic) and cryptocurrency. To investigate this, we compute the correlation matrix for all labels, shown in Figure 3.6. A topic correlation matrix for binary (0/1) topics shows the pairwise association between topics, specifically how often one topic is related to another across observations. A ϕ coefficient, which is shown for each pairwise relationship, of one would indicate that topics always co-occur, zero that they are independent, and negative one that they are anti-correlated. Expectedly, the highest correlations are between political and news content and money and cryptocurrencies. We observe a weak correlation between medical, political, and news content. We also observe a weak correlation between manosphere, political, and news content. In particular, we find a weak correlation between gambling content and non-YouTube monetization. This indicates that gambling-related channels have more information on outbound links to purchase products or services, similar to channels with cryptocurrency content. In general, we find that content that may infringe copyright and AI-generated content appears throughout the board, independent of other topics, having only a slight association with political, money-related, and child content. Lastly, while there is no strong correlation, we find that kids content co-occurs with potentially problematic categories. As noted by past work, this can particularly harmful for young audiences [173].

From Innocuous to Problematic

When listing a channel for sale on Fameswap, users choose a category that best fits the channel's content. Choosing a category improves the searchability of the channel in the marketplace. As observed in Figure 3.3, there are 18 categories that cover a wide range of topics. More than 88% of the listings have categories. We manually verified the validity of these self-assigned categories and found them to be largely accurate. Using these self-assigned categories, we investigated channel transitions into potentially problematic categories, as shown in Figure 3.7. We find that 404 (37%) of repurposed channels later displayed problematic content in their channels. Of these, 25% had ideological content and 12% had financial content. In particular, the origin category (from Fameswap) had little to do with the types of problematic content that would appear in the channel. We see that every innocuous category can ultimately disseminate problematic content.

However, not all channels necessarily transitioned into problematic categories. During our observation period, there were many channels in which we detected no objectionable content even after being repurposed. For example, a sports page became a musician's page, and an entertainment page later became an influencer's travel vlog. At the same time, some repurposed channels would later become channels associated with companies or brands. Although they were not posting objectionable content, it raises the question of whether there is a need for more transparency, given that people may rely on metrics such as subscriber counts to judge the authenticity of a company.

Rate of Suspension

iven the granularity of our observations for Fameswap channels, we estimate the rate at which YouTube suspends these accounts. To do so, we compute the survival function for two groups, repurposed Fameswap channels and those not yet repurposed, using the Kaplan-Meier estimator. As shown in Figure 3.8, all Fameswap channels exhibit an average survivability of approximately 85% after 200 days (95% CI: 79%–87.5%). Based on a log-rank test, there is no significant difference between repurposed channels and those that have yet to be sold or repurposed. Using our Social Blade scrapes, we computed the percentage of deleted or suspended channels between January and March 2025. After 90 days, we observed that 3.3% of the channels were deleted. That is, 96.7% channels from the larger Social Blade sample ($\sim$1.4M) survived. These results indicate that Fameswap channels have slightly lower survivability ($\sim$91% at 90 days). Given that 37% of the repurposed channels displayed some potentially problematic content, we expected a lower survival. However, these results indicate that repurposed channels and channels for sale generally

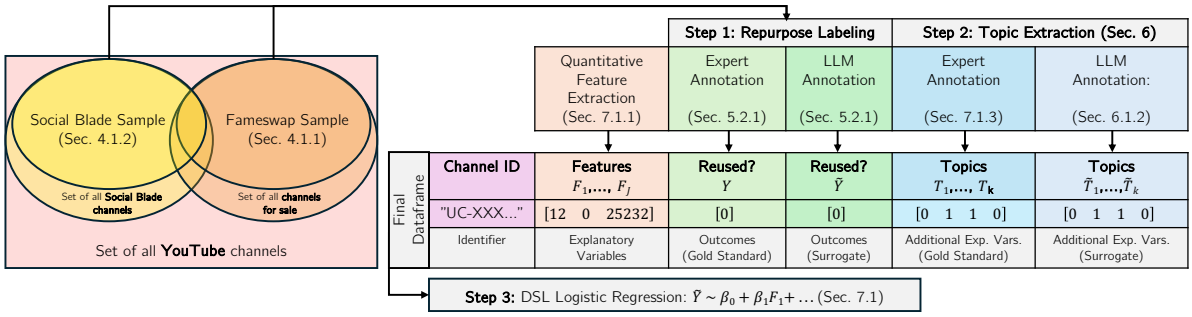


Figure 3.9: End-to-end experimental procedure, including samples, annotations, featurization, and regression.

avoid penalties.

3.7 Indicators of Channel Repurposing and Prevalence

We next identify the indicators that correlate with repurposed channels in our two channel samples: channels originating from Fameswap and those collected in the wild from Social Blade. To do this, we employ a logistic regression model with the same set of features and compare each sample against a baseline. We confirm that many topics discussed in Section 3.6 are commonly associated with channel repurposing. In addition, we identify quantitative features, such as video likes and posting behavior, that also predict channel repurposing. The goal of these experiments is to quantitatively understand what features, if any, are commonly associated with channel repurposing. Additionally, we want to test whether the potentially problematic topics that we annotated appear at a significantly higher rate across repurposed channels compared to regular channels.

3.7.1 Regression Model

We use a logistic regression model to examine the relationship between channel characteristics and channel repurposing, modeling the log-odds of the outcome as a linear combination of the predictors, in our case:

$$\log\left(\frac{p}{1-p}\right) \quad = \quad \beta_0 \; + \; \beta_1 F_1 \; + \; \cdots \; + \; \beta_j F_j \; + \; \beta_{j+1} T_1 \; + \; \cdots \; + \; \beta_{j+k} T_k \quad ,$$

where β_0 is the intercept (log-odds outcome when all predictors are zero), F_i represents quantitative features extracted from each channel, and T_i represents an indicator for each topic. In this model, a one-unit change in a covariate X_i changes the log-odds outcome by β_i; equivalently, the odds change by a factor of e^{β_i}.

Quantitative Feature Selection

We take the mean and standard deviation of all observed video views, likes, and comments. We also compute the mean and standard deviation of the number of videos posted weekly, video lengths (seconds), and the time between each video posted (days). Finally, we compute the time between the first and last video the channel has posted (months) and the oldest video the channel has, which would typically mark when the channel first started uploading content. Finally, we take the total number of videos. Table 3.1 presents the descriptive statistics for each sample. Video deletions are likely an important feature. However, because we do not have frequent snapshots of Social Blade channels (given the sample size), we miss all uploads and deletions that take place between our two Social Blade snapshots. However, we argue that our current results are robust to the omission of this covariate given that the time between videos may still capture this signal.

Using Text Annotations with Valid Estimates

We combined small-scale qualitative annotations with large-scale LLM predictions, using the latter as surrogate classifiers to enable scalable analysis. LLMs have significantly advanced computational social science [154], with growing evidence supporting their reliability across disciplines [154, 174] and platforms like YouTube [166, 175]. However, using LLM-generated labels in downstream analyses introduces measurement error—mismatch between predicted and true labels [176]. To address this, we follow Egami et al.'s framework for statistically valid inference with LLMs [176], using the Design-based Supervised Learning (DSL) framework. The process involves: (1) predicting labels with an LLM, (2) sampling a subset for expert annotation, and (3) combining predictions and gold-standard labels in DSL regression. This method requires that expert-coded samples have known, non-zero sampling probabilities. In our case, we created two such datasets: one in which we annotated whether a channel was repurposed or not (Section 3.5.2) and one where we extracted topics (Section 3.6.1).

Expert Topic Annotations

We create an expert-annotated ("gold-standard") dataset of topic extraction, by annotating a sample of 260 channels (5% out of 5,158) randomly drawn from the three groups for which we extracted topics with an LLM: repurposed Fameswap channels, repurposed Social Blade channels, and a baseline described above. These channels had a combined total of 14,354 videos, which a coder manually annotated. To create a valid expert-annotated dataset for channel repurpose, we leverage the Social

Table 3.1: Descriptive statistics for the regression features used in the Fameswap and Socialblade models.

	Fameswap (n=1,084)		**Socialblade** (n=1,074)		**Baseline** (n=3,000)	
	Mean	SD	Mean	SD	Mean	SD
(Intercept)	–	–	–	–	–	–
View Ct. (Mean)	155,519	1,104,304	119,214	1,016,701	211,258	1,085,125
View Ct. (SD)	448,550	1,552,437	195,928	1,100,404	442,672	2,061,892
Like Ct. (Mean)	5,838	36,336	1,120	8,357	1,657	9,986
Like Ct. (SD)	14,959	63,326	1,761	9,201	3,290	15,435
Comm. Ct. (Mean)	84.65	1,021.2	45.09	270.5	88.28	400.0
Comm. Ct. (SD)	246.24	2,144.2	73.72	303.6	166.17	752.0
Vid. Len. (Mean)	862.91	4,215.2	518.60	1,207.4	616.16	1,113.2
Vid. Len. (SD)	1,803.2	19,831	580.82	1,489.1	803.19	8,011.8
T. Btw. Vids. (Mean)	17.73	58.02	119.73	328.23	72.11	213.58
T. Btw. Vids. (SD)	41.14	106.6	206.89	381.47	115.11	200.46
Vids./Wk. (Mean)	5.53	8.82	3.01	5.59	3.65	18.43
Vids./Wk. (SD)	4.86	10.09	2.37	6.22	3.08	20.60
T. 1st to Oldest Vid.	17.65	22.82	57.72	64.77	84.15	62.78
Oldest Vid. Age	20.06	23.08	65.78	68.18	132.43	56.53
Subs. Count	202,981	1,013,686	40,952	169,278	103,427	1,002,724
Total Videos	158.01	474.26	127.09	806.19	373.66	1,365.7
Content Features						
Non-YT Money	0.347	0.476	0.372	0.483	0.432	0.495
AI-generated	0.105	0.307	0.066	0.248	0.022	0.147
Political	0.210	0.408	0.171	0.377	0.187	0.390
Religious	0.203	0.403	0.194	0.396	0.154	0.361
News	0.176	0.381	0.154	0.361	0.129	0.335
Medical/Health	0.090	0.286	0.121	0.327	0.153	0.360
Cryptocurrency	0.111	0.315	0.039	0.194	0.014	0.118
Gambling	0.053	0.225	0.021	0.143	0.017	0.129
Money/Stocks	0.160	0.367	0.113	0.316	0.073	0.260
Kids	0.047	0.212	0.062	0.241	0.075	0.264
P. Copyright Infr.	0.251	0.434	0.345	0.476	0.426	0.495
Manosphere	0.036	0.186	0.011	0.106	0.006	0.078

Blade sample we annotated in Section 3.5.2 and annotate a 10% sample of repurposed Fameswap channels.

3.7.2 Regression Results

We conduct two logistic regressions, first we model repurposed channels from Fameswap against a baseline drawn from Social Blade using the `R` package, `DSL` [177]. We report the coefficients in Table 3.2.

Takeaways from Quantitative Features: Channels whose oldest video is more recent

are more likely to be repurposed, as are those with a longer interval between oldest and most recent uploads. So, buyers seemingly favor channels that look established (i.e., showing a sizeable upload time span), but those are not necessarily truly long-standing. Indeed, a genuinely mature channel would feature *both* an old first upload and a long production span.

Irregular upload intervals (high standard deviation among "time-between-videos") and modestly larger subscriber bases both predict repurposing, implying that erratic schedules and an existing audience are typical markers of repurposed channels. Mean views/likes do not have statistically significant effect on predictions. *Lower* comment counts are, surprisingly, predictors of channel repurposing. A plausible hypothesis is that repurposed channels may disable or limit comments to avoid scrutiny. Subscriber count is a positive but modest predictor of channel reuse: repurposing seems less about absolute reach and more about fast content throughput in lucrative niches.

Takeaways from Qualitative Features: Repurpose odds are primarily content-driven. AI-generated content (β=0.6–1.1) is the strongest predictor of repurposing, probably because automated pipelines scale cheaply after a channel is acquired. However, we rely on disclosure to tag channels, as described in Section 3.6, so so the coefficient may also reflect self-disclosure bias. Cryptocurrency shows the next-largest effect (β=1.7 on Fameswap), followed by gambling. We hypothesize that these categories are prone to reuse because of their monetization potential. In line with previous work, we find a significant amount of off-platform monetization [77, 79], albeit only for Fameswap. Surprisingly to us, medical/health is negatively associated, suggesting either stricter platform scrutiny or lack of prominent health-related narratives. Religious content and potentially copyright-infringing content are associated with repurposing. A possible explanation could be that these types of content are used to grow channels due to their appeal (e.g., free TV shows), similar to kids content, and we are identifying trace videos prior to deletion.

3.8 Discussion and Conclusions

Our results suggest that channel repurposing involves sold channels (such as those advertised on Fameswap), as well as channels that may not have been for sale (such as most of those found through Social Blade). In either case, we find that channels with large numbers of subscribers—220M+ for Fameswap and 43M+ for Social Blade—completely repurposed their channels, erasing any perceivable association with their prior identity. Across most Fameswap accounts, channel repurposing seemed to go unnoticed, as indicated by the subsequent growth in subscribers. The median time between a channel being listed for sale and being repurposed was 30 days. However,

Table 3.2: Regression results for Fameswap and Socialblade models.

	Fameswap Coefficient (SE)		**Socialblade** Coefficient (SE)	
(Intercept)	0.8332***	(0.0982)	0.3401***	(0.0824)
View Ct. (Mean)	0.0000	(0.0000)	0.0000	(0.0000)
View Ct. (SD)	0.0000	(0.0000)	0.0000	(0.0000)
Like Ct. (Mean)	0.0000	(0.0000)	0.0000	(0.0000)
Like Ct. (SD)	0.0000***	(0.0000)	0.0000	(0.0000)
Comm. Ct. (Mean)	−0.0006*	(0.0003)	−0.0005	(0.0004)
Comm. Ct. (SD)	0.0002*	(0.0001)	0.0000	(0.0002)
Vid. Len. (Mean)	0.0001*	(0.0000)	0.0000	(0.0000)
Vid. Len. (SD)	0.0000***	(0.0000)	0.0000	(0.0000)
T. Btw. Vids. (Mean)	−0.0040*	(0.0021)	0.0002	(0.0002)
T. Btw. Vids. (SD)	0.0049***	(0.0013)	0.0022***	(0.0003)
Vids./Wk. (Mean)	0.0057	(0.0057)	0.0193*	(0.0117)
Vids./Wk. (SD)	0.0050	(0.0052)	−0.0047	(0.0053)
T. 1st to Oldest Vid.	0.0243***	(0.0043)	0.0224***	(0.0027)
Oldest Vid. Age	−0.0702***	(0.0038)	−0.0375***	(0.0025)
Subs. Count	0.0000*	(0.0000)	0.0000*	(0.0000)
Total Videos	−0.0001*	(0.0001)	−0.0002	(0.0001)
Content Features				
Non-YT Money	0.5754***	(0.1395)	0.0713	(0.0936)
AI-generated	1.0935***	(0.2057)	0.6283***	(0.2000)
Political	0.1320	(0.1749)	−0.1145	(0.1303)
Religious	0.9767***	(0.1608)	0.3910***	(0.1187)
News	0.6627***	(0.1918)	0.4305***	(0.1356)
Medical/Health	−0.4690**	(0.1866)	−0.3025**	(0.1300)
Cryptocurrency	1.6575***	(0.2491)	0.5112*	(0.2424)
Gambling	1.3716***	(0.3274)	0.2759	(0.2728)
Money/Stocks	0.2661	(0.1776)	0.1741	(0.1451)
Kids	0.7105***	(0.2130)	0.3357**	(0.1341)
P. Copyright Infr.	0.4544***	(0.1444)	0.2826**	(0.0943)
Manosphere	0.7057*	(0.3687)	0.4109	(0.4444)

some took as long as 200+ days.

On Fameswap, 37% of channels posted potentially problematic ideological and financial content. Before these accounts were repurposed, they belonged to various innocuous content categories, ranging from humorous content to sports and celebrity gossip. Figure 3.1 illustrates an example of a typical problematic transition, in which a channel that previously shared 'interesting facts' is sold and repurposed into a news channel featuring contentious geopolitical actors: Russia, Ukraine, Zelensky, etc. Due

to space constraints, we cannot discuss all remarkable cases individually, and instead, quantitatively show that potentially problematic content is significantly related to channel repurposing.

By leveraging a large sample of YouTube channels, from a random Social Blade sample, we estimated that 0.24-0.25% of channels were repurposed between January and March 2025. These results imply that there were another 16,000+ repurposed channels (out of the 68M captured by Social Blade) that we did not capture, likely with tens or hundreds of millions of subscribers.

Lastly, engagement metrics offer little indication that a channel has been or will be repurposed. Instead, video upload behaviors (i.e., time between videos) and potential gaps in their video history are better indicators for potential repurposing. However, these indicators may not be detectable or readily interpretable by everyday users.

Our results suggest that social media audiences are becoming more commoditized (i.e., have less differentiation, have widespread availability, and compete mainly on price). From the lens of transaction cost economics, social media account markets are evolving from forums (unassisted markets) into assisted markets, with lower barriers to entry and increased liquidity [111]. This transition typically makes outsourcing more attractive and, as a result, fosters specialization, a pattern that frequently emerges across cybercriminal endeavors [178–180]. In our context, this may signal the emergence of actors who specialize in harvesting organic audiences.

We do not claim nor expect these markets to be free of fraud. Many vendors may use artificial engagement to inflate their numbers and sell accounts at a premium. However, these markets are increasingly becoming more transparent, allowing prospective buyers to access more information. This continuous evolution, coupled with the apparent success of these markets, seems to indicate that a substantial portion of buyers find these markets useful.

We found evidence that channels may be repurposed into disinformation outlets. Additionally, generative artificial intelligence (GenAI) seems to play an important role in the automation of content creation, as we found evidence of GenAI-created satirical political content. In general, repurposed channels show a high correlation with AI-generated content. Putting misinformation consumption in terms of supply and demand, scholars have argued that AI does not increase the demand for misinformation [181]. However, our results add nuance to these claims. By facilitating the cultivation of audiences that can then be served mis/disinformation, GenAI may be helping increase the demand for disinformation indirectly. Furthermore, we hypothesize that GenAI enables actors to more easily produce diverse content across niches and languages, facilitating the cultivation of specific audience demographics—an asset that can be strategically leveraged when a channel is repurposed to target vulnerable populations.

There are many legitimate reasons to change handles, titles, and descriptions. Likewise, benign users may want to change their handles without drawing unwanted attention. However, current changes are imperceptible to subscribers, which leads to potential abuse. We do not suggest that platforms such as YouTube start monitoring and banning accounts from second-hand social media account marketplaces or seek to deplatform these marketplaces. These interventions will likely only succeed in making nefarious activity harder to measure [182]. Instead, a simpler start could be to explicitly call out changes in channels above a certain number of subscribers, with varying degrees of notification depending on the channel size and the significance of the change. Ultimately, improving visibility into major channel changes rather than suppressing second-hand marketplaces themselves offer a path forward, preserving user awareness while maintaining the ability to monitor this evolving ecosystem.

3.8.1 Limitations and Future Work

A main limitation of our study is that we rely on video metadata to infer topics, without analyzing the video content itself. Thus, when labeling channels as political, medical, or news-related, we do not assess whether the content is actually harmful or mis/disinformation. Fact-checking video content remains underexplored, with most work focused on news articles [183]. Assessing harm is also inherently subjective. While some cryptocurrency or money-themed channels promoted scams, we cannot generalize this finding to the entire financial category. Still, identifying channels with potential for harm is a first step. Future work could analyze transcripts, video frames, or external URLs to refine these labels and estimate the volume of problematic content rather than relying on binary classifications.

Although Fameswap is likely a key player in the second-hand social media account ecosystem, further work is needed to assess its influence. For instance, the extent of overlap across marketplaces or between marketplaces and forums is unknown. This motivated our prevalence estimation, which avoids limiting analysis to marketplace samples. As discussed in Section 3.3, Fameswap verifies ownership using unique strings. Future work could develop tools to detect these in the wild. Finally, our study focuses only on YouTube. While we expect our findings to generalize to platforms like TikTok, Twitch, or Twitter/X, this remains to be tested.

3.9 Ethics

This study involved various stakeholders and careful ethical consideration of various decisions, particularly data collection. Ultimately, our decision to conduct and publish this study is guided by the fact that it is the first to document repurposing

of social media accounts, which we find may affect millions of people. All data collected were stored on a private server hosted at our university and accessed only by members of the research team, unless otherwise indicated.

Fameswap. Fameswap is a website allegedly registered in the US with 12 years of activity. Purchasing a Fameswap membership raises ethical concerns, as it supports a service that some individuals can use to carry out questionable practices, as we found in this study. However, Fameswap is *not illegal* nor do we claim that Fameswap is an active collaborator in the practices we uncovered. Furthermore, as long as 1) alternative options are not available, and 2) the monetary amounts considered are small—especially when compared to the potential scientific value of the work in assisting with counter-measures—there is precedent in the research community, even for goods and services that are illegal, such as "booters" (DDoS-for-hire services) [184], goods from spam-advertised websites [185, 186], and even hacking services [187], with purchases ranging from a few hundred dollars to over a thousand. All these numbers are in line with or exceed our own expenses: USD 348 for a one-time monthly and yearly subscription.

The Fameswap platform does not explicitly prohibit scraping and there is also recent precedent of account registration to access data, even in problematic platforms [188–190]. The site's `robots.txt` disallows nothing (i.e., `Disallow: <empty>`. We recognize that our web scraping activities increased the platform's server load and bandwidth usage; however, we kept our requests to a minimum, collecting at most tens of pages per day. We believe our impact on Fameswap was negligible.

We considered whether to publish Fameswap's name, as it may attract attention to it. We decided to name Fameswap as it has been named in recent work [77, 110] and may encourage researchers to study similar platforms and other commercialized accounts beyond YouTube.

Users of the Fameswap Marketplace. In this study, we collected metadata from YouTube accounts listed for sale on Fameswap, as well as metadata from those accounts through YouTube. We did not attempt to deanonymize any Fameswap user nor did we employ user-level data beyond counting the number of vendors on the platform. We did not interact with users (e.g., via direct messages). We do not report quotes. We blur account and channel identifiers, where necessary, to prevent driving traffic to these channels and for user privacy.

Social Blade and the Internet Archive. Similar to Fameswap, we sought to mitigate our impact on the platforms from which we collected data: the Wayback Machine and Social Blade. Where possible, we opted to use existing data, rather than query it ourselves. We did not scrape Social Blade, but instead sampled from the data that Horta Ribeiro and West had collected [143]. To collect data from the Internet Archive, we followed their API guidance and limits [191].

YouTube. We employed yt-dlp to assist in obtaining data from YouTube during this period. It is important to note that historically there has been some concern over the use of yt-dlp. In 2020, the Recording Industry Association of America, issued a takedown notice to GitHub under the Digital Millenium Copyright Act (DMCA), requesting the removal of the project and its forks, arguing that it violated German copyright law [149]. Nonetheless, the takedown was reversed [149], and is worth noting that the main concern was using yt-dlp for video downloads, not metadata (which was our use). To date, yt-dlp remains a useful tool for academic projects, even for video downloads [151].

YouTube Users. For all YouTube channel data, we sought to incorporate the privacy guidelines and considerations proposed by Beadle et al. [192]. Although YouTube does not meet Beadle et al.'s definition of social media data[192, 193], we still applied the same privacy considerations as with other social media sites. We did not use emails or phone numbers in our analysis, even though some accounts volunteer this information in their channel metadata. We also focused our analyses on larger channels (>1,000 subscribers), as we believe that they have a lower expectation of privacy, given their large audience.

Research Team. Labeling the content in each channel required data annotation which was conducted primarily by two members of the research team and one external member. All annotators were briefed about the potential content they would encounter. Each annotator participated voluntarily, eagerly, at their leisure, and with the freedom to stop at any point without repercussions. Because we only focused on metadata, the chances of encountering disturbing content were minimized.

Chapter 4

Evaluating the Reliability of Online Anonymous Market Measurements

---

This chapter is adapted from our paper:

> [37] Alejandro Cuevas*, Fieke Miedema*, Kyle Soska, Nicolas Christin, and Rolf van Wegberg. "Measurement by Proxy: On the Accuracy of Online Marketplace Measurements". In Proceedings of the 31st USENIX Security Symposium (USENIX '22), August, 2022. USENIX Association. **These authors contributed equally*.

4.1 Motivation and Goals

While a central question in this thesis is whether profile signals can meaningfully represent user quality, particularly in adversarial environments, any attempt to answer that question requires a robust ground truth for success or failure. For online vendors, that ground truth is typically operationalized as revenue or sales. However, unlike in traditional e-commerce, where transaction data may be accessible via APIs or third-party aggregators, studies of darknet marketplaces rely almost exclusively on public scraping—making revenue estimation an inherently error-prone task.

This chapter addresses a key challenge in studying profile signals in online anonymous marketplaces: how to reliably estimate financial outcomes when only partial, noisy, or proxy data are available. Online anonymous marketplaces have been the focal point of numerous measurement efforts of the underground economy [33, 34, 178, 194–196]. To gain insight into the size and scope of illegal activities on these markets, and how these evolve over time, most of the earlier work captured the markets'

nature and their size—investigating the types of illicit products traded, and deriving the amount of listings, vendors and estimating its revenue.

Although these established insights help us understand trends in volume and types of crimes facilitated by online anonymous markets, the vast majority of earlier work is limited by their common measurement approach. All perform their analysis based on data collected through web scraping—i.e., collecting the content of public web pages displayed by the markets. This scraping is done in a measurement environment that is both inherently challenging (markets often run on low-availability servers [34] with high latencies due to the use of Tor or i2p hidden services), and even adversarial due to the market operators' extensive use of rate-limiting mechanisms such as CAPTCHAs, or their attempts to detect and ban automated activity [197]. As a result, researchers have to take missing and incomplete data for granted.

Furthermore, because they generally do not have access to the markets internal databases, researchers must use certain proxies—e.g., reviews instead of documented transactions, or listing counts—when performing analyses of, for instance, economic volumes. This "measurement-by-proxy" results in additional errors, whose size and influence on the results of the analysis are unknown. Because most of the approximations are due to missing, rather than incorrect, data, we know that many online anonymous market measurements can provide reliable lower bounds on economic activity. But by how much are they underestimating actual activity?

The potential for measurement errors does not only influence scientific research. If the confiscation of illegal assets by law enforcement is based on projected revenue calculated based on only data measured by proxy, the seized amount will often be lower than the actual turnover of the seller. In short, estimating the size of measurement error on these marketplaces, as well as what influences these errors, is not only important to validate the outcomes of previous work, but, more importantly, understanding the origins of these errors should also help shape best practices for measurements of these marketplaces moving forward.

We make the following contributions:

- We provide the first overview of measurement methodologies used in online anonymous market research and show that very few papers explain their scraping and pre-processing routines.
- We build a framework to reason about online anonymous marketplace data collection and projections. Specifically, we mathematically define a model to express possible sources of inaccuracies in online anonymous market measurements.
- Using back-end data from a seized market, we empirically measure coverage statistics and find that scraped listings differ significantly from not-scraped listings on features such as price, product category and visibility.

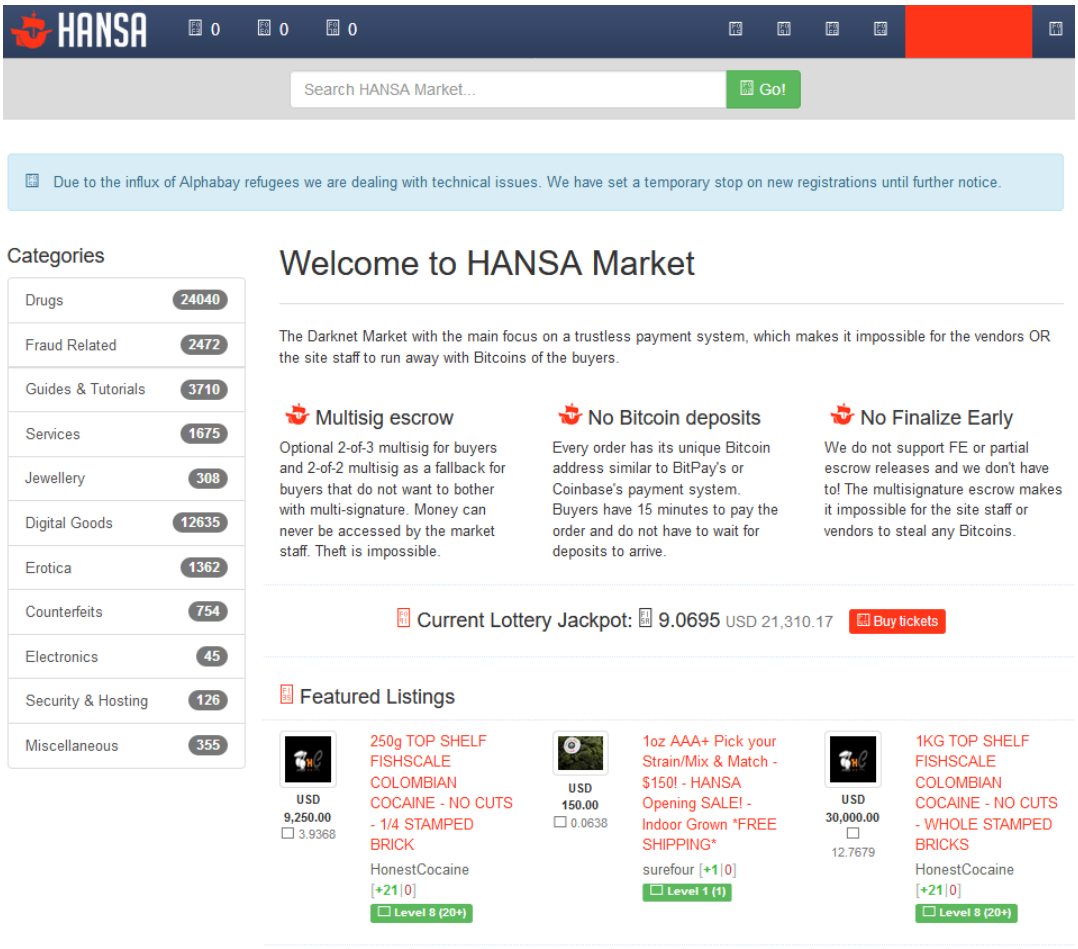


Figure 4.1: Landing page of the Hansa marketplace.

- We validate revenue calculation approaches and show that taking reviews as a proxy for transactions can lead to underestimating the total market revenue by a factor of four.
- Through simulations seeded by actual market data, we estimate the coverage impact of various scraping methodologies, rate limits, and the precision of abundance estimation techniques.

This chapter lays the methodological foundation for Chapter 5, where we ask whether profile signals can predict vendor success or failure. Having validated our revenue estimation method, we can now use it to train and evaluate predictive models on real marketplace data.

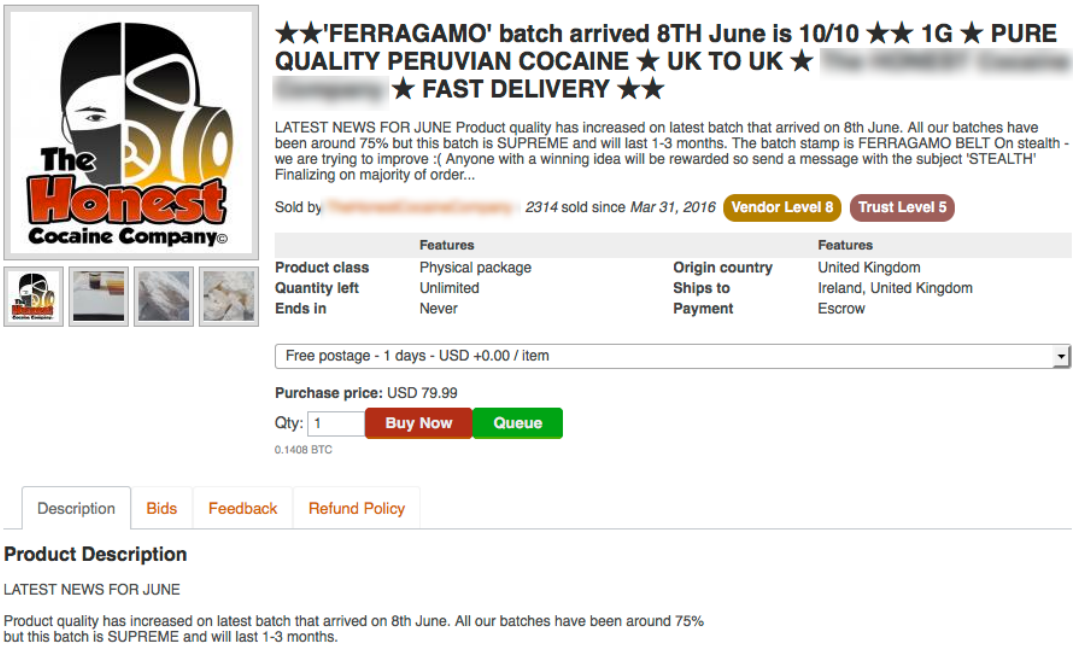


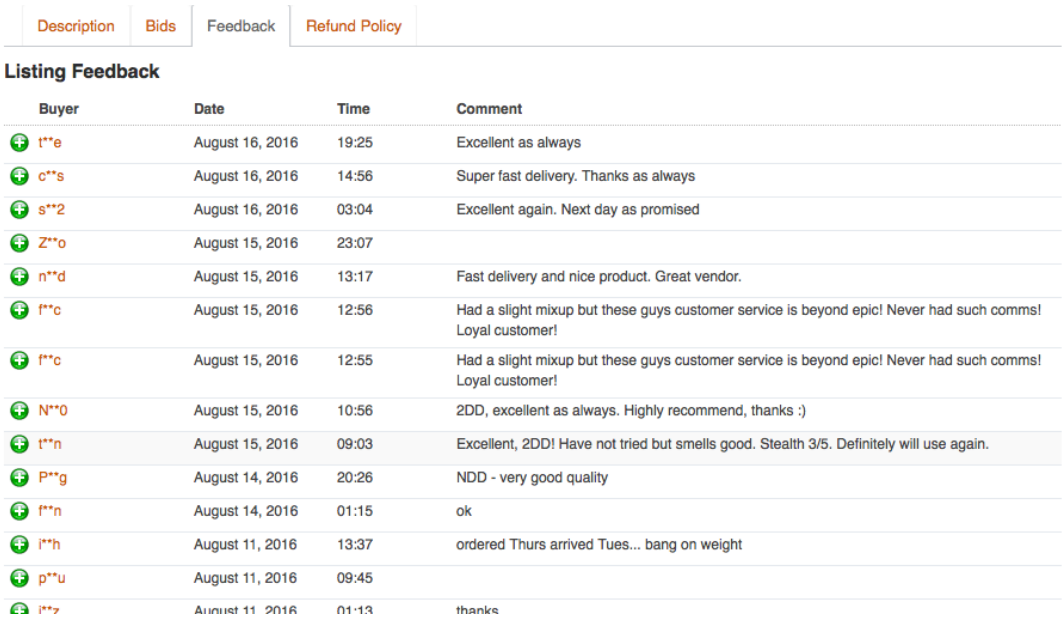
Figure 4.2: Item page of the Alphabay marketplace. Hansa, like many other dark web marketplaces, features a similar layout and design. Given the lack of Hansa screenshots, we use Alphabay as an example.

4.2 Measuring Marketplaces

An extensive body of work studying online anonymous markets has provided us with substantial insights into market economics. Since 2013, over 60 papers covering a broad range of disciplines have used data from online anonymous marketplaces or their dedicated forums. From analyzing the first modern[1] online anonymous market, Silk Road, in 2013, to evaluating a whole ecosystem of competing markets, measuring marketplaces has evolved from studying a single market to analyzing market economics [33, 34, 178, 198], security practices [199], buyer and vendor behavior [200–202] and the impact of police interventions [35, 203]. Research using scraped data of online anonymous marketplaces has also shed light on the relationship between online and offline drug trade [195].

We first summarize how, at a high level, researchers can in turn exploit publicly available market data to produce measurements in Section 4.2.1. Then, in Section 4.2.2, we survey the literature for methods used in previous research.

[1]That is, the first online market to rely on a combination of network anonymization (Tor) and distributed cryptocurrency (Bitcoin). Other "proto-markets," such as The Farmer's Market, existed prior to Silk Road, but did not rely on the combination of cryptocurrencies with anonymizing technology, and were arguably far less influential.



Description | Bids | Feedback | Refund Policy

Listing Feedback

Buyer	Date	Time	Comment
t**e	August 16, 2016	19:25	Excellent as always
c**s	August 16, 2016	14:56	Super fast delivery. Thanks as always
s**2	August 16, 2016	03:04	Excellent again. Next day as promised
Z**o	August 15, 2016	23:07	
n**d	August 15, 2016	13:17	Fast delivery and nice product. Great vendor.
f**c	August 15, 2016	12:56	Had a slight mixup but these guys customer service is beyond epic! Never had such comms! Loyal customer!
f**c	August 15, 2016	12:55	Had a slight mixup but these guys customer service is beyond epic! Never had such comms! Loyal customer!
N**0	August 15, 2016	10:56	2DD, excellent as always. Highly recommend, thanks :)
t**n	August 15, 2016	09:03	Excellent, 2DD! Have not tried but smells good. Stealth 3/5. Definitely will use again.
P**g	August 14, 2016	20:26	NDD - very good quality
f**n	August 14, 2016	01:15	ok
i**h	August 11, 2016	13:37	ordered Thurs arrived Tues... bang on weight
p**u	August 11, 2016	09:45	
i**z	August 11, 2016	01:13	thanks

Figure 4.3: Review page of the Alphabay marketplace, which we use as an example given the lack of Hansa screenshots, as described in Figure 4.2. Revenue can be estimated for each item, based on data present on the item and reviews pages.

4.2.1 Background

Anonymous online marketplaces are similar to regular online markets on the clear web such as eBay, Alibaba, or Amazon Marketplace: they serve as a platform for vendors to post listings about products or services for buyers to purchase.

Figures 4.1, 4.2, and 4.3 show the Hansa and Alphabay marketplaces as examples, whose main features carry over to most online anonymous markets.[2] The landing page users initially reach (Fig. 4.1) often features a menu with product categories, a search bar and an overview of popular listings on the market, as well as listing counts by category.

A more precise estimate of revenue can *a priori* be obtained by looking at each listing in more details. Specifically, Fig. 4.2 represents a typical listing page: title, description, geographical origin, vendor information, price, and, in some cases, total number of sales (2,314 here). To get a more precise picture of revenue over time, one may need to look at the feedback received by the vendor about the relevant item (Fig. 4.3): the review timestamps can provide an approximate idea of the purchase dates. However, if buyers are not required to leave feedback for every single purchase, using reviews as a proxy for transactions will result in under-counts.

[2]Alphabay was reportedly one of the largest online anonymous markets ever. It was seized in July of 2017, in a one-two punch that involved Hansa.

4.2.2 Literature survey

We next present an overview of measurement methodologies used in online anonymous market research. As we are interested in methodologies for scraping and preprocessing, which proxies and heuristics were used, and if/how external validation was done, we focus this overview on papers that performed this complete process – from data collection to analysis – and that investigated one or more complete markets. We thus exclude papers that use existing data, papers that focus on just one product category or country, as well as papers based on other methods, e.g., user surveys or interviews.

Scraping methods. The first step in acquiring and analyzing dark net market data is to *scrape* the relevant markets; that is, to capture copies of the web pages describing item listings, vendors, and feedback so that they can be subsequently used for further processing and analysis. Relatively few authors [33, 34, 204] provide extensive details on their scraping methods: number of scrapes, frequency, crawling mechanics, size, design goals or explanations of failed scrapes. Baravalle et al. [205], Baravalle and Lee [206], and Hayes et al. [207] all describe the technical implementation of the scraper, however they do not explicitly mention the number of scrapes they collect nor the frequency. Dittus et al. [195] and Aldridge et al. [194, 208] use a single-shot scrape and merely provide details on the hyperlinks the scraper collected and followed. Similarly, Dolliver [209] uses a single scrape but only discusses the scraper design and mechanics. Van Wegberg et al. [178] describe the number and frequency of their scrapes, but do not discuss their scraper design. This finding of limited disclosure of crawling approaches in research is similar to the general survey on crawling methods in research from Ahmad et al., who sampled 350 papers that use a crawling methodology and found that 36% of their sample can be classified as *not repeatable* [210].

Post-processing. Once a market has been scraped, the relevant pages need to be post-processed before they can be analyzed. Basically, this means 1) parsing each page to extract salient information – e.g., listing title and vendor name, and 2) "cleaning up" the parsed data. More precisely, we look for discussion of parsing, deduplication, recoding, review-to-item listing matching, and completeness validation. Surprisingly to us, post-processing pipelines are seldom discussed across previous work. Completeness validation are most often discussed [34, 195, 204, 206, 211], but techniques are not standardized: authors instead employ a variety of custom strategies to assess the completeness of their datasets. Similarly, only a few authors describe their parsing procedures [34, 204, 206] and deduplication methods [33, 34, 204]. In two cases, the authors describe their recoding procedures [195, 209].

Proxies and heuristics. Parsed and cleaned data are then analyzed to provide insights about the market. This is where, for instance, researchers extrapolate from re-

views to get a sense of economic revenue. Revenue can be defined as `price` × `sales`. Since both these features are not always directly scrapable, proxies have been used for analysis. We did find some consensus across the use of proxies and heuristics.

As hinted above, authors frequently use reviews left on listings as a proxy of transactions [33, 34, 194, 195, 208]. However, Celestini et al. cast doubt on this procedure [204]. Because some buyers will not leave a review, the number of reviews should always be considered as a lower bound for the number of sales. Most authors [33, 34, 178, 194, 206, 208] use the listing price as a proxy for the paid sales price. This proxy has two drawbacks, which are discussed in the aforementioned papers.

First, "holding prices," where vendors increase listing prices astronomically to signal an item is out of stock, are a known phenomenon across online anonymous markets, and authors employed various heuristics, mostly grounded in domain expertise and manual analysis, to filter/include them [34, 194, 206]. Second, the listing price changes over time, which means that a later price might differ from an earlier sales price. Only Soska and Christin [34] account for this by using the listing price scraped closest in time to a review timestamp.

Procedures to estimate the number of vendors either count the number of scraped pages directly [204, 208], or conditioned on activity, defined as having a listing in a given period of time [33, 34]. Some authors identify wholesale (as opposed to retail-level) vendors by looking at quantity [209] or listing price [194]. In terms of proxies not related to transactions, there is an apparent consensus in using shipping to/from destination for determining geographic location of items [33, 195, 204, 205, 209]. Item categorization (e.g., to determine whether a product is a narcotic, a prescription drug, or a weapon) sometimes relies on the marketplace's advertised categories [33, 204], sometimes on machine learning models [34, 178], or sometimes on manual analysis [195, 209].

External validation. Validating the reliability of collected data is an important step in online measurement studies. Past work employs a variety of external data sources to assess the reliability of the collected data. For instance, Soska and Christin compare the data they collected against data contained in trial evidence, criminal complaints, and leaked pages [34]. Van Wegberg et al. also used criminal complaints for validation [178]. Similarly, Tai et al. also use court records in the context of vendor tracing across marketplaces [202]. Tai et al. complement their evaluation with a publicly available (at the time) crowd-sourced vendor database [212]. Last, Wang et al. compare their collected data against past studies [200].

Closest to our work, Rossy et al. use data collected by police following shutdown operations [213], and two efforts use ground truth data from back-end sources. Van de Laarschot and Van Wegberg use data from Hansa [199], and Bradley uses (partial) data from Silk Road 2.0 [214]. Interestingly, neither effort uses this back-end data for

validation, but instead relies on it as ground truth for analysis.

Validating the collected data by using internal data is in specific cases also possible. On some marketplaces, the incremental identifier that is used in the database to uniquely identify for example a listing or user, is also used in the url or visible on the page. McCoy et al. the authors leverage the fact that each page on an underground market is identified by an increasing ID to crawl it[215]. This is a good strategy to check how comprehensive a measurement is, and should be discussed as an option in this chapter (obviously for cases where this ID is available).

For the validation of our own measurements, we will use all papers that have used Hansa data (either scraped or the database). These are Kruithof et al. [216], Lewis [217], Dittus et al. [195] and Van de Laarschot and Van Wegberg [199].

4.3 Methodology

We first formalize an abstraction for online anonymous marketplaces in Section 4.4. This abstraction can be used to test the impact of different hypothetical scenarios – e.g., what coverage do we get as we scrape more? Additionally, with accurate parameters, we can extend the insights that we derive from the specific marketplace we study, Hansa, to other marketplaces. We will later use this abstraction to model data collection and data analysis methods in simulated experiments.

We leverage three datasets for our analysis of losses when measuring marketplaces. The first dataset consists of scrapes collected from the public view of Hansa (Section 4.5.1). The second dataset is the Hansa database, i.e., the administrator's view, seized in the Hansa takedown operation by the Dutch National Police (Section 4.5.2). The third dataset is a set of simulated marketplaces that are scraped with different scraping procedures and parameters (Section 4.5.3).

We provide the first empirical measurement of scraping coverage based on ground truth data from Hansa in Section 4.6.1. By matching listings, reviews and users in a scrape to the same objects in the database at that moment in time, we measure both instantaneous and cumulative coverage. This experiment first confirms that not all objects are captured. In Section 4.6.2, we divide the objects in groups of *scraped* and *not-scraped* objects and show that significant differences exist between them, evidencing different biases in scraped data.

Revenue calculations are a key part of marketplace research. To better understand the impact of the biases that originate from incomplete scrapes and conservative heuristics, we calculate the revenue of one month of Hansa's revenue based on different data sources and different proxies and heuristics in Section 4.7. Based on these different revenues, we can define the different loss categories and their size. For example, how much revenue do you underestimate by using reviews as a transaction,

versus orders as a transaction?

Finally, we conduct a series of simulations to understand the collection loss incurred by different scraping approaches in Section 4.8. We compare the coverage of one and two-shot scrapes and we estimate the impact of scraping consistency on coverage. We then explore the effectiveness of different abundance estimators. Certain pages may yield higher coverage than others (e.g. a listing with many reviews vs. one with none). Thus, given the adversarial environment of online anonymous marketplaces and the heavy impact that rate limiting has on coverage, we evaluated the design of a scraper that splits its scraping budget between rescraping listings with most feedback growth and discovering new listings.

4.4 Modeling Marketplaces

To reason about marketplace data collection and projections, we first need to mathematically define a model to express what a market is. We describe the model components in Section 4.4.1. We then describe data collection and analysis methods in Section 4.4.2 and Section 4.4.3, and the types of losses that arise from those functions in Section 4.4.4.

4.4.1 Model Components

Our model describes the relationships between the components of a marketplace: 1) the *states* of a marketplace, 2) the core *objects* of a marketplace, 3) the *views* that actors – such as a marketplace customer, vendor, or administrator – have of the state, and 4) the means by which states are *altered*. For instance, we consider a scrape to be a representation of one state, which observes various objects – such as reviews – through the public-facing view of the marketplace – i.e., what a customer would see – which does not alter the state.

States. The *state* of the marketplace, denoted σ_t for each time t, contains all of the information currently stored on the marketplace's back-end servers. Our focus in this work is on centralized marketplaces where a state takes the form of a database, which contains tables on marketplace objects. The marketplace *transcript* at time t is the complete history of all states from the beginning of the marketplace until t, namely $\mathrm{T}_t = \bigcup_t \sigma_t$. If the marketplace does not support deletions of states then $\mathrm{T}_t = \sigma_t$.

Objects. *Objects* are the core elements which constitute a marketplace. They are contained in a state, can be seen in a view, and can be altered through an operation. Objects include *users* (containing both vendors and customers), *item listings*, *reviews* and *transactions*. While there may be other objects that exist in a marketplace database

(e.g., cryptocurrency wallets), a marketplace at least contains these. The objects themselves can have different attributes related to them. For example, a listing can have attributes such as price, shipping origin and item description.

Views. At any point in time, a marketplace offers different views to different actors. Most commonly, a customer can observe the marketplace state σ_t from a public view, which we denote σ_t^{public}. This view allows the actor to observe item price and previous reviews but may not have any information on hidden listings, or on listings deleted before time t. On the other hand, a marketplace administrator may be able to see all the information from the marketplace. The administrator view provides access to the collection of states σ_t^{admin}, which if complete, represent the marketplace transcript.

For the remainder of the chapter, we assume that scrapes always rely on public views of the marketplace, while a marketplace take-down by law enforcement allows access to either the complete transcript (e.g., if the administrators kept backups of old states), or at least a partial view of the transcript. While out of scope for this work, it would also be desirable to consider vendor and moderator views, as law enforcement has been known to infiltrate these accounts, which represent a practical vantage point through which different signals can be extracted from a state. While these views are not as comprehensive as the administrator view, they should provide more information than is available in public views.

Operations. The state of the marketplace evolves via the *insertion* and *deletion* of objects, where *updating* the marketplace state is modeled as a deletion followed by an insertion. These operations affect σ_t, and thus all views of the marketplace and imply that future states are generally neither a proper subset nor a superset of previous states. Some operations can also affect specific views of the marketplace state. For instance, a *hide* operation on a listing, affects the public view but not the administrator view. On the other hand, the deletion of database backups or logs affect the administrator view but not the public view.

4.4.2 Data Collection

We define data collection functions as those which aim to retrieve the state of the marketplace at a time t and with a given *view*. The most common collection function is the *scraping* function, which uses $view = public$. We model a scraper that collects marketplace information from time m to time n as:

$$S_{m \to n}^{public} = \bigcup_{i \in [m,n]} x_i \leftarrow s(\sigma_i^{public}) \,. \tag{4.1}$$

Here, $s(\cdot)$ is a scraping function that takes in a marketplace state and returns the

subset of data sampled according to a certain distribution. Typically, this function will either return the empty set when no information is collected on a particular state, or pieces of data representing the collective information on a few pages that were scraped.

4.4.3 Data Analysis

Data analysis is using data that has been collected, to measure any characteristic of the marketplace. Analysis functions mostly focus on taking the objects available in the public view (item listings, users and reviews) to approximate the objects available in the admin view (transactions). For instance, if one uses reviews as a proxy for transactions, we formally have:

$$|Tr_l^{admin}| \geq |R_l^{public}|\,, \tag{4.2}$$

that is, the number of actual transactions Tr for a listing l in the admin view will always be greater or equal to the number of reviews R for l present the public view. In other words, the number of reviews is a lower bound for the number of transactions. As discussed in Section 4.2, the functions applied to transform the "raw" collected object files to analyzable datasets are often overlooked in data analysis. These are mostly functions that combine different approximated states to one approximated transcript of a marketplace.

4.4.4 Losses

We define two broad types of loss in our model: collection loss and inference loss. First, *collection loss* results from any process which causes the data collection of a state to be different from the true state. Formally, between times m and n, for a given *view*, we have:

$$\textit{Collection Loss}_{m\to n}^{view} = \left[\bigcup_{i\in[m,n]} \sigma_i^{view}\right] - S_{m\to n}^{view}\,. \tag{4.3}$$

(In the present discussion, $view = public$, but the loss definition generalizes to other views.) There are numerous sources of collection loss, including technical sources of loss (e.g., network errors, rate-limiting, backup loss), scraping-related losses (e.g., scraper design and website layout), and simply data loss that occurs over time due to data updates (e.g., deletion of objects from public view). In practice, collection loss can be defined as $1 - coverage$.

Second, we consider *inference loss*. For instance, to infer transactions, we need to match reviews to their corresponding listing. In this process, we may find matching and/or duplication issues which can lead to loss. For instance, attempting to

detect when two *a priori* different reviews match the same sale (i.e., the buyer simply updated their feedback message) may lead to a loss, when this matching process reaches an incorrect conclusion.

4.5 Datasets

For our analysis we leverage three sources of data: Hansa scrapes (Section 4.5.1), Hansa database (Section 4.5.2) and simulations (Section 4.5.3).

4.5.1 Public View – Hansa Scrapes

We built our scraper using Scrapy [218], on top of Tor [219]. We scraped Hansa 17 times between late 2015 and mid-2017, collecting a total of 332,795 pages amounting to 39.5 GB of data.[3] The precise scrape dates can be found in Appendix B.4. The scrapes provide a picture of Hansa during three periods of time: its initial stages (late 2015), its mature stage (mid-2016), and its peak prior to takedown (mid-2017). Out of the 17 scrapes, 3 of them failed due to authentication problems (due to cookies being invalidated), leaving 14 scrapes for analysis. Following the scraping, we proceeded to parse the pages and deduplicate entries. Below, we describe each of these processes.

Scraping procedure. We designed the scraper with *reliability* (to reduce data loss) and *stealth* (to prevent evasion) as primary goals. Our scraping algorithm was *depth-first* across parallel Tor circuits. To build the scraper we first performed a manual analysis of Hansa's layout. We then built a set of regular expressions for the URLs in the marketplace. This also allowed us to restrict certain requests to be sent when following links – e.g., add items to cart, checkout, etc. On session start, we provided the scraper with a session cookie manually obtained after solving a CAPTCHA. Scraping sessions ranged from a few minutes to a few days. When carrying out requests, our scraper randomly selects among a set of pre-built Tor circuits as a way of bypassing anti-DDoS mechanisms by "spreading the load" over multiple connections.

Ideally, we would want our scraper to instantaneously capture a snapshot of a marketplace, and to do so frequently. This would allow us to capture changes in the marketplace state as they happen and avoid missing objects that may be changed or deleted as time passes. In practice, however, we need to limit our requests so that we 1) do not alert the marketplace's operators and resultingly get blocked, and 2) do not significantly impact marketplace operations by flooding it with traffic. We performed approximately 12 requests per minute.

[3]The sanitized scrape data can be found at `https://arima.cylab.cmu.edu/markets/`

Timeline. During our initial scrapes (late 2015, early 2016) we observed slow growth in daily revenue – on average ∼$2,000 per day. As a result, we decreased our scraping rate throughout 2016 and early 2017, where Hansa had modest growth, and remained far behind competing marketplaces, notably Alphabay. Finally, following the Alphabay takedown in July 2017, Hansa saw a surge in popularity, so we began scraping frequently again.

Parsing and deduplication. We then extracted information from our scrapes through a parsing process. We iteratively adapted our parser to account for changes in the Hansa website over time which caused information, such as data fields, to be added, modified, and/or removed. One of our main parsing objectives was to ensure reviews are correctly paired with item listings, since this forms the basis of revenue calculation.

Scraping provides a snapshot of the marketplace (public) view at one point in time. Subsequent scrapes capture new information as well as substantial duplicate information. Deduplicating listings and vendors is trivial since they have unique identifiers. However, review deduplication is more challenging. We consider a review to be a duplicate if the author,[4] message, and timestamp[5] are the same and correspond to the same listing. We note that the review editing feature that Hansa provided may have caused a few overcounts given that it alters the timestamp of the review.

Author / Scrape Date	Vendors	Listings	Reviews	Est. revenue
Kruithof et al. (2016/1/11 → 2016/1/15)	219	4,829	–	–
This work (→ 2016/1/17)	282	5,987	2,847	$134,145
Lewis (2016/12/10→ 2016/12/16)	–	43,841	–	∼$3,000,000
This work (→ 2016/12/14)	840	21,185[6]	64,123	$2,885,133
Dittus et al. (2017/6 → 2017/7)	2,300	51,800	91,900[7]	–
This work (→ 2017/7/7)	1,639	48,330	186,893	$10,305,493

Table 4.1: Comparisons between Hansa studies. We include counts of reviews without price information. However, we omit them when estimating revenue.

[4]Regardless of username length, Hansa only displayed the first and last character of a review author with three asterisks in-between, e.g. `a***b`.

[5]Hansa originally provided timestamps with a one-minute granularity, before switching to a one-day granularity.

[6]We skipped 27,145 listings, unable to confirm their scrape date.

[7]The review discrepancy is likely caused by the fact that Dittus et al. focus on scraping "product catalogs," missing reviews left on vendor pages.

External validation. We first validated the completeness of our scrapes by comparing them to information contained in other work on Hansa. Table 4.1 summarizes this comparison. Kruithof et al. conducted a scrape between January 11th and January 16th, 2016 [216]. Lewis conducted a scrape between December 10th and December 16th, 2016 [217] and Dittus et al. conducted a scrape "in late June to early July 2017" [195].

For all three datasets, we can directly compare our review counts since reviews are timestamped, which allows us to drop all reviews which do not fall in the scraping dates mentioned by the authors. However, in terms of listings and vendors, we can only do direct comparisons with Kruithof et al. and Dittus et al.'s datasets, since we have a Hansa scrape on January 17th 2016, and on July 7th 2017. This is because vendor and listing pages are not timestamped, so we cannot determine how many listings or vendors were present at the time of Lewis's scrape. Instead, we approximate the listings and vendors we had at the time of Lewis's scrape by only counting listings (and their corresponding vendors) which we had seen prior to July 7th and had more than one review. Table 4.1 shows that our scrapes mostly match measurements of earlier work. This is reassuring, given the scraping gap between 2016 and 2017.

4.5.2 Admin View – Hansa Database

We next use Hansa data obtained by the Dutch National Police on July 20, 2017 when the market was taken down [220]. At that time, the Dutch National Police had been running Hansa through a covert operation for exactly a month, starting on June 20, 2017. Using this data raises ethical considerations that we discuss in Section 4.10.

This data we have at our disposal is, in practice, a copy of the Hansa "back-end" database, that consists of 64 tables created by the marketplace administrators, as well as 76 back-up tables containing data from specific, earlier time periods. Using our earlier notations, we thus have both the "final state" of the marketplace, $\sigma_t^{admin} = \sigma_t$ (where $t =$"July 20, 2017") and some of the $\sigma_{t'}^{admin}$ for $t' < t$. We focus on quantifying measurement loss that occurs when we rely on scrapes of public views to reconstruct the entire market transcript (see Section 4.4 for definitions). For this analysis, we only need the data that pertains to the main objects (see Section 4.4.1) of the back-end database: listings, reviews, users, orders and transactions. Because older data was deleted as time went by, the final state of the Hansa market is not identical to the complete transcript of the market. Fortunately, the presence of back-up databases allowed us to partially recover that transcript. To that effect, we took the following preprocessing steps.

First, we noticed that a number of objects were present in different back-up tables.

Object	Time period	Records	Highest ID	Missing (%)	After filtering
Listings	2015/3/19–2017/7/20	123,143	123,969	0.67%	123,133
Reviews	2015/3/19–2017/7/20	258,184	260,853	1.02%	258,184
Users	2015/3/18–2017/7/20	419,323	432,287	3.00%	419,323
Orders	2015/6/17–2017/7/20	312,128	589,038	47.01%	192,708
Transactions	2016/1/28–2017/7/20	1,686,919	1,715,485	1.67%	505,883

Table 4.2: Marketplace objects from Hansa back-end

For each object type (e.g., orders, users, ...), we combined all of these records into a single, merged "complete" table. Whenever we found multiple records corresponding to a single object, we kept the most recent record.

Second, we then pruned these complete tables to ensure they only hold data pertaining to "finalized" purchases, as opposed to aborted attempts. For instance, we filtered out of the complete order table entries referring to 1) orders without an associated transaction (money transfer), 2) orders that were declined by the seller, and 3) orders that were refunded. Similarly, we removed transactions between internal wallets to avoid double-counting transactions.

Third, we checked data completeness in each table. Each table has an incremental unique identifier, which we can use to infer the amount of records purged from the database, simply by comparing the record count with the highest unique identifier.

Table 4.2 summarizes the outcome of our data processing. It shows the time period data is available from, the amount of records, the highest identifier, the percentage of missing data and finally the total amount of records available for analysis after filtering. The order table seemingly only holds roughly 50% of all orders, even when all the available backup tables are used. However, plotting the data over time, in Figure 4.4, shows a much more nuanced picture: order data is sporadically highly available, and sometimes completely missing. This shows that even after seizing a marketplace, one does not necessarily possess the ability to completely recreate the whole transcript of everything that happened during the marketplace's lifespan. In contrast with Van de Laarschot and Van Wegberg [199], who used the same dataset, we do not reconstruct purged orders by using their reviews as a proxy.[8]

[8]Van de Laarschot and Van Wegberg used feedback to reconstruct missing order data, whereas we - for investigating differences between (scraped) reviews and orders - turned to data from the previously untapped and more complete transaction table.

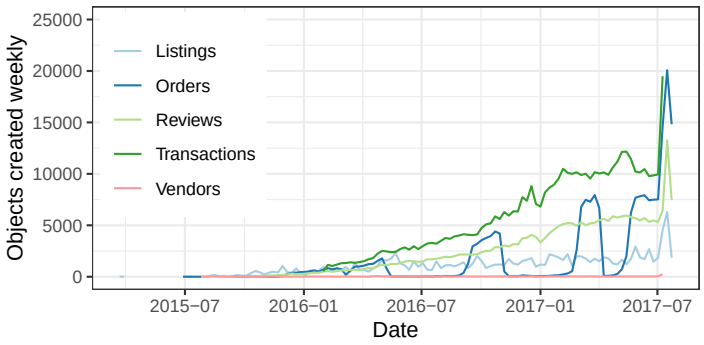


Figure 4.4: Weekly counts of objects from the Hansa back-end

4.5.3 Public View – Simulation

To derive insights on the impact of different scraping regimens and abundance estimation techniques on the quality of revenue estimation, we first generate (by simulation) fictitious marketplaces, that are similar to the Hansa marketplace [9], i.e., they feature similar objects and similar statistical parameters. As we discuss in Section 4.8.5, with the right choice of parameters, such simulations could reproduce other markets like Alphabay, Evolution, White House Market, Silk Road, etc. We then simulate different scraping routines on these markets. We begin with a formal description of the marketplace generation and scraping simulation processes, based on the model defined in Section 4.4. Figure 4.5 shows the entire process.

Following our model in Section 4.4.1, our simulation consists of four main *objects*: listings, vendors, reviews, and transactions. Additionally, we implement *operations* on each of these objects. Vendors and reviews can only be created, whereas listings can be created, deleted, set to hidden, or set to visible. Our simulations need five *inputs*: probability spaces, assignment functions, growth functions, a shaping function and a scraping function.

Probability space ❶ The probability space determines the sampling probability of each operation, e.g., probability a listing gets deleted, that a vendor is "created" (i.e., appears on the market), etc. Since the Hansa database provides final counts for the objects and operations we defined, we use this information to empirically define a probability space.

Assignment function ❷ Our objects have ordering and a set preference. For instance, vendors can exist in isolation, however listings must be created by a vendor, and reviews must belong to a listing. The way that each object is assigned to another is important in the context of hide and delete operations. For example, the distribu-

[9]The code used for simulations can be found at: https://github.com/aledcuevas/dnm-simulation

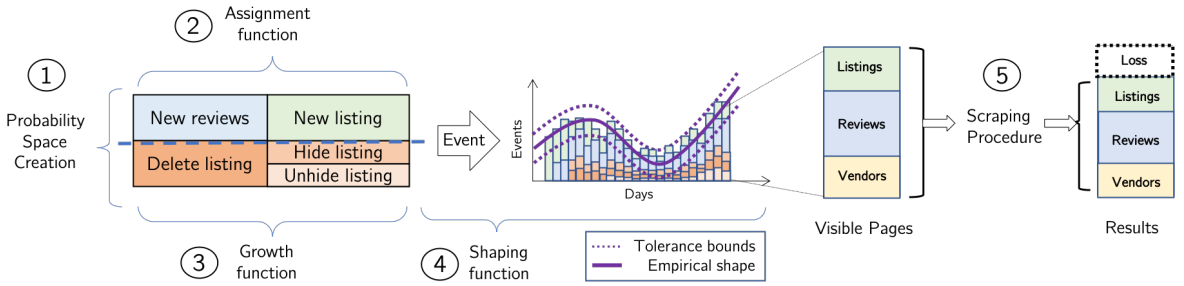


Figure 4.5: Steps involved in the generation and scraping of a simulated online marketplace.

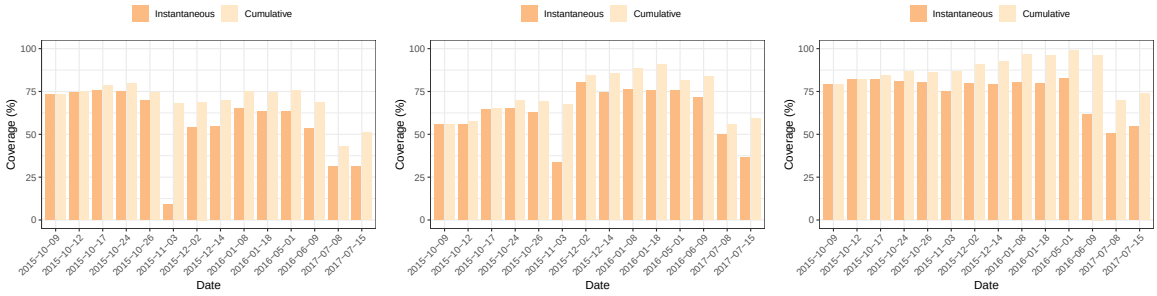
tion of reviews over listings can greatly impact a scraper's coverage, since the deletion of a listing with many reviews will cause a bigger loss if the scraper did not manage to capture this information before deletion. Thus, we define distributions for each assignment function (i.e., reviews to listings, listings to vendors).

Growth function ❸ Certain operation probabilities depend on the quantity of the objects in the market. For example, the probability of deleting a listing is zero when no listing exists. However, as the market grows and the number of listings increases, the probability of a delete operation will also increase. As such, we define a growth function which adapts our probability space as the quantity of objects increase.

Shaping function ❹ Once we have the probability of each operation, we need to add a time abstraction for the occurrence of events. For this, we employ a shaping function. Its purpose is to organize (or *shape*) the sequence of operations that take place over the lifetime of the marketplace simulation. Without a shaping function, each operation corresponds to a state transition from $\sigma_t \rightarrow \sigma_{t+1}$. Shaping allows the state transition to be over *epochs* corresponding to a number of operations.

Here, we define each epoch to represent a day. We allow a certain number of operations to take place before we proceed to the next epoch. So, we compute the moving average of the objects over the lifespan of the marketplace in days and summed them to derive an approximate shape for our events. Then, we define *tolerance bounds* around the average and allow the number of allowed events to be picked uniformly within the bounds. The tightness of the bounds determines the variability between each simulation. The simulation ends once either a certain number of operations have taken place or a certain number of epochs have passed. We also allow tolerance bounds around the allowed number of operations/days.

Scraping procedure ❺ Last, we define a simulated scraping procedure. Our basic scraping procedure is parameterized by the frequency at which scrapes are conducted, the number of requests the scraper is allowed to conduct, and an error probability characterizing the risk of failure of the request. The scrapes are instantaneous.



(a) Listings (b) Reviews (c) Vendors

Figure 4.6: Per-scrape instantaneous and cumulative coverage.

For each request, the scrape has access to the public view of the marketplace, that is, all public listings, vendors, and reviews; we call these *pages*. A page is scraped by uniformly drawing from the list of public pages. Each page retrieval counts towards the request cap, as well as a failed request.

Simulation setup. We summarize the marketplace simulation and describe the parameterization we used. The probability spaces determine the frequency at which operations take place. Because we do not know the precise ordering of operations in Hansa's transcript, we assume that the probability of a specific operation is equal to the number of times the operation occurred over the total number of operations. The assignment function determines how objects are assigned to their parent sets. We compute the empirical distributions of object assignments (e.g., distribution of reviews across listings) from the back-end to handle this sampling. Since we do not know how the conditional probabilities of operations evolve as the number of objects vary, we lack empirical data to parameterize our growth function. Instead, we assume that probabilities scale linearly. For our shaping functions, we allow the tolerance bounds to be within $\pm 25\%$. That is, in a given epoch, we allow a minimum of 75% and maximum of 125% operations over our empirical values. Lastly, we allow $\pm 1\%$ bounds for the number of operations/days. These bounds are much narrower since we have more precise information to parameterize them.

4.6 Coverage and Bias

We measure scraping coverage by comparing scrapes and back-end data in Section 4.6.1. We then measure differences between the *scraped* and *not-scraped* objects to to empirically uncover scraping bias in Section 4.6.2.

4.6.1 Scraping Coverage

We define coverage as the percentage of objects from a scrape that can be matched to the database on the scrape date. We measure listing coverage, review coverage and active vendor coverage.[10] For each scrape, we parsed and deduplicated the captured pages as explained in Section 4.5.1. This results in listing, review and vendor tables, which we use for our analysis. We compare these tables to the tables on listings, reviews and users directly derived from the Hansa back-end database (see Section 4.5.2). More precisely, we use the object creation dates to slice the Hansa back-end data into 42 sub-tables: one for each of the three object type (listings, reviews, vendors), on one of each of the 14 successful scrape dates. Each sub-table contains the relevant object data present in the market up to the corresponding scrape date.

We match listings between both datasets (back-end and scrapes) using the "listing ID," a numerical identifier present both in the database, and in the listing URL observable in the public view. We match vendors using vendor name. We match reviews using the tuple (review date, buyer name, vendor name, review message). For each of the 14 scrapes, we calculate the listing (L), review (R) and vendor (V) coverage using the following procedure. The input is an array T of 14 dates, the sets from a scrape (L_t^s, R_t^s, V_t^s) and the sets from a database slice $(L_t^{db}, R_t^{db}, V_t^{db})$. Then, for each $t \in T$, the coverage of that object type is calculated by taking the intersection between the scrape set and the database slice set, followed by calculating the percentage of the intersection to the database slice total size. For listings this is for example: $(L_t^s \cap L_t^{db}) / L_t^{db} \times 100$.

Figure 4.6 shows the coverage over time for each object. The mean coverage is 56.61% for listings, 62.66% for reviews and 74.71% for vendors. Looking at the scrape dates, there is a large time gap between the 12th scrape (2016/6/9) and the 13th scrape (2017/7/8). Unsurprisingly, the coverage of the last two scrapes is as a result much lower than the average of the first twelve scrapes. This can be explained by these scrapes not capturing all of the listings and reviews that have been created and then hidden or deleted for the public view in the time between scrapes. A different explanatory factor can be the increased size of Hansa, which grew from 28,700 listings and 20,100 reviews in mid-2016 to 112,800 listings and 233,600 reviews in mid-2017, making it more likely for a scrape in 2017 to be unable to capture all objects in one go.

In general, the vendor coverage is the highest type of coverage with almost 75% of all active vendors being captured on average by scrapes. Comparing the listing coverage and the review coverage over time, we observe the review coverage to be lower than the listing coverage for the first six scrapes. From December 2015 onward,

[10] We distinguish active vendors, as public views do not provide information on inactive vendors – i.e., those that have no listings on Hansa.

	Tests			Scraped listings $n = 61,248$				Not-scraped listings $n = 61,885$			
Variable	Test	Statistic	p-value	M	μ	σ	min-max	M	μ	σ	min-max
usdPrice	M-W U	1.6×10^9	0.00	30.00	390.18	2,739.08	0.01–3.2×10^5	66.48	625.50	6,508.99	0.01–1.0×10^6
views	M-W U	1.4×10^9	0.00	637.00	2820.82	12,569.63	0.00–270,251	232.50	1,536.93	5,438.36	0.00–251,554
numReviews	M-W U	1.8×10^9	≤ 0.001	0.00	2.90	18.71	0–1,313	0.00	1.30	11.47	0.00–2,114.00
ageListing	M-W U	1.7×10^9	≤ 0.001	239.00	267.64	206.86	5–728.00	207.00	224.96	149.12	1.00–855.00
isHidden	χ^2 test	5.9×10^3	0.00	0.00	0.05	0.22	0–1	0.00	0.20	0.40	0–1
isDeleted	χ^2 test	2.4×10^4	0.00	0.00	0.05	0.22	0–1	0.00	0.20	0.40	0–1
soldNoReview	χ^2 test	558.61	≤ 0.001	0.00	0.05	0.21	0–1	0.00	0.02	0.15	0–1
category	χ^2 test	9.0×10^3	0.00								

Table 4.3: Results of the Mann-Whitney U and χ^2 tests between scraped and not-scraped listings

however, the review coverage of each scrape is higher than its listing coverage. This could indicate that while a scrape captures less of the total inventory of listings, the listings it does capture are responsible for a larger proportion of all available reviews.

To give insights on how subsequent scrapes influence the cumulative coverage, Figure 4.6 shows the cumulative coverage when all scrapes are combined. While instantaneous scrape coverage does not improve, the increase in cumulative coverage shows that consecutive scrapes capture different objects. Thus, in most cases the combination of two consecutive scrapes leads to a higher cumulative coverage than the average of the two scrapes separately. The cumulative coverage of our scrapes for the market up to and including 2017/07/15 is 50.83% for listings, 59.49% for reviews and 73.93% for vendors. Hence, the empirical collection loss on Hansa is 49.17%, 40.51% and 26.07% for listings, reviews and vendors respectively. The average of these coverages weighted by their counts is 53.84%, meaning that on average just a bit more than half of all available objects was scraped.

4.6.2 Scraping Bias

We just showed that even after 14 scrapes, a non-negligible number of listings, reviews and vendors have still not been captured. From the back-end data, we also know that listings could be hidden and reviews deleted, making them disappear from the public view. In what way then is a scrape a truly random sample from the total population of available objects?

To answer this, we analyze the differences between scraped and not-scraped listings. Differences between scraped and not-scraped vendors and reviews come down to whether or not the corresponding listings are scraped. Indeed, comparing the characteristics of scraped and not-scraped vendors shows that 99.95% of the scraped vendors have a listing and 98.86% have a listing that is scraped. For reviews (given the

necessary pairing between a review and its listing) the percentages are even higher, with 100% of the scraped reviews having their paired listing scraped and 99.84% having the corresponding vendor scraped. This means that whether a review or vendor is scraped ultimately depends on whether the listing is scraped. This is because a review is scraped only when the vendor or the corresponding listing is scraped, and a vendor is scraped when A) it has a listing that B) is scraped. (See Tables B.6 and B.7 in the appendix for the descriptive statistics and tests for vendors and reviews.)

We next explore features that could be correlated with the chance of an object being scraped (e.g., the object being hidden) and features that can influence revenue calculations (e.g., the price of the object). To make sure we test features that have small inter-dependencies and thus capture different variations of why an object is not scraped, we performed an exploratory factor analysis on the listing features. As we did not discover any latent factors, we will not use the factors nor loadings themselves. The analysis and descriptive statistics of the factor analysis can be found in Appendix B.2. The subset of features then is `numReviews`, `ageListing`, `views`, `usdPrice`, `isDeleted`, `isHidden`, `category` and `soldNoReview`.

We performed Mann-Whitney U [221] and Chi-Square [222] tests between the scraped and not-scraped groups, to test for significant differences. The results in Table 4.3 show that *all* features differ significantly between the scraped and the not-scraped listings. Since not-scraped listings have less `views` and a lower number of reviews, `numReviews`, on average, this could point in the direction of a scrape being biased through "popularity". This is supported by a lower average `usdPrice`, as lower priced products are seen and sold more as they are more popular than higher priced listings. The features `ageListing`, `isHidden` and `isDeleted` influence the scraping process as we would expect: the longer a listing is available (and not hidden or deleted) on the market, the higher the probability the listing is scraped. The feature `soldNoReview` (i.e., the listing had sales, but no reviews) is relevant for a specific type of listing, namely *custom* listings [33]. Such listings sell a specific (larger) quantity and are created for a single buyer, who often does not leave a review.

Surprisingly, a larger percentage of the scraped than the not-scraped listings was bought without anyone leaving a review.

Finally, comparing the categories of scraped and not-scraped listings, we found that while on average $\approx$ 46% of a category is scraped, "Digital Goods" listings were scraped more often ($\approx$ 77%), while "Weed" listings were more often not scraped ($\approx$ 35%).

4.7 Revenue Calculations

We next compare projected revenues from our scraped data, to the actual revenues we can infer from the back-end database.

Projected revenue. Projecting market revenue from scraped data requires the use multiple proxies and heuristics. First, we detect and remove holding prices. Second, we pair reviews to listings, to approximate the actual price paid by the advertised price closest in time to when the review was left. Multiplying the number of reviews left every day by the listing prices gives us daily revenues in Bitcoin, which we convert to US Dollars using exchange rates from Coincap [223] for the corresponding dates. From there, we get the total revenue for a listing by summing these daily revenues over the lifespan of the listing; and the total projected revenue for the entire market, by summing the revenues for all listings.

Actual revenue. We next compute the *actual* market revenue from the Hansa back-end database. Because the transaction table only holds data from 2016/1/28 onward, we add revenue from order data for 2015/6/17–2016/1/27 to the revenue from transaction data for 2016/1/28–2017/7/20. For the revenue computation to be perfectly reliable, we would need the complete marketplace transcript; the Hansa back-end database, albeit very comprehensive, is not perfectly complete, as described earlier. However, based on the missing data percentages from Table 4.2 we assume that it is a very close approximation of ground truth data.

Loss. As discussed earlier, projecting revenue from scrapes produces two loss types: (i) an inference loss, due to using proxies and (ii) a collection loss, due to using data with incomplete coverage of reviews and listings. To estimate the size of the inference loss, we reproduce our projection calculations using, this time, data from the Hansa back-end database that would have been publicly available for scraping. In essence, this allows us to simulate what we would have gotten if we had "perfect scrapes" that captured all the information ever made publicly available by the market. Since we know, from Table 4.2, that review and listing data is 98.98% and 99.33% complete, respectively, the difference between our earlier projected revenue computation and this computation with perfect scrapes will approximate the inference loss well.

The total market revenue projected from scrapes is $13,149,373. When the revenue is calculated based on all the reviews available in the back-end database ("perfect scrapes"), this number rises to $27,385,346. The final number of total marketplace revenue for Hansa from transactions and orders is, however, $50,056,008. Shortly stated, inference loss causes a 50% drop, and collection loss seem to cause another 50% loss, resulting in a projected number that is only slightly more than a quarter of the actual market revenue.

Where does the loss come from? We next attempt to discover the causes for these

losses. We use one month – March 2017 – for this, since full order data is available for that month, Hansa had matured enough that, at that point, it was generating millions of revenue each month, but was not yet growing exponentially as it did later in 2017.

We calculate the revenue that month based on five different inputs: 1) the scraped reviews 2) the reviews from the database 3) the orders with the single quantity price 4) the orders with the item price 5) the orders with the full paid price (incl. shipping). The difference between 1) and 2) reflects the collection loss for this time period. The difference between 2) and 3) captures the inference loss from using reviews as a proxy for sales (orders), when not all customers leave reviews. The difference between 3) and 4) is the inference loss coming from assuming unit quantities for each inferred transaction. Finally the difference between 4) and 5) is the inference loss due to ignoring shipping costs.

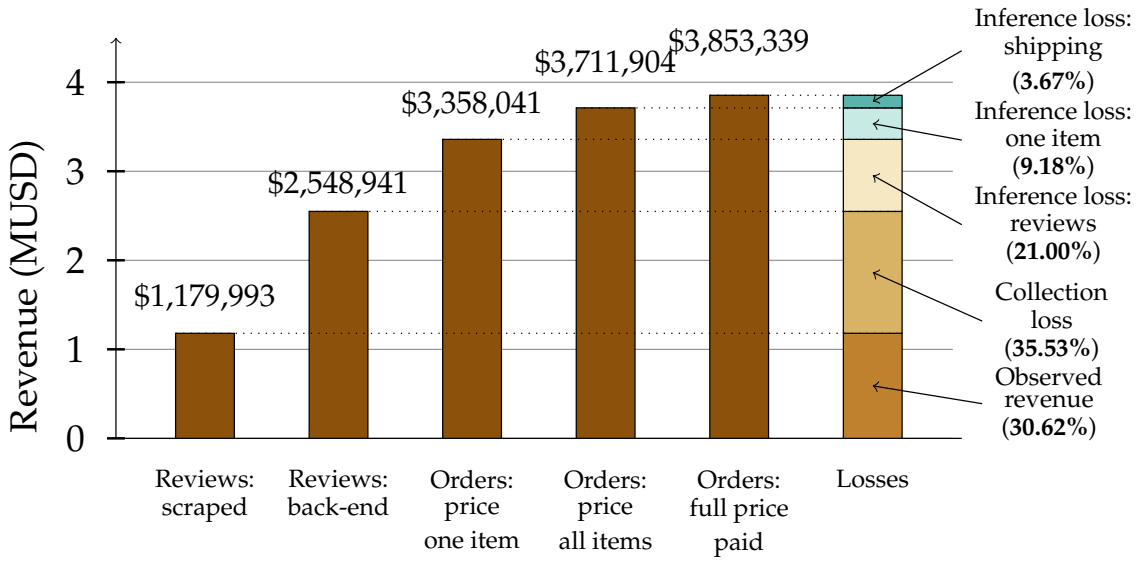


Figure 4.7: Calculation of Hansa's March 2017 revenue with different inputs

Figure 4.7 shows these revenue calculations based on different inputs. The gap between scraped reviews and reviews from the database is about a factor of two – $1,179,993 and $2,548,941 respectively. This collection loss of 53.71%, is in line with our findings in Section 4.6. The inference loss when using reviews as a proxy for sales is 21%, which translates to $809,101 in revenue. The difference between orders with the price for a singular quantity (3) and orders with an item price (4) is $353,863 (9.18%) in revenue, and the final difference between the orders with full price paid and orders with item price is just $141,435 (3.67%).

Take-aways. In short, achieving good scraping coverage is essential to get reliable estimates. Transactions without reviews present a major challenge. Without additional information from the market (e.g., the total number of sales for an item, as

displayed by Alphabay), it is impossible to infer whether the transaction occurred. The extent of this problem depends on the "social norms" of the market: the original Silk Road, for instance, reportedly strongly incentivized buyers to leave a review [33], whereas, evidently, compliance is a lot looser on Hansa. Finally, assuming away shipping costs and orders for multiple quantities of the same item seems to bear little impact on the projections.

4.8 Simulation

Through simulations (see Section 4.5.3) we explore marketplace coverage when varying the frequency, consistency, and rate-limiting of scrapes (Section 4.8.1 and Section 4.8.2). We present a comparison of abundance estimators in Section 4.8.3. Last, we propose and test a new, popularity-driven, scraper design.

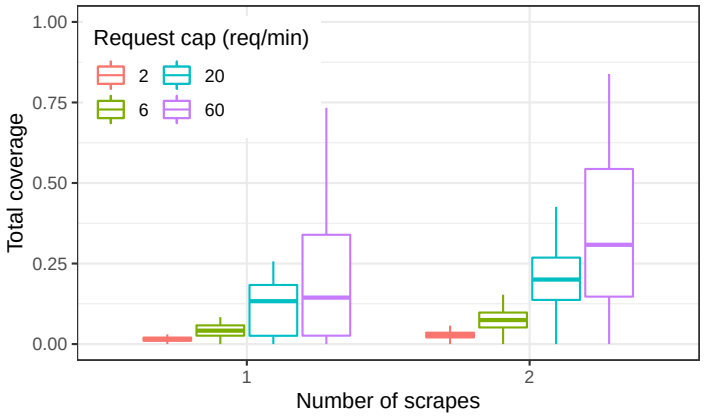


Figure 4.8: Distribution of coverage for one and two-shot scrapes simulated across different request limits.

4.8.1 Coverage of One and Two-shot Scrapes

We first quantify the coverage loss for our simulated marketplaces. Given that many studies rely on only one or two scrapes [194, 195, 209, 224], we compute the coverage distribution for both scenarios. First we simulate markets where only one scrape is available; we repeat this simulation for every single day the market is live. We then compute the expected coverage for each possible day in the simulation. Then, we simulate markets where two scrapes (taken on different days) are available. We run this simulation for every possible pair of days among the days the market was live. We then compute the expected coverage for all possible combinations of scrapes.

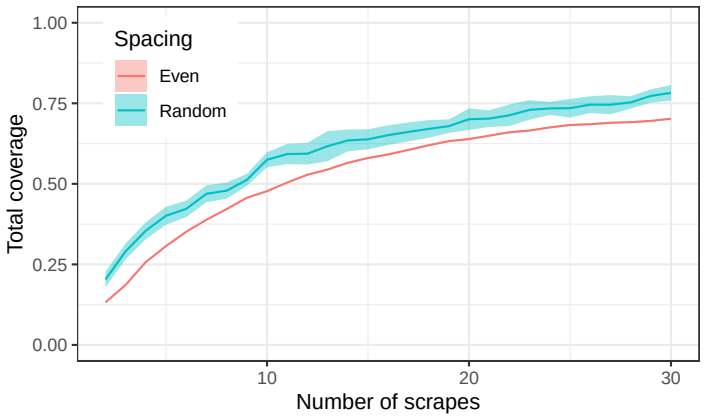


Figure 4.9: Scraping coverage as the number of scrapes increases, with evenly spaced scrapes and randomly spaced scrapes. The shaded area is the 95% confidence interval.

Further, we conduct these experiments with different page request limits: 2 req./min. (2,880 daily), 6 req./min (8,640 daily), 20 req./min (28,800 daily), 60 req./min (86,400 daily). In total, we simulated 2,800 scrapes for one-shot scrapes, and over 1,897,000 two-shot scrapes.

Figure 4.8 shows the results, using box plots with 95% confidence intervals. Even when scraping a page every second, the median coverage is low in the one-shot case (0.144) and only moderately better in the two-shot case (0.308). The theoretical maxima are 0.733 and 0.840 for the one and two-shot cases, respectively. However, in practice, 60 req./min. is rarely achievable due to the presence of anti-scraping mechanisms (e.g., CAPTCHAs, temporary bans, rate-limiting, etc.) [197].

4.8.2 Coverage and Scraping Consistency

We next seek to understand how coverage increases as the number of scrapes increase. Further, given that most past work we reviewed does not follow a consistent scraping schedule, we want to differentiate the impact on performance between consistent and inconsistent scraping routines. So, we compare the final coverage of all pages obtained between: 1) evenly spaced scrapes and 2) scrapes which are done at random intervals. For both settings, we calculate the coverage as we increase the number of simulated scrapes from 3 to 30. For each setting, we conduct simulations until our results converge into a narrow 95% confidence interval; this amounts to over 30,000 simulations.

Figure 4.9 shows that increasing the number of scrapes yields diminishing returns as the number of scrapes increases, mirroring Soska and Christin's findings [34]. We

find that not following a scraping routine is not necessarily detrimental to the coverage. However, it is important to caveat these results with the fact that the random scraping days were computed with *a priori* knowledge of the lifetime of the market. For continually growing markets (until takedown), such as Hansa, later scrapes have a greater chance of contributing more information to the final coverage. Thus, the more scrapes we have around periods of time growth, the better the coverage. On the other hand, if objects are frequently removed from the public view (e.g., deletions), then a consistent scraping routine might perform better since it has greater chance of catching data before the public view changes. In essence, we do not expect to see major differences in coverage between studies that did not follow a consistent scraping routine, as long as their scrapes are not concentrated in the early stages of the market.

4.8.3 Comparison of Abundance Estimators

We have evaluated scrape coverage using the ground truth contained in the backend data. In practice, however, public views do not always provide features to help us determine the size of the population for each object.[11] Instead, past work has relied on abundance estimators to calculate scraper coverage or collection loss. For instance, Soska and Christin used the Schnabel estimator [225] to estimate coverage [34]. Coverage estimations can then be used to extrapolate revenue, missing data, or adjust scraping regimens.

Abundance estimators, however, have not been evaluated in the context of online marketplaces. Thus, we proceed to evaluate the Schnabel estimator, along with the Lincoln-Petersen (LP) estimator, and the Jolly Seber (JS) estimator on our simulated marketplace. These estimators are part of a family of methods known as "mark and recapture," derived from tagging and recapturing experiments used to estimate wildlife populations [226]. A summary of these algorithms is given in Appendix B.3. At a high level, LP is the simplest estimator and assumes the population is constant, and estimated from two population samples; Schnabel extends LP to account for repeated sampling; Jolly-Seber extends these algorithms to a situation, like here, where the population changes over time.

We implemented each of the three estimators and used them in our simulation. We validated the LP and Schnabel estimators using the capture histories of northern pike data [227] in the R `FSAdata` package and the procedure described by Ogle [228]. For the JS estimator, we used the implementation provided by the `MARK` package, a well-known and widely used package for mark-and-recapture models [229].

[11] Most markets list the total number of items; some give the number of vendors; very few give the number of transactions per listing.

Algo. \ Coverage	Bi-Weekly Low	Bi-Weekly High	Monthly Low	Monthly High	Quarterly Low	Quarterly High
Jolly-Seber	0.501	0.081*	0.451	0.163*	0.401	0.338*
Lincoln-Petersen	0.219*	0.226	0.251*	0.249	0.358*	0.356
Schnabel	0.603	0.455	0.583	0.457	0.57	0.467

Table 4.4: Avg. error when estimating the number of listings across scraping intervals and using either a low request limit (2 req./min) or a high request limit (20 req./min).

We performed experiments in six different settings, varying the frequency and coverage of our scrapers. We tried three scraping frequencies: bi-weekly, monthly, and quarterly. We paired these with either a low request limit (2,880 requests per scrape; 2 req./min.) and a high request limit (28,880 requests per scrape; 20 req./min.). For each simulated scrape, we estimated the population of listings in the market based on prior captures and recaptures. We then computed the average collection loss for each scraper configuration across all our simulations. We repeated the simulations until we narrowed our 95% confidence interval; this took over 9,000 simulations.

Results. We summarize our results in Table 4.4. We observe that the JS estimator performs best in scenarios where our scrape has higher coverage. The JS estimator provides the best estimates when scraping frequently and with high coverage. However, the LP estimator performs better when coverage is poorer. This is because higher estimates are preferable when there is low coverage, and the LP estimator provides high estimates when there is low coverage. Surprisingly, the Schnabel estimator, which yielded good results in earlier work [34], performs here quite poorly across all settings.

4.8.4 Popularity-Driven Scraping

As explained in Section 4.6.2, certain pages are more critical to achieve good coverage than others. For instance, a listing page with a lot of reviews is more important to scrape properly than a listing with zero reviews. Previous work has hinted that, in terms of popularity, listings and vendors follow long-tailed distributions [34]. Thus, we hypothesize that one may achieve good coverage by primarily focusing on the most popular vendors and listings. While, ideally, one would want to scrape everything, it may not be possible: marketplaces have been deploying increasingly strict anti-scraping measures, which limit the ability of a third party to collect information [197]. We next explore whether "popularity-driven scraping" provides good coverage when facing a limited scraping "budget".

More precisely, we assume that we are given a limit ℓ on the number of requests

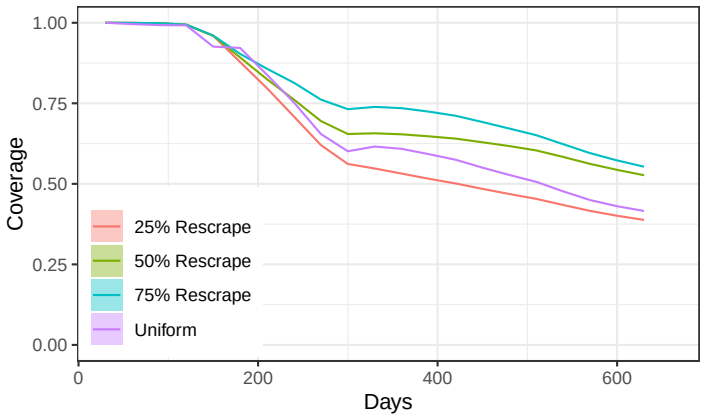


Figure 4.10: Average scrape coverage through a simulated market's lifetime for various popularity scraping budgets and compared to our uniformly random scraping baseline.

our scraper can issue (e.g., 2 requests per minute), and that we control the proportion ρ of previously seen pages we want to scrape again. We sort listings by popularity, i.e., by the number of reviews they have.[12] We rescrape the most popular listing pages until we hit $\rho\ell$ pages; we then scrape $(1-\rho)\ell$ pages we had not seen before.

We simulate three different parameters choices for ρ: 25%, 50%, and 75%, with $\ell = 6$ req./minute, over a 30-day interval. We conduct experiments until we sufficiently narrow our 95% confidence interval; here, this takes slightly over 20,000 simulations. We compare popularity-driven scrapers against our baseline, which is to scrape uniformly at random from the set of available pages. We present the mean coverage at each scrape date over all our simulated markets.

Results. Figure 4.10 shows that the scraper with $\rho = 75\%$, performs the best, with an average coverage of 0.765, followed by the scraper with a $\rho = 50\%$ rescraping budget (average coverage of 0.725). The baseline, random scraper, achieves a 0.674 coverage. Perhaps surprisingly, a scraper $\rho = 25\%$ budget performs *worse* than the baseline, with a 0.638 coverage. In short, a popularity-driven scraping approach can substantially increase coverage—as much as 10% higher than the baseline—as long as it is properly parameterized. Also, the difference in coverage widens as the market grows, which, in Hansa, was the case toward the end of the market's life.

4.8.5 Extrapolation

An optimal scraper for Hansa is contingent on a set of features that may not be shared by other markets. Hansa was a market with no established deletion policy, as op-

[12]For the first scrape, all listings are assumed to be equally popular.

posed to others. For instance, Dream Market deleted reviews older than 150 days [230]. Likewise, the recently deposed Russian-language Hydra Marketplace[13] purged reviews older than 240 days.

Thus, a scraper that follows a consistent routine would likely ensure more reliability and coverage. Hansa also experienced a burst of growth, following the Alphabay takedown, which occurred and lasted for a small period of time towards the end of the market. With constrained resources (i.e., limited number of scrapers and number of allowed requests), an optimal routine would have sporadically scraped Hansa during its slow period and aggressively scraped during its meteoric rise.

However, not only it may be hard to establish this routine *a priori*, but other markets follow different patterns, even following takedowns. For instance, Soska and Christin [34] show that older markets like Pandora or Agora had various bursts of revenue throughout their lifetime. These different growth patterns may call for different routines. Thus, when facing a new market, researchers may want to simulate different possible growth patterns and market lifetimes and choose the most robust strategy.

Lastly, a popularity-driven approach is an efficient choice for studies where we can infer where high-yield objects will be located (such as a revenue estimation study). For example, reviews on Hansa were largely concentrated among a handful of vendors, which is intuitive since listing popularity on anonymous markets has historically tended to follow Pareto-like distributions.

4.9 Discussion

This work brings up ethical considerations, especially as they relate to the use of seized data, which we discuss next. Second, while our results show that scraping as a measurement approach can introduce significant losses, we explain why this chapter should not be seen as an indictment of scraping—quite the contrary. Third, we discuss other contexts such as fora and other online shops. Last, from our observations, we derive a set of best practices for scraping online markets.

Value of scraping. While our results show that scraping can result in significant loss, ground-truth data is rarely, if ever, available. Seized back-ends are rare—and may be very far from complete when they exist. We discovered that Hansa's database holds many features unavailable in the public views. However, a major drawback is that this database only contains a *single* record for each object. Absent any back-up (which were available here, due to the Hansa administrators espousing questionable data retention practices), one would only be able to see the *latest* version of each ob-

[13] No relationship to an older Hydra market active in 2014–2015[34].

ject. On the other hand, a consistent weekly scraping regime could have captured 108 versions of each object in Hansa's lifetime. Doing so allows to understand historical price developments, vendor PGP-keys changes, and vendor geographic shipping information – all important data points for revenue analysis and vendor matching [202].

Other contexts. The issues of incomplete data and the usage of proxies and heuristics for (revenue) calculations are not limited to the domain of online anonymous marketplaces. Other marketplace contexts, such as online fora (e.g. hacker fora) or specific web shops (e.g. pharmaceutical websites), also face the challenge of doing empirical marketplace research in adversarial contexts. This has two consequences.

First, online anonymous marketplace research can learn from approaches on these other types of marketplaces. Different internal and external validation techniques from other works could also be applied. Two notable examples are calculating completeness of a scrape through leveraging unique marketplace identifiers (e.g., changing URLs or sales counters [186]) and cross-referencing tables and checking concordances between transactional data and metadata [215].

Second, the present study can serve as a model for these other contexts. As Portnoff et al. [231] note in their analysis of an underground forum marketplace: "an analysis relying on both private and public data vs. just public may reach different conclusions about the revenue of a market." More broadly, Andreas and Greenhill in "Sex, Drugs, and Body Counts" show how scientific measurement errors often motivate inappropriate policy choices [232]. As all these types of fora or unlicensed shops primarily deal in illegal offerings, precision is of the utmost importance.

Best practices. Our findings can inform future online anonymous markets measurement studies, both for study design and for reporting results. First, we recommend *frequent* and *periodic* scraping to mitigate the impact of scraping errors, rate limits, and data deletion. When describing data collection, studies should disclose when the scrapes were obtained and the number of requests that were sent. To contextualize the potential coverage of their scrapes, studies should try to estimate the size (i.e., pages) of the site. While abundance estimation can help, markets may offer metadata that provide a better starting point for estimation. For instance, markets may disclose the number of vendors, items, or even the number of orders that each vendor has fulfilled. These results can then be complemented with estimates derived from Jolly-Seber or Lincoln-Petersen models for high and low coverage assumptions, respectively.

In the face of limited scraping budgets (e.g., as caused by anti-scraping mechanisms), future studies should consider identifying and focusing their scraping on high-yield portions of the website, rather than scraping in a breadth-first fashion. Rate of growth can be measured through observed changes in subsequent scrapes (as

we described in Section 4.8.4) or through metadata (e.g., a leaderboard of reputable vendors). Further, we also recommend that future studies provide more detail on their scraper *design*. We found that scraper design is often either not discussed or described with insufficient detail in the literature (Section 4.2.2). Yet, understanding how the scraper traverses pages, the number of requests it performs, or how it adapts to adversarial scraping environments are all important details that help contextualize the coverage of the measurements, and subsequently its impact on estimation.

Last, our research showed that the "measurement-by-proxy" approach provides a *very* conservative lower bound for revenue estimations on online anonymous marketplaces. If the assumption of similar review-to-transaction ratios holds for a newer marketplace (e.g., feedback is neither mandatory nor automatically purged over time), our loss factors from Section 4.7 can help calculate an upper bound for revenue projections. That way, future research can take the biases we discovered into account and reason about the impact of calculating revenue based on scraped data on measurement outcomes.

4.10 Ethics

Our data and measurements share similar ethical considerations as previous work on the external measurement of websites which may be engaging in illicit activities [33, 34], as well as in the usage of seized back-end data [199, 233, 234]. For our scraping measurements, we followed Martin and Christin's recommendations [235], and took proactive steps to minimize direct and indirect consequences that our measurements may have had on marketplace participants and on Tor users. (For instance, we purposefully limited our scraping regimen, did not interact with marketplace actors, etc.) Our lightweight, non-intrusive scraping approach had a twofold intent: to avoid alerting the operators and to reduce the burden on the network. The former point is particularly important, both from a scientific perspective (to prevent effecting a change on that which is being measured), as well as from an ethical perspective (to not induce any potentially harmful effects in the ecosystem). In terms of the back-end data, this chapter focused on validating external measurements of underground marketplaces. As such, we focus our analyses on the marketplace itself, rather than individual users. Similar to earlier work [199, 233], all of our analyses of the back-end data were conducted on-site at Dutch law enforcement agencies, and the data was stored and protected under their safety and security guidelines. The data was made accessible to us for academic research purposes. Extracting aggregate data points for our tables and figures was done under strict supervision through one specific monitored channel. A Dutch law enforcement privacy-officer vetted that the data contains no personally identifiable information.

As we obtained the approval of the Dutch Public Prosecution Service for our analysis, the Delft IRB viewed this work as outside their jurisdiction and were satisfied with this assessment. The three authors at US institutions did not directly interact with back-end data. The Carnegie Mellon IRB had earlier opined, and confirmed, that scraping marketplace data (without personal identifiers) did not constitute human-subject research.

Most importantly, this study does not, and does not seek to, provide any legal proof of criminal conduct.

Chapter 5

Evaluating Reputation Systems in Online Anonymous Marketplaces

---

This chapter is adapted from our paper:

> [38] Alejandro Cuevas and Nicolas Christin. "Does Online Anonymous Market Vendor Reputation Matter". In Proceedings of the 33rd USENIX Security Symposium (USENIX '24), August, 2024. USENIX Association.

5.1 Motivation and Goals

This chapter evaluates the predictive power of profile signals in online anonymous marketplaces (OAMs), focusing on whether reputation systems and other vendor-visible signals can explain or forecast financial outcomes. Building on the revenue estimation methodology validated in Chapter 4, we model vendor success and longevity using data from eight darknet marketplaces and associated forums, spanning over a decade of activity.

On markets where "seller anonymity is guaranteed, and no legal recourse exists against scammers, one would expect a certain amount of deception." [33] Yet, the market capitalization of online anonymous marketplaces (OAMs) has massively grown since their inception in 2011, with individual vendors in these platforms that operate multi-million dollar operations [34, 178]. This alone seems to indicate that the reputation and feedback systems in place in these marketplaces are overall working as expected.

However, scam stories abound in underground forums. Goods that do not match their description, dangerously adulterated drugs, and unfulfilled orders are among

the most common complaints. So, which is it? Do these marketplaces provide enough signals for buyers to distinguish between high and low quality vendors? Or do buyers have to resort to other signals to make this determination?

Answering these questions is especially important in the context of underground markets, where hazardous substances (e.g., narcotics) are often being sold, with potential deadly consequences for buyers [32]. A key argument in defense of these markets is that, by enabling buyers to avoid dangerous vendors and/or products, reputation systems help with harm reduction compared to alternatives (e.g., street sales). However, this claim assumes that these reputation systems provide a useful signal.

Surprisingly, despite substantial research demonstrating the importance of reputation in driving sales in traditional online marketplaces [1, 236], there has been significantly less exploration of what drives success in OAMs. While prior work has found some correlations between market or forum-derived features and performance [36, 237, 238], they have only studied narrower contexts: carding forums [36, 237], or B2B cybercrime vendors in a single market [238]. Furthermore, despite ample evidence that buyers use Reddit-like forums to provide additional vendor reviews, no prior work studies the link between forum-derived features and success in OAMs. Last, despite prior work examining listing and vendor longevity in OAMs [33, 34], no prior work tests *which* factors impact survivability of vendors in these markets.

We fill these gaps by exploring the predictive power of various signals on out-of-sample predictions of financial success and longevity of a vendor from a OAM. We 1) use multivariate survivability models to test the role of various covariates on the disappearance of a vendor, and 2) use explainable machine learning models to predict the disappearance and wealth tier that a vendor will belong to in a future state of the market. We conduct our experiments on eight OAMs and two types of forums, with activity spanning from 2011 to 2023.

We argue that long-term vendor success, as determined by accrued wealth and permanence in the market, is a good proxy for the vendor selling acceptable products. On the other hand, sales cessation and vendor exit are good proxies for vendor failure. As such, the ability to predict success or failure likely helps explain the risk associated with a specific vendor. Moreover, the success of our models without extensive parameter tuning or feature engineering highlights the robustness of the approach and its potential for deployment in low-resource monitoring environments.

We offer the following contributions:

- We quantify the impact of various market and forum-derived features on vendor longevity and find that feedback scores (including imported product reviews from other markets) have a significant impact on increasing longevity across most markets we study;

- We find that (both positive and negative) reputation signals from forums explain vendor survivability, but overall have little predictive power for vendor success;
- We demonstrate we can build a *generalizable* model to predict, more accurately than raw feedback, which vendors may leave the market in the short-term (1–3 months);
- We find that future financial success is predictable, particularly for the top/bottom 25% of vendors, and even *on previously unseen markets*.
- We find that features external to the market, and time-series representations of features not only fail to increase the predictive power, but instead often *decrease* it.

Our results show that reputation signals play a role in shaping vendor trajectories. In proportional hazards models, higher feedback scores are associated with lower dropout risk, suggesting that reputation systems do help push out low-quality vendors over time. However, when applied to predict vendor disappearance or success in the short term, these signals perform poorly. Instead, our models that integrate a broader set of features——such as past sales volume, vendor age, and activity trends—outperform raw feedback scores in both vendor longevity and revenue prediction tasks. These models generalize across different markets and timeframes, identifying high-risk vendors and future top earners with notable accuracy, even in unseen environments.

Interestingly, we find that forum-derived reputation signals—such as vendor mentions or reputation within Reddit-like underground forums—add little to predictive performance. Manual inspection reveals that forum interactions are often noisy, performative, or inconsistent across vendor size. This finding underscores a recurring theme in the thesis: the visibility or prominence of a signal does not necessarily reflect its informativeness or reliability.

Our results have several implications. Our models can be used by law enforcement agencies for early identification of important vendors on emerging markets. In particular, by achieving high predictive accuracy even when applied to a new, previously unseen market, our models can make monitoring and intervention efforts targeting online criminal ecosystems more efficient. Furthermore, our results shed light on the viability of strategies that involve "poisoning" the reputation of vendors inside OAMs and across forums. Our results also empirically validate the role of reputation systems in OAMs. On one hand, we find evidence that a functioning feedback system may help online marketplaces reduce harm for drug consumers—and that it can be improved by looking at other signals. On the other hand, we find that discourse and reputation signals from external forums may not be as useful to identify

bad actors.

By demonstrating that conventional reputation systems are imperfect but improvable, this chapter reinforces a core argument of the thesis: profile signals must be evaluated not only by their theoretical value but also by their empirical performance. Our findings show that richer, multi-feature models can surface higher-fidelity representations of vendor quality.

This chapter also motivates the next. While vendor disappearance is a useful proxy for failure, it is ambiguous: vendors may exit due to fraud, personal circumstances, or platform bans. To better understand how profile signals relate to misbehavior, Chapter 6 shifts focus to account suspension as a more direct outcome. There, we apply the same modeling approach to cryptocurrency peer-to-peer marketplaces, assessing which signals predict platform-enforced penalties.

5.2 Background and Related Work

We next provide an overview of reputation systems in the broad context of general online commerce, before focusing on idiosyncracies of anonymous markets; and discuss measurement and inferencing work on OAMs and forums.

5.2.1 Reputation & Feedback Systems

Online marketplaces initially faced significant skepticism, particularly from economic theorists. The asymmetric information between buyers and sellers, as well as the lack of incentive from one party to guard against risk, can indeed drive markets to failure [1]. Traditional online marketplaces (e.g., eBay) overcame these challenges by employing reputation systems. Reputation systems are essential in creating trust, particularly in two-sided marketplaces (i.e., markets that serve as platforms to connect independent buyers with independent vendors, such as eBay). The promise that good (resp. bad) behavior in the present may be rewarded (resp. penalized) in the future by increased (resp. decreased) sales is how reputation systems incentivize buyers to act in good faith. There is substantial economic literature in conventional markets that empirically demonstrate how vendors with better reputation attract more buyers and higher prices, while the converse holds true for disreputable vendors [236].

OAMs face similar challenges as conventional online marketplaces, but with some particularities. First, dispute resolution is less robust. None of the parties (particularly buyers) have any legal recourse when facing a scam. Second, vendors often have access to buyers' private information (e.g., shipping address) and can leak this information in retaliation [239]. Third, illicit goods—particularly, narcotics—typically

have high price and quality dispersion [240]. This quality uncertainty is exacerbated by the lack of incentive for buyers to guard buyers against risk (moral hazard).

Nonetheless, OAMs have persisted and thrived, which indicates that they have managed to create systems of trust between buyers and sellers. Platforms offer a variety of features to create trust, including escrow, discussion forums, feedback scores, automated reviews, and various signaling mechanisms such as badges [241, 242]. In two surveys, OAM buyers reported that the existence of reputation systems fostered their engagement [243, 244]. Yet, it is unclear which specific signals are most important in creating trust and drive vendor success.

5.2.2 Performance in Criminal Markets

Prior work [36, 237, 238, 245] has attempted to measure and explain the factors that drive success in criminal markets, particularly in OAMs and sales-driven criminal forums.[1] While OAMs and criminal forums offer slightly different transaction experiences for buyers, they share many similarities and have been broadly studied using similar theories. For instance, researchers have analyzed vendor signals through Gambetta's signaling theory [246] to identify and explain buyer preference in carding forums [36, 237], while van Wegberg et al. applied it to explain B2B vendor performance in OAMs [199]. Several papers have attempted to characterize vendor trajectories in OAMs [199, 245], or have studied conversations and actors in forums to identify "key players" [247–250]. Similarly, others have found links between observable features (e.g., vendor position in their social network) and private features (e.g., amount of private messages received) [251–253].

Ultimately, this body of research attempts to identify which vendors will become successful directly (e.g., sales volume when feedback can be used as a proxy) or indirectly (e.g., number of private messages when sales proxies are elusive). Unfortunately, the results have not yielded a clear picture of what drives financial success. Van Wegberg et al. posited that who the vendor is matters more than product differentiators [238]. Holt et al. found that signals like badges in forums seemed to drive more feedback [237]. Décary-Hétu et al. found correlations between vendors' sales and their network features but not with their forum features [36]. Furthermore, even though buyers have long used forums to review OAM vendors [254–257], the literature shows a gap on how reputation and/or influence signals from forums affect OAM vendor success. Despite prior work modeling the survivability of vendors and listings [33, 34], factors that accelerate the disappearance of vendors and listings re-

[1] We distinguish between sales-driven criminal forums whose primary intent is to connect buyers and vendors in private transactions, and forums that serve a complementary role to OAMs, e.g., to discuss vendor experiences.

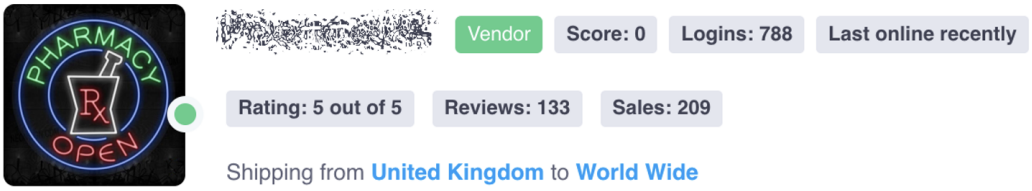


Figure 5.1: Profile page of a vendor in the Nemesis marketplace.

main unknown. Last, Bradley explored the resiliency of the OAM ecosystem, as well as that of vendors within it. Closest to our work, they observed how reputational damage may reduce vendor capacity to trade [214]. They also employed a qualitative approach on forum data to assess the impact of law enforcement operations [214]. We use similar techniques but apply them toward vendor financial success.

5.3 Methodology

We next describe how we obtained and processed data from the markets and forums we analyze, how we extract the features our analysis uses, and discuss data validation.

5.3.1 Data Sources

Marketplaces: High-confidence inference from web scrapes requires robust processing and validation strategies [37]. Hence, we use peer-reviewed and validated datasets when possible. For the Silk Road, Pandora, Silk Road 2.0, Agora, and Evolution markets, we use the Soska and Christin [34] dataset; for Hansa Market, the Cuevas et al. [37] dataset; for Alphabay, the van Wegberg et al. [178] dataset. In addition, we collected and processed a market active at the time of writing, Nemesis, along with its internal forum. Figure 5.2 shows the revenue of all markets (scaled to be on the same time axis). Figure 5.1 shows a user profile page in the Nemesis marketplace, which is similar to other online anonymous marketplaces. The profile page shows the vendor's reputation score, reviews, sales, among other information.
Subreddits: Many OAMs used Reddit as a discussion platform until they got banned in 2018 [258].[2] We collected and processed data from the subreddit

[2]After the 2018 ban, a Reddit alternative, Dread, emerged, but it does not feature data relevant to the markets we study—in particular, Nemesis discourse is all but banned on Dread due to a feud between administrators.

Table 5.1: Overview of collected and processed marketplace data.

Marketplace	#Vendors	#Feedbacks	Est. Revenue	First Seen	Last Seen	Activity Length	#Snapshots
Silk Road	2,336	605,744	$62,334,431	2011-11-27	2013-08-19	631 days	133
Pandora	459	89,065	$12,239,165	2013-11-02	2014-10-13	345 days	140
Silk Road 2.0	1,202	687,375	$121,529,265	2013-11-27	2014-10-29	336 days	195
Agora	1,961	234,272	$40,857,567	2013-12-24	2015-02-11	414 days	161
Evolution	2,352	464,146	$43,993,997	2014-01-13	2015-02-18	401 days	43
Alphabay	6,101	1,736,127	$218,971,605	2014-12-31	2017-05-26	877 days	33
Hansa	1,309	153,400	$13,149,373	2015-08-21	2017-07-15	694 days	14
Nemesis	372	18,794	$6,388,411	2022-03-09	2023-01-31	328 days	10

`/r/HansaDarknet-Market` which contains 264 posts, 3,613 comments, from September 2015 to September 2017. This subreddit was used to discuss matters related to the Hansa marketplace (e.g., news, policy updates), by vendors to advertise products, and by buyers to describe their experiences with vendors. We also collected and processed `/r/DarkNetMarkets`, with 125,300 posts, and 1,850,533 comments, ranging from October 2013 to September 2017. Similar to `/r/HansaDarknet`, this subreddit discussed vendor quality across a variety of markets, among other topics.

Nemesis Forum: The Nemesis forum similarly employs a Reddit-style interface, with various sub-forums such as `/n/AskNemesis` for platform questions and `/n/Cocaine`, for discussions related to cocaine vendors. Creating a Nemesis marketplace account also creates a Nemesis forum account, so that marketplace and forum handles are identical (for both buyers and sellers). We collected 4,018 posts and 12,710 comments from March 2022 to February 2023.

5.3.2 Data Processing and Validation

We scraped Nemesis from November 18th, 2022 to February 1st, 2023 at a rate of about 32 pages/min (or roughly 46,000 pages per day). We employed a Scrapy-based breadth-first scraper.[3] Similar to previous work, we attempted to proxy sales by matching feedback to item listings. Given that we began scraping the market relatively early in its development, we were able to match over 99% of collected feedback to listings. This is facilitated by Nemesis' design: feedback left on vendor pages links to the item page featuring the review. However, Nemesis presents a unique challenge: some individual item listings feature various quantity options (e.g., a listing "High-quality Cocaine" may offer "1g at $10," "15g at $125," and "1kg at $5,000"). The most conservative approach would be to assume that each sale is for the lowest priced option, giving us a lower bound on sales, but potentially vastly underestimating the

[3]Due to parallelization across multiple scraping agents, the breadth-first order is not always respected in practice.

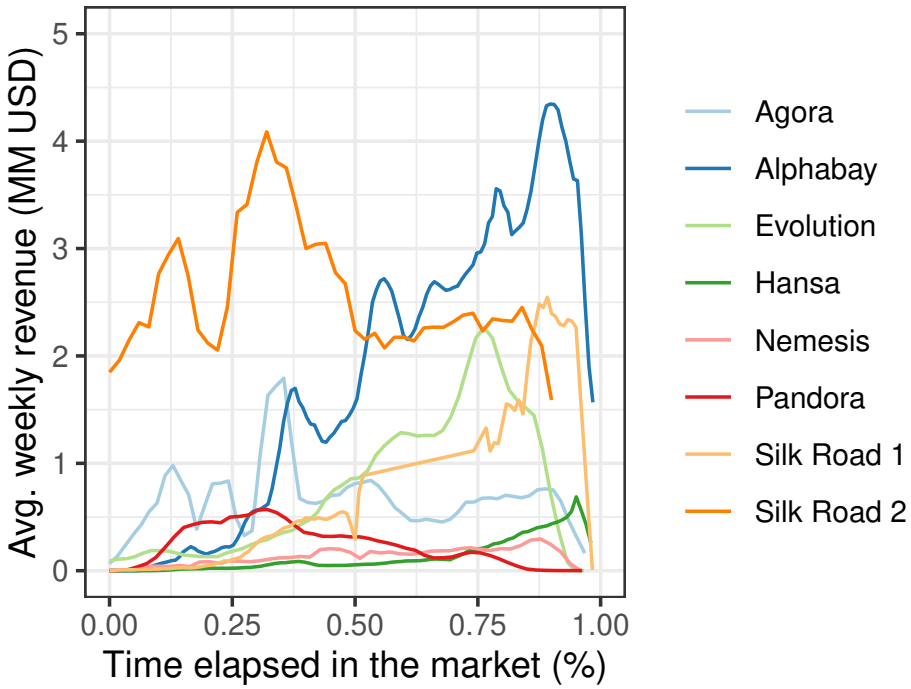


Figure 5.2: Revenue over time for all markets scaled to the same time axis. Each point is a four-week rolling window average.

sales. Instead, we experiment with taking the mean and the median price when a list of options is provided. While Nemesis vendors do not seem to use "holding prices," i.e., abnormally high prices signifying a lack of stock, we applied the same heuristic as Soska and Christin to filter these out, such that our data is consistent with theirs [34].

Furthermore, to validate our processing and inference, both authors independently parsed the raw HTML scrapes and estimated the revenue for each vendor. Our estimates of revenue using the minimum listing price are within 6% of each other. We did not find public analyses of Nemesis to which we could compare our revenue against as Cuevas et al. did for the Hansa market [37].

5.3.3 Extracting Features

Our goal is to understand the impact of reputation systems on the financial success or disappearance of vendors across markets. Over time, markets have displayed a variety of attributes and badges for vendors, such as measures of "level," "score," "experience," and/or whether a vendor has undergone some measure of verification. Markets have also employed a variety of feedback scales (e.g., 1–5 stars, positive/neutral/negative, etc.). Furthermore, vendors have also utilized various differentiators, such as alternate media of communication (e.g., Telegram, ICQ, etc.), describing terms

of service, refund policies, and avenues for customer support. Prior work has found that some platform-specific attributes may be used to explain sales performance (e.g., customer support), for a type of goods (i.e., cybercrime-related), within a specific market (Alphabay) [178].

We hypothesize that we can use generalizable features or attributes to explain and predict the performance of vendors across OAMs. We focus on features common across markets, using the basic objects that support these markets: feedback/reviews, listings/items, and vendors [37]. We also explore the impact of capturing time variations across these features, as the market evolves. Lastly, we investigate the impact of forum-based features.

5.3.4 Ethics of Data Collection and Release

Our Institutional Review Board (IRB) deemed our study not to be human-subject research. Nonetheless, our work still has important ethical implications. The collection and release of our data follows the principles outlined by Martin and Christin [235], and the same approach as previous OAM research, especially Soska and Christin [34] and Cuevas et al. [37]. Whenever possible, we prioritized the use of existing peer-reviewed datasets for replicability and to abide by the same ethical considerations as prior work. However, we also collected data from a new marketplace, Nemesis market. For this, we balanced data accuracy with stealth (to avoid impacting the studied ecosystem) and low impact on the Tor network (using a light-weight crawler). We also contribute fast Tor relays with long uptime to compensate for our use.

Marketplace and forum data contain discussion of potentially illicit activities; and forum data may inadvertently leak information about buyers and/or sellers. After consulting our IRB and general counsel, an unlimited public release is undesirable. However, we can follow the lead of other researchers. Indeed, data for seven of the eight marketplaces we study are already publicly available [34, 37] *upon request*, for non-commercial use, using the IMPACT portal.[4] This allows researchers to vet possible uses of the data before releasing it. We will adopt the same strategy for our own (Nemesis) data.

Last, we also relied on Reddit data, which was publicly available through Pushshift [259] at the time of writing. However, since then, Reddit updated its API policies which has affected the availability of these data through Pushshift [260]. Instead, these data may be accessed through independently hosted torrents [261].

[4] https://www.impactcybertrust.org.

Base Features

As a starting point, we define and extract features common across our markets. Our initial set of feature categories are:

- **Revenue features**: mean/median order value, cumulative revenue, time of first sale.
- **Feedback features**: count, average count per week, and feedback score.
- **Listing features**: item diversity, within-category price z-score, time of first listing, main category of goods sold.

Temporal Features

We can make some of the above features more expressive by adding a time dimension. Using just the base features defined above, yields a matrix of shape $(nr\_vendors \times nr\_features)$ up to a time T in the market. However, we could also build time series for several of the above features by tracking their evolution over time. That is, we break T into a series of time steps t_i, resulting in a matrix of shape $(nr\_vendors \times nr\_timesteps \times nr\_features)$. For example, for period of length T, rather than just having the total revenue up to time T, we instead consider the revenue per time step (e.g., week) t_i up to T.

Forum Features

Forums provide a platform for customers to discuss experiences with vendors, or suspicions that a vendor has been compromised by law enforcement [254, 255]. Forums also provide signals on vendor notoriety (e.g., if a vendor's posts garner a lot of attention) or influence (e.g., by looking at their interaction network size). As such, forum signals may help predict vendor success and longevity.

For Hansa, we consider `/r/HansaDarknet-Market` and `/r/DarkNetMarkets`, similar to prior work [214]. Posts we consider in `/r/DarkNetMarkets` refer to vendors who existed in Hansa, but may not always directly relate to a sale taking place on Hansa. Further, there is no definitive way of mapping users from Reddit to the Hansa marketplace. For this reason, we do not attempt to build interaction networks between users. Instead, we only extract comment "sentiment" (good, neutral, and bad) about vendors. We first try automated methods: named-entity recognition for vendor discovery, and sentiment analysis. However, pilot testing showed that these methods perform poorly in these forums, as described below.[5] Instead, we opt for a manual analysis process.

[5]Whether typical sentiment analysis packages could be made to work, using specialized training sets, is an open question, that is likely to be answered in the affirmative. However, performing such

We first use a fuzzy matching search for vendor names across posts to find a set of candidate posts and comments that may refer to a given vendor. We find 2,294 posts and a total of 13,002 comments under these posts. One coder independently goes through the posts and comments and 1) confirms that the match was appropriate, and 2) determine the sentiment of the comment or post given to a vendor. For validation, we randomly select 10% of the posts and comments and have a second coder qualify these posts. That way, we also ensured that the first coder was not missing entries. The coders then compared their results and derived a Cohen's Kappa of $\kappa = 0.58$, moderate agreement. The disagreements mainly stemmed from three types of issues:

- **Unclear interpretation (e.g., conflicting sentiment)**. "[REDACTED]'s bud would actually look reallly nice if his buds werent so compressed."
- **Unclear attribution (e.g., acronym mapping)**. "I remember KK having issues with oily batches." posts anymore?"
- **Lack of understanding in lingo**. "50% FE [from this vendor]seems tarded to me. Anyone else?"

While a moderate agreement is not ideal, the examples above illustrate the difficulty of both attributing and interpreting signals in fora, even when done manually. Unsurprisingly, our automated efforts to extract entities (through named-entity recognition models) and to extract sentiment (by leveraging sentiment analysis models) failed to provide useful results. To mitigate the sources of disagreement, we introduced a "neutral" code for comments, rather than just "positive" and "negative." However, we chose to exclude "neutral" mentions as we found them to be of little use (e.g., "Please give me one example of shilling for [REDACTED]" conveys little signal). Furthermore, when the attribution was not clear, we decided to omit the comment. Using these guidelines, the first coder coded the rest of the dataset, and we focused on comments that had clearer signals, such as "I love that Yoda out of all those strains!! That Skywalker from [REDACTED] is fire as well!!" In total we found 843 positive (677 unique vendors) and 263 (210 unique vendors) negative comments.

For Nemesis we can derive social networks of vendors and buyers given that the forum and marketplace aliases are the same. We can see a vendor listings, as well as their posts and comments. Thus, we create a directed interaction network for comments, whereby an edge is formed when a comment is left as a reply to a comment/post. We also quantify the number of posts and comments made by each vendor, as well as the up-votes they receive. Due to the large number of interactions between buyers and vendors on the forum, we did not attempt to manually code the sentiment of these interactions.

retraining would have required a labeled OAM forum dataset in the first place, which was not available.

Table 5.2: Cox Proportional Hazards Regression across all 8 markets, where exp(c) indicates the hazard rate increase per unit increment. The regression was stratified based on the wealth quartile the vendors belonged to at the end of the market.

	Silk Road 1				**Pandora**				**Silk Road 2**				**Agora**			
Covariates	$\exp(c)$	SE	z	p	$\exp(c)$	SE	z	p	$\exp(c)$	SE	z	p	$\exp(c)$	SE	z	p
Avg. Feedback Value	0.50	0.16	-4.25	**<.005**	0.69	0.24	-1.53	0.13	0.51	0.10	-6.34	**<.005**	0.61	0.07	-7.48	**<.005**
Presence in Other Mkt.	–	–	–	–	0.85	0.14	-1.16	0.24	0.59	0.09	-5.95	**<.005**	0.67	0.07	-5.73	**<.005**
Mainly Digital	0.89	0.19	-0.62	0.53	0.78	0.37	-0.69	0.49	0.56	0.26	-2.24	**0.02**	0.80	0.21	0.29	0.29
Mainly Category A Drugs	1.12	0.18	0.61	0.54	1.10	0.34	0.29	0.77	0.87	0.23	-0.59	0.55	1.31	0.19	1.42	0.16
Mainly Category B Drugs	1.30	0.17	1.50	0.13	0.91	0.34	-0.27	0.78	0.78	0.24	-1.06	0.29	1.10	0.20	0.50	0.62
	Evolution				**Alphabay**				**Hansa**				**Nemesis**			
Covariates	$\exp(c)$	SE	z	p	$\exp(c)$	SE	z	p	$\exp(c)$	SE	z	p	$\exp(c)$	SE	z	p
Avg. Feedback Value	0.58	0.12	-4.44	**<.005**	0.67	0.05	-7.91	**<.005**	0.50	0.19	-3.59	**<.005**	0.36	0.18	-5.48	**<.005**
Presence in Other Mkt.	0.59	0.07	-7.26	**<.005**	0.68	0.05	-7.29	**<.005**	0.60	0.14	-3.77	**<.005**	1.32	0.31	0.89	0.37
Mainly Digital	0.64	0.21	-2.14	**0.03**	0.47	0.11	-7.09	**<.005**	0.78	0.41	-0.61	0.54	0.33	0.63	-1.75	0.08
Mainly Category A Drugs	0.69	0.21	-1.74	0.08	0.75	0.11	-2.69	**0.01**	1.90	0.40	1.62	0.10	0.61	0.65	-0.76	0.45
Mainly Category B Drugs	0.74	0.21	-1.43	0.15	0.75	0.11	-2.70	**0.01**	1.58	0.40	1.14	0.26	0.45	0.65	-1.23	0.22

Listing Categorization

Each market provides a different categorization of goods. For cross-market comparability, we use the listing category classifier from Soska and Christin [34]. This classifier predicts whether a listing pertain to one of the following categories: Opioids, Ecstasy, Psychedelics, Cannabis, Digital Goods (e.g., malware, cybercrime, carding), Prescription-based Drugs, stimulants, Benzodiazepines, Dissociatives, Other (which combines drug paraphernalia, weapons, electronics, tobacco, sildenafil, and steroids [34]), and Miscellaneous (everything that does not fit in any of the above categories).

5.4 Survivability Drivers

We first explore the impact of reputation scores on vendor survivability using a Cox proportional-hazards regression. We include in our model covariates that capture the main type of goods that the vendor offers, as well as whether they operate in a different market. Last, we control for the effects of wealth by stratifying our experiments.

Past work has measured the survivability of vendors by employing Kaplan-Meier models and using the observability (i.e., reachability of the page or last observed activity[6]) of vendors and listings to define the "death" event [33, 34, 245]. However, Kaplan-Meier models are univariate and do not allow us to observe the effect of various covariates, nor can they be used with continuous variables.

[6]Some marketplaces present a "last seen" field in vendor profiles that seems to track login activity.

5.4.1 Experimental Setup

We define our death event to be the last week that a vendor has an observable sale (as observed by the feedback timestamp) in the market. If the vendor had a sale in the last two weeks of the market, we consider the vendor to have remained alive until the market end. We do this to account for collection errors during the days preceding a market takedown operation. Using statistical terminology, all vendors who did not die prior to the end of the market are "right-censored."

We are interested in the effects of reputation scores on the survivability of a vendor. To explore this effect, we also include covariates that may impact the survivability, namely, the main type of goods sold by the vendor, as well as their presence in other markets. To determine presence in other markets, we matched case-insensitive handles. Tai et al. show this approximation is acceptable, given the absence of ground truth and infrequent occurrences of impersonation [202]. Last, we account for the wealth tier the vendor belonged to during our last observations. More specifically, we encode our variables as follows:

- **Average feedback value (FB)**: the mean value of all the feedback the vendor received. If the market does not use a 5-point scale, we transform the scores using a min-max scaler.
- **Presence in other markets (POM)**: encoded as an indicator variable, 1 indicates the vendor's name exists in a different (contemporary or earlier) market, 0 and if not.
- **Main category**: Soska and Christin [34]'s classifier distinguishes between 10 categories of goods. To reduce the number of covariates in our model, we re-label the categories into a smaller set considering the potential harm to users [262]. We distinguish between category A drugs (potentially more harmful): opioids, ecstasy, prescription, stimulants, benzodiazepines, and dissociatives; category B drugs: psychedelics and cannabis (potentially less harmful); and digital goods (D). We exclude miscellaneous goods such as counterfeit goods and weapons, as their sales volumes are very small. We then create three indicator variables, where a 1 indicates the main category of goods sold by the vendor, between category A drugs (MA), category B drugs (MB), and digital goods (MD).
- **Wealth tier**: vendors are divided into quartiles based on the revenue they accumulated at the time of our last observation. We encode this as 1 to 4, where 4 corresponds to the highest 25% earners. Because this variable is correlated with survivability, we do not include it as a covariate. Instead, we stratify our model based on the four tiers.

The hazards regression formula for all markets is then:

$$h(t) = exp(\alpha + \text{FB} + I(POM) + I(MA) + I(MB) + I(MD)) \,.$$

Last, we run two additional experiments with the features derived from forum data, namely the `/r/HansaDarknet-Market` and `/r/DarkNetMarkets` subreddits and the internal Nemesis forum. For Hansa, we encode the variables as two indicator variables which capture negative and positive mentions. We choose indicator variables as the encoding for two reasons. First, plenty of users refer to vendors by aliases or abbreviations (e.g., "YD" for YOURDEALER). Our fuzzy matcher is not able to catch these instances so such users are underrepresented. Further, we noticed that in some threads, users almost exclusively mention a vendor by name, whereas in other threads vendors are introduced by name once and subsequent comments only refer to them using pronouns. Thus, we smoothen the effect with indicator variables. For Nemesis, we add as covariates the vendors' degree and various centralities, as well as the number of posts made and the number of posts deleted. However, experiments involving betweenness, eigenvector, and closeness failed to converge, so we omit them.

The hazards regression formula for the extended variables in Hansa and Nemesis are as follows:

$$h(t) = exp(\alpha + I(\text{Pos.Mention}) + I(\text{Neg.Mention})) \,,$$

and

$$h(t) = exp(\alpha + \text{Deg.Cent.} + \text{Bet.Cent.Nr.Posts} + \text{Nr.Del.Posts}) \,.$$

5.4.2 Results

We find that the average reputation score of each vendor is significantly ($p < .005$) associated with a decrease in the hazard rate across all markets except Pandora, as seen in Table 5.2. The interpretation for the exponential of the coefficient ($\exp(c)$) is that, for example, a one-unit increase in the average feedback value on Silk Road 1, corresponds to a 50% decrease in the hazard rate. We also observe the same significant reduction in the hazard rate on vendors who had a presence in other markets. We find more mixed effects on the category of drugs being sold. That is, whether the seller mainly class A "harder" or class B "softer" drugs has mixed impact on the hazard rate across markets. Vendors who focused on digital goods, however, were more consistently correlated with lower hazards with some significant effects ($p < .05$) observed in Silk Road 2, Evolution, and Alphabay.

In our extended experiments for Hansa, we found that positive mentions of vendors across subreddits decreased the hazard rate by 30% significantly ($p = .01$), as

Table 5.3: Cox Proportional Hazards Regression on forum features extracted for Hansa and Nemesis.

	Hansa-Extended			
Covariates	$\exp(c)$	SE	z	p
Positive Mention	0.70	0.15	-2.43	**0.01**
Negative Mention	0.80	0.24	-0.93	0.35
	Nemesis-Extended			
Covariates	$\exp(c)$	SE	z	p
In Degree	0.99	0.01	-1.18	0.24
Out Degree	0.02	1.02	1.71	0.09
Nr. of Posts	0.99	0.01	-0.47	0.64
Nr. of Upvotes	1.00	0.00	-0.26	0.80
Nr. of Del. Posts	1.04	0.06	0.65	0.52

observed in Table 5.3. In the case of Nemesis, we did not observe significant effects across the measures of centrality that we tested, nor across the number of posts or deleted posts that vendors had.

5.4.3 Reputation Slander Attack

By leveraging the results from our survivability analysis we can conceptualize the cost and potential impact of a reputation attack. Past work suggested interventions that exacerbate information asymmetries in these markets to push them to failure [263, 264]; and showed that reduction in reputation may affect vendors' trade capabilities [214]. An example proposed by Franklin et al. in IRC-based markets was to use Sybils to slander the reputation of vendors [265].

Our results indicate that a slander campaign may only work if done through product reviews within the market and not in forums. In our model, we did not observe that negative mentions had an effect on survivability. Forum signals may in fact be too noisy to a prospective buyer. For instance, vendor visibility across posts could also help advertising. Likewise, negative comments are not always unilaterally accepted, instead they often draw debate and alternative experience reports from other buyers. This phenomenon was also noted by Morselli et al., when exploring conflict resolution techniques in criminal forums [239]. On the other hand, product review scores have a marked impact on survivability. Based on these results, we can infer the theoretical cost and impact of the attack as follows: we calculate the cost of de-

creasing a unit of average review score based on the lowest item cost. Let CA be a vendor's current average score, TF the total number of reviews they have received, L the lowest review that can be given, and F the number of feedback needed for the attack. Then,

$$\frac{CA \times TF + L \times F}{TF + F} = CA - 1 \,,$$

and

$$Cost = F \times \text{Item Cost} \,.$$

Solving for F gives us the cost of a reputational attack on a given OAM vendor by increasing their hazard. As an example, the vendor "YOURDEALER" (one of the largest vendors in Nemesis, at the time of writing) has an average feedback score of 4.99 from a total of 742 reviews, and their lowest priced item is $9. It would take 254 1-star reviews for a total cost of $\sim \$2,286$ to reduce their average rating by 1 unit and thus increase their (predicted) hazard by 64%. In practice, less 1-star reviews might be sufficient to cause fear in future vendors. Furthermore, the cost could be further reduced by conducting these attacks early in a vendor's career.

5.5 Predicting Success and Longevity

We now explore whether we are able to predict the financial success of a vendor, and the variables that drive their success. For interpretability, we use standard decision tree-based models. We train and test a standard prediction model which does not capture time variation across variables, and a model which does. We then repeat our experiments with the additional variables from Hansa and Nemesis. Last, we explore the generalizability of our models by training and testing with different market combinations.

Given a set of observable features from a vendor at a given state of the market, our first goal is to predict the wealth tier (i.e., revenue quartile) to which the vendor will belong at some point in the future. We then repeat this process by incorporating temporal features and forum-derived features for Hansa and Nemesis.

We do not attempt to predict revenue directly because revenue estimates are noisy and can often be heavily biased by collection and inference factors [37]. Consider the case of Nemesis, where vendors can choose to create a listing with various price options, or create one listing per offering. Using feedback as proxy for sales, we have no way of inferring which option the buyer used. Thus, the range of potential revenue that we could estimate for the vendor is wide, depending on what price we choose to use for our proxy. Furthermore, using quartiles allows for evenly balanced prediction targets.

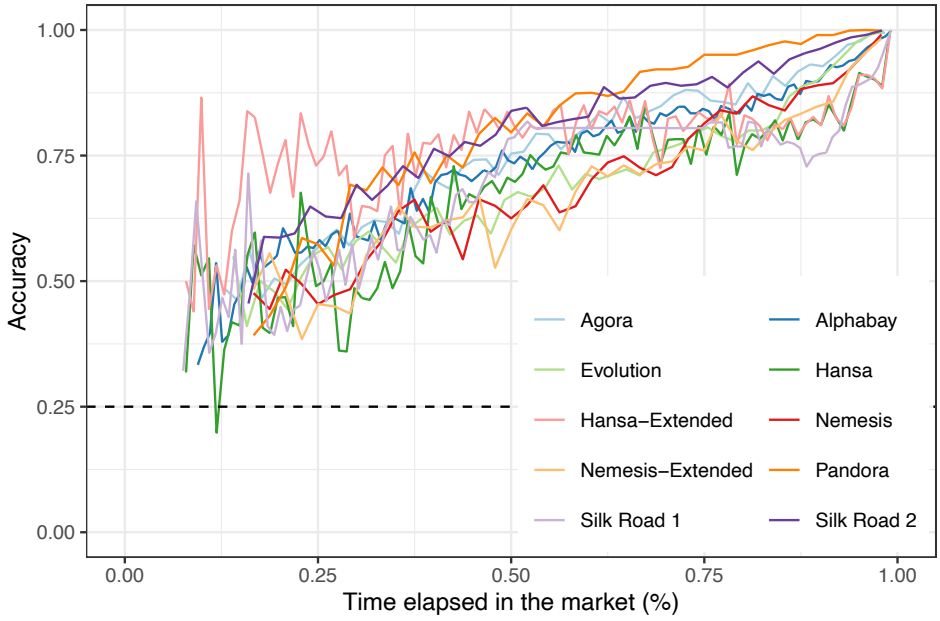


Figure 5.3: Accuracy of the RF model using base features in predicting the end-of-market revenue quantile a vendor belongs to, across markets. Labels are balanced. Timesteps are scaled to market lifetime percentage for visualization. Decrease in accuracy is due to new vendor entrancy. "Extended" includes subreddit/forum features.

Experiments Setup

For each market, we split the market into weekly intervals. We label each vendor with the quartile they belong at the end of the market (i.e., the last week the market was active prior to a takedown, or in the case of Nemesis, the last week for which we have data collection). Then, we iteratively split our dataset into observation intervals up to a given week. At each time step, we train a model based on the state of the market at that time. As we include observations of the market, new vendors appear and the features evolve.

We first train a Random Forest Classifier (RF) on the observable vendors' base features (described in Section 5.3.4). A Random Forest model is an ensemble estimator that fits decision trees to various sub-samples of the data [266]. We also train two additional classifiers on Hansa and Nemesis with the additional features extracted from their corresponding forums.

We hypothesize that time and time variation of features carry signals which will improve our estimation task. For instance, we may want to capture vendors with first-mover advantage, or the momentum of sales that a vendor has from one time step to the next. Our base features can be made more expressive by adding a time

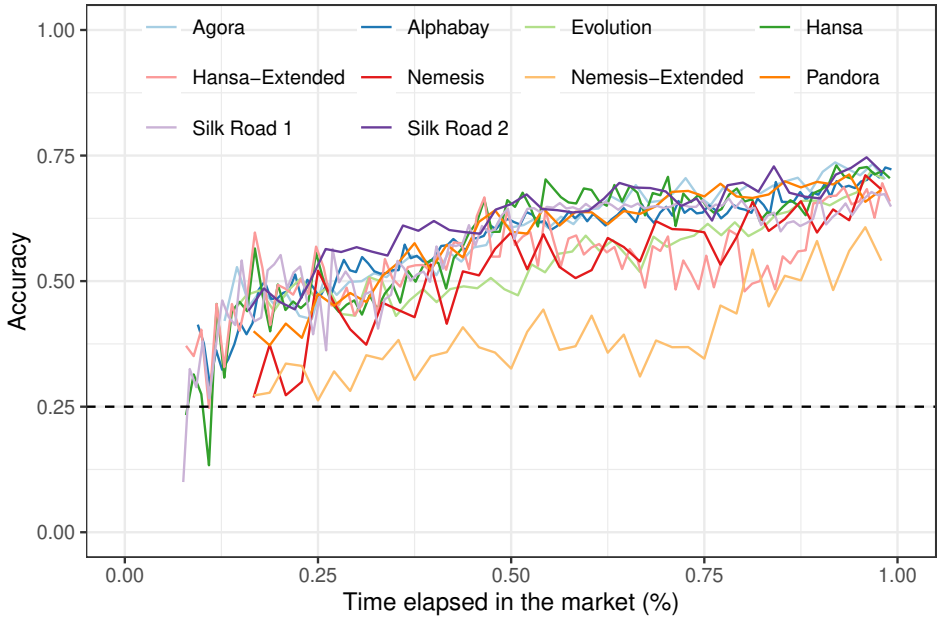


Figure 5.4: Accuracy of the TSF model using temporal features in predicting the end-of-market revenue quantile a vendor belongs to, across markets. Labels are balanced. Timesteps are scaled to market lifetime percentage for visualization. Decrease in accuracy is due to new vendor entrancy. "Extended" includes subreddit/forum features.

dimension. Using only the base features defined in Section 5.3.4, we have a matrix of shape $(nr\_vendors \times nr\_features)$ up to time T in the market. However, we could also build a time series for some of the features by tracking the evolution of features over time, as described in Section 5.3.4. That is, we break T into a series of time steps t_i, and end up with a matrix of shape $(nr\_vendors \times nr\_timesteps \times nr\_features)$.

To conduct a classification task on our time series data, we train a Time Series Forest classifier (TSF). A TSF model extends a RF classifier by sub-sampling the input time series into slices of random lengths (denoted as "windows" in the model) and extracting the mean, the standard deviation, and the slope. Each of these windows can provide insights into the temporal characteristics of the input time series, allowing us to explore what windows and features were the most relevant in the prediction [267]. Similar to the RF classifier, the sub-trees in TSF choose a label using hard voting.

We repeat the same process we defined with our RF model for all markets. We use out-of-the-box parameters for our models: 100 estimators and maximum depth of 4 for both the RF and the TSF. TSF has an additional parameter: the number of windows. For this, we choose the number of timestamps as the number of windows.

We use a 75/25% train test split.

Results

Our models perform better *without* the temporal features, as observed in Figures 5.3 and 5.4. Even when the model has access to almost a complete view of the market, the average accuracy plateaus at 70% for our TSF model. This indicates that having temporal features is detrimental to the model's performance, which is possibly caused by the hard voting mechanism the TSF model uses. The model may be learning features from earlier portions of a vendor's performance that may seem to indicate future success. Since it weighs these features equally to more recent data, the newer observations are unable to affect the final prediction.

On the other hand, we see that the RF model converges to a perfect accuracy as the market evolves. At 20% of the market's lifetime, we achieve over 40% of accuracy in predicting the vendors who would accrue the most wealth by the end of the market. At 40% of the market's life time our accuracy is mostly over 60%. And by 80% of the market, we have over 75% accuracy, and over 90% accuracy for two markets.

Last, we find that the additional forum features seem to have little effect on the prediction accuracy. Hansa's forum features decrease the accuracy of the model. Nemesis shows the opposite. In either case, the effect is small in the RF model. On the other hand, we see a significantly higher negative effect in the TSF model. In the case of "Nemesis-Extended," we see an accuracy of 5–10% decrease at each time step, as well as less convergence towards the end of the market. In this case as well, the forum activity habits of vendors of different sizes may not be sufficiently distinct for these features to carry a meaningful signal, which ultimately confounds the model.

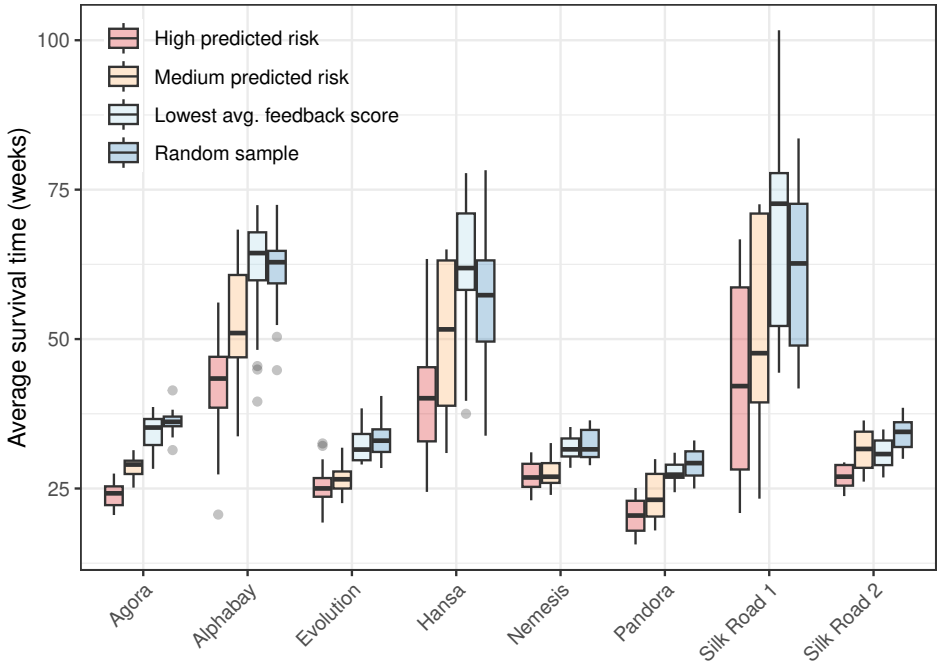
5.5.1 Predicting Vendor Disappearance

We now attempt to build a model to predict whether a vendor will leave the market or is at risk of doing so. Similar to before, we consider a vendor to have disappeared from the market at the time they stop receiving feedback. To do this, we employ the base features we described in Section 5.3.4. Furthermore, we design our experiments to combine observations across markets for generalizability.

Experiment Setup

Our goal is to identify the vendors who are on the brink of leaving the market. To do this we design a classifier that attempts to predict one of the following: 1) whether a given vendor will leave the market in the next month (high-risk), 2) whether the

Figure 5.5: Average survival time in weeks for each group, across all market timesteps. Vendors who stop having sales after a given week are removed from the sample. The high/mid-risk groups across markets were assembled with the same classifier, which was trained with samples from all markets.



vendor will leave the market after the first month but before the third month, or 3) whether the vendor will still be active after the third month.

For each market, we split the market lifetime into weekly intervals. At each stage of the market, we label the vendors according to the labels above. We then combine data from different markets. However, given that our prediction goal is not end-of-market revenue, we do not combine them based on the percent of revenue accrued by the market. Instead, we naïvely combine vendors from different markets based on the amount of time elapsed in the market. That is, we combine observations from a vendor from market M at week W with a vendor from market M' at week W. Further, as vendors disappear from the market, we remove them from our sample (so as to not overfit on already disappeared vendors).

We then train and evaluate an RF model and a TSF model with default settings, similar to our experiment setup in Section 5.5, at each timestep. We note, however, that both models perform poorly (F_1 ¡ 0.25 for high/mid-risk vendors) due to class imbalance (i.e., not a lot of vendors disappear within 1–3 months). Therefore, instead of evaluating our model based on label prediction, we collect the label probabilities

for each class. To do this, we pick the classifier with the best F_1 score for high/mid-risk labels, trained on a subset of data from all markets. We then use this classifier to assemble, for each timestep, the 20% of vendors with: 1) the highest probabilities in the high-risk class, 2) the highest probabilities in the mid-risk class, 3) the lowest average feedback score, 4) and a random sample. We take the average survival time for each of these groups, at each timestep of each market. We chose 20% as it provided a big enough sample size for each market, while reducing the overlap between vendors across groups.

Results

Across all markets, the group of vendors assigned the highest probabilities of being at "high-risk" indeed had shorter lifespans as compared to the other groups, as seen in Figure 5.5. Vendors in the "mid-risk" category also had shorter lifespans than the other groups, except for Silk Road 2. We observe that a low average feedback score seems to carry some signal of quality, given that vendors with low average feedback score have, for the most part, slightly shorter lifespans than a random sample. However, we observe that a low average feedback score, alone, may not be clearly indicative of a near-term disappearance from the market. For instance, established vendors may have a dip in their average feedback score in a given week, but that may not necessarily shorten their lifespan significantly. Our main finding is that making a prediction on the lifespan of a vendor, may depend on more variables beyond just average feedback scores.

Figure 5.6: Comparison of average prediction accuracy for the revenue quantile that a vendor belongs to by the end of the market. Accuracy is averaged across market stages. We train on $n-1$ markets and test on the holdout. This plot compares training performance on the $n-1$ markets with test performance on the held-out market.

Figure 5.7: Comparison of test scores across categories for predicting vendor revenue quantile by market end. Category segmentation includes only vendors whose primary goods fall into the given category. The baseline line shows performance without category segmentation.

5.6 Generalizability and Feature Importance

For our vendor disappearance model, we trained our model by mixing vendor observations from different markets because the event of interest is whether a vendor stopped receiving feedback. That is, the labels are not significantly different across markets. We claim generalizability for this model, given that we used a single classifier, trained on traces from all markets to do predictions for each of these markets across each of the markets' lifetimes.

For our financial success model, however, our labels are the end-of-market revenue quantiles for a given market. Thus, we cannot directly combine vendor observations from different markets For instance, vendor V from market M may belong to Q1 with \$1M revenue, whereas vendor V' from market M' may belong to Q3 with \$1M of revenue. Furthermore, in our vendor disappearance model we were able to directly combine our traces based on the time elapsed in the market. However, revenue is trickier. Consider the case of Silk Road and Hansa. Silk Road was the first successful market, facing little competition in its early stages. Hansa on the other hand, was a market that had little traction for over a year, and gained most of its revenue following the Alphabay takedown. If we combine vendors' data from the first month of Silk Road with the data from the first month of Hansa, we are combining two disparate market environments. Instead, we combine market's data when their environments were most similar.

Thus, to explore the generalizability of our financial prediction model. We design an experiment where we train a model on $n-1$ markets and predict on an unseen market. Further, we combine cross-market observations by combining traces at stages where the markets had accrued a similar revenue percentage. We perform this experiment with all vendors, and also by segmenting vendors by the main category they sold. Last, we discuss feature importances across each model.

5.6.1 Experiment Setup

We want to repeat the experiments defined in Section 5.5 by training a model on a set of markets and testing our prediction on an unseen market. To combine observations across different markets, we explore a simple heuristic: splitting the data by the percent of revenue accrued by the market. That is, we iteratively split each markets' data at the time they accrued 10%, 20%, ..., 90% of the revenue at the time of their last observation. We then iteratively combine the data of $n-1$ markets to train our model, leaving one market out completely (which we call our *holdout market*). We train and evaluate each model on these $n-1$ markets using a 75/25% train/test split. We then test the performance of our model on our holdout market. Finally, for each model,

we conduct an ablation study by iteratively removing each of the feature categories described in Section 5.3.4: listing features, revenue features, and feedback/reputation features.

We hypothesized that our classifier accuracy could be improved by segmenting vendors by category. We used the same labeling of Section 5.4. We combined vendors who sell mainly category A ("harder") drugs, B ("softer") drugs, and digital goods. For each market, we followed the same procedure as mentioned above, except that we only trained and tested on one category at a time.

5.6.2 Results

Our financial success model generalizes well to 6/8 other markets even when using a naïve heuristic to combine markets' data, as seen in Figure 5.6. Across all markets, the average accuracy during evaluation stayed consistent. This means that even when we shuffled vendors from different markets during our train/test split, we were able to maintain a consistent accuracy of over 70% for all markets except Alphabay, which was 67%. Furthermore, in 5/8 markets we observed similar performance between the accuracy during training/evaluation and the accuracy during testing. This means that our model was able to perform well when doing prediction on vendors from a completely unseen market. The results from Nemesis indicate generalizability across time, given that Nemesis is significantly more recent than some markets (e.g., it appeared 10 years later than Silk Road).

With regards to our category segmentation approach, we observe mixed results across markets, as seen in Figure 5.7. In general, we do not see significant improvements/deterioration in performance over our baseline across markets. This could be due the categories being too broad, due to a reduction in the size of the training set, due to our current approach at combining market segments, and/or due to different category performance dynamics across markets (e.g., market X is more popular for drug A, whereas market Y is more popular for drug B). Nonetheless, we believe some form of segmentation is useful, but will likely require market-specific optimizations.

In Table 5.4, we show the precision, recall, and F1 scores for our experiments with each holdout market, across our 4 revenue quantiles. The model performs best when doing predictions on the lowest/highest earners (Q1 and Q4). Because the overall market revenue follows a power law distribution, the middle portion (Q2 and Q3) are harder to distinguish. Last, in Table 5.5, we show that the absence of revenue-related features decreases accuracy the most. When only reputation features are excluded, accuracy is barely affected. When revenue and reputation features are both excluded, the model suffers the biggest loss. Listing-related features have little impact on the model.

Table 5.4: Average classification metrics across our 4 labels (wealth tiers). The holdout is the market on which we predict while training on the others. Labels are balanced across classes. Each metric is the average score obtained across our 10 experiments.

Wealth Tier	**Q1 ($x \leq$25%)**			**Q2 (25%$< x \leq$50%)**			**Q3(50%$< x \leq$75%)**			**Q4(75%$< x$)**		
Holdout Market	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Silk Road	0.77	0.82	0.79	0.63	0.72	0.67	0.67	0.62	0.64	0.88	0.80	0.83
Pandora	0.83	0.55	0.65	0.47	0.50	0.48	0.54	0.64	0.58	0.81	0.84	0.82
Silk Road 2.0	0.76	0.31	0.43	0.22	0.16	0.18	0.23	0.19	0.20	0.54	0.93	0.68
Agora	0.64	0.76	0.67	0.53	0.66	0.58	0.63	0.62	0.62	0.92	0.69	0.77
Evolution	0.74	0.99	0.84	0.57	0.78	0.65	0.56	0.52	0.54	1.00	0.54	0.69
Alphabay	0.81	0.96	0.88	0.74	0.78	0.76	0.77	0.64	0.69	0.87	0.83	0.85
Hansa	0.48	1.00	0.65	0.26	0.38	0.31	0.28	0.20	0.23	1.00	0.40	0.57
Nemesis	0.68	0.96	0.79	0.65	0.72	0.68	0.66	0.65	0.64	0.99	0.62	0.76
Average:	0.71	0.79	0.71	0.51	0.59	0.54	0.54	0.51	0.52	0.88	0.71	0.75

Table 5.5: Ablation study of our revenue prediction model on holdout markets. We exclude combinations of features and quantify the accuracy decrease on the model.

Excluded Feature(s)		**Avg. Accuracy Decrease**			
Feature Set 1	**Feature Set 2**	**Min.**	**Max.**	**Mean**	**Std.**
Revenue	–	0.01	0.31	0.16	0.08
Reputation	–	<0.01	<0.01	<0.01	<0.01
Listing	–	<0.01	0.02	<0.01	<0.01
Revenue	Reputation	0.05	0.44	0.27	0.12
Revenue	Listing	0.04	0.31	0.16	0.09
Reputation	Listing	0.02	0.02	<0.01	<0.01

5.6.3 Explaining and Improving Performance

We hypothesize that the poor performance on Silk Road 2 and Hansa is due to the unique environments that these markets faced, as seen in Figure 5.2. Essentially, vendors in these markets may be considered to be out of distribution. Hansa had unremarkable economic activity until two months before its takedown. Following the Alphabay takedown, about 5,000 users a day flocked to Hansa [268]. Intuitively, models that were trained in economic activity from markets that did not experience the same trajectory are bound to have poor performance, as observed in Figure 5.3. In the case of SR2, this market had strong performance from the beginning likely due to its brand recognition after the original Silk Road's takedown. However, its performance gradually degraded due to a series of issues (arrest of moderators and a hack) [269], as opposed to gradually ramping up. Because of this, a naïve model that

trains on markets dissimilar to Silk Road 2 yields lower quality results.

While segmentation may not offer substantial improvements, a set of approaches can be adopted at other stages of the pipeline. For consistency, we conducted our holdout experiments by training on $n - 1$ markets. However, some markets have uncommon trajectories (e.g., SR2, Hansa). In a practical setting, we may need to curate our training set based on the target market. For example, if the target market was born in response to a takedown, or faces more/less competitors, we ought train our models on markets with similar characteristics. With regards to our prediction goal, we naïvely consider everybody in a quartile to have the same label (a classification task). Instead, we could design our model to be a regression tree over vendor revenue percentiles to preserve relative ordering within vendors; we could also define arbitrary cutoffs (e.g., top 5% of vendors). Last, our models can be improved with traditional machine learning optimizations: testing other models (potentially trading explainability for accuracy) and finetuning parameters.

5.7 Discussion

Our results from Section 5.4 indicate that reputation, derived from feedback scores, plays a role both in the financial success as well as in the longevity of vendors, although in different forms. Our proportional-hazards regression shows that average feedback scores in the market have a significant impact on the survivability of a vendor. Across the board, we see that a 1-unit increase in average reputation reduces the hazard rate of vendors across the markets. However, this regression leverages a full view of the market. That is, as a whole, feedback scores can explain the disappearance of vendors.

On the other hand, our results from Section 5.5.1 show that the average feedback score is not the best predictor of a vendor leaving the market in the short term. That is, as the market progresses, vendors with lowest average score may not necessarily leave the market. This effect surfaces on the markets that have a longer lifetime (i.e., SR1, Hansa, Alphabay). Instead, our model, by leveraging more vendor features, better identifies vendors at a higher risk of disappearing. Furthermore, our model generalizes across markets and time, given that it was trained on vendor observations from 8 different markets spanning 12 years.

With regards to the financial success of vendors, we demonstrated that our predictions generalize across most markets. The average feedback score seems to play a role in predicting their future wealth. However, it is not the main predictor. Rather, past financial performance is a better predictor of future financial performance. In part, we hypothesize that this is the case because scaling criminal operations is hard, particularly for drug-related items [270]. Thus, vendors who demonstrate capacity

to scale their business early (as demonstrated by large sales) often become dominant vendors. Another reason why sales volume and history are likely drivers of success is because these signals are hard to fake. Décary-Hétu et al. noted that signals which could be cheaply purchased had little impact in predicting sales [36]. Frequent sales, over time, ultimately create an attractive signal for buyers who want to reduce risk. When segmenting vendors by category, however, we do not observe a significant difference across markets. We believe that segmentation may help, but may need to rely on other approaches to combining market data and accounting for vendor offering diversity. Similarly, our time-series model performed significantly worse and may also benefit from ddiferent feature engineering approaches. Remarkably, we did not perform any parameter tuning; instead, we employed out-of-the-box defaults. We did not employ sophisticated feature transformations nor models. Rather we focused on explainability and establishing a performance lower bound.

Across our experiments we did not observe a significant effect from signals derived from forums, neither from the co-located forum (for Nemesis), nor from the external forums (subreddits). During our manual analysis, we observed that forum signals are predominantly noisy. We observed small vendors that frequently used forums for advertising. A vendor who posts a lot can easily build an impressive social network through their interactions, despite not driving sales. We also observed large vendors who were not mentioned even once, and who also did not engage in any of the forums. Furthermore, negative reviews in forums were often not unilaterally accepted but often raised discussions from other users in the community, a similar finding to Morselli et al. [239].

5.7.1 Interventions and Policy Takeaways

Our results can help improve interventions in two ways. First, our prediction model can readily be used in new markets to identify vendors who will become big earners. Early identification allows for monitoring efforts to be more efficient, particularly as OAMs and criminal forums increasingly adopt adversarial anti-scraping mechanisms [197]. Cuevas et al. demonstrated that focusing scraping efforts on more popular vendors using a naïve algorithm improved coverage and inferences substantially [37]; our results build on that approach. Our prediction model ought to be taken probabilistically: not as a definitive answer, but as a tool that can help navigate uncertainty. Second, we show that slander attacks may be viable and cost-effective, particularly when done early in a vendor's career. Our findings suggest, however, that slander attacks ought to be done through low score feedback orders and not through slander in forums.

With regards to market design and policy, our results demonstrate that existing

reputation systems within these markets carry a signal that can help reduce harm in the long run. However, this signal is imperfect and may not have a strong enough effect in the short term. On one hand, the continued success of these markets are testament to the fact that existing reputation systems are, however crudely, culling out low-quality vendors. On the other hand, our simple classifier demonstrates that there are other signals which seem to more readily identify vendors who might disappear from the market. While there are a variety of benign reasons why a vendor may leave a market, there are some quite harmful ones, such as vendors who sell dangerously adulterated drugs. A model or signals which can more quickly alert buyers of these situations can substantially reduce harm in the long run. Policies which consider the regulation of two-sided marketplaces (particularly for drugs), ought to consider the reputation system design as well.

5.7.2 Limitations and Future Work

First, we do not test a large number of covariates through our proportional hazards model, because a "one-in-ten/twenty" rule (1 covariate for every 10/20 deaths) is advised for proportional hazards model [271]. Thus, while we identified a set of meaningful covariates contributing to vendor survivability, there may be other latent factors which our model does not capture. Second, our financial success prediction model only predicts the wealth quantile that a vendor will belong to. Within the top 25% of vendors there may be significant variance in revenue. Third, we only tested the impact of external reputation signals for one market (Hansa) and social network features for one market (Nemesis). Our manual review of these signals indicates high noise, particularly as it relates to the success of vendors. However, these features may correlate with other vendor attributes which future work may explore. In our study, we saw less accuracy from the TSF model which sought to capture time-based feature changes. However, it may be useful to explore other feature engineering approaches that incorporate temporal features. Furthermore, we leveraged qualitative analysis to extract signals from forums in an effort to collect high-fidelity signals. However, scaling this work manually is inefficient. Current off-the-shelf named-entity recognition and sentiment analysis techniques did not perform well on our dataset. However, advances in large language models, particularly for coding textual data [272, 273], may allow our forum analyses to scale.

Chapter 6

Using Public Signals to Identify Risky Vendors in Cryptocurrency P2P Markets

---

This chapter is adapted from our paper:

> [39] Taro Tsuchiya, Alejandro Cuevas, and Nicolas Christin. "Identifying Risky Vendors in Cryptocurrency P2P Marketplaces". In Proceedings of the ACM Web Conference (WWW '24), May, 2024. Association for Computing Machinery. https://doi.org/10.1145/3589334.3645475.

6.1 Motivation and Goals

This chapter examines the limitations of profile signals in cryptocurrency peer-to-peer (P2P) marketplaces by evaluating their ability to predict account suspension due to fraud, abuse, or illicit activity. Building on the modeling techniques developed in the context of online anonymous marketplaces, we turn to Paxful[1] and LocalCoinSwap[2]——two of the most active P2P cryptocurrency platforms——and ask: can the signals shown in user profiles meaningfully reflect risk? If profile signals, particularly those associated to quality (e.g., ratings) and those associated to security (e.g., verifications) effectively predict account suspension, then we consider them to be informative to users. And if not, what signals should platforms surface instead

A P2P cryptocurrency exchange is a market that facilitates trade between buyers

[1] https://paxful.com/
[2] https://localcoinswap.com/

and sellers of cryptocurrency assets. *Vendors* post advertisements to buy/sell cryptocurrencies through various payment channels, such as bank transfers, gift cards, and mobile payments. *Customers* browse the market and respond to ads to initiate transactions. These platforms are different from centralized cryptocurrency exchanges such as Binance or Coinbase, where the platform matches buyer and sellers through an order book. Instead, buyers and sellers to transact directly, making trust an essential component of every transaction. As such, P2P cryptocurrency marketplaces more closely resemble marketplaces such as eBay, Craigslist, and online anonymous marketplaces [33, 34, 38].

Similar to the aforementioned online marketplaces, malicious actors may attempt to defraud users. They may reverse payments, send fake/manipulated gift card receipts, harass users to release payments or block the release of cryptocurrencies. For the platform to be trustworthy, users should thus be provided with information that allow them to assess counterparty risk. Most online marketplaces (including Paxful and LCS) use feedback-based reputation systems, where customers give vendors feedback to assess the vendors' credibility. However, these systems are susceptible to various types of attacks and manipulation such as self-promoting, whitewashing, retaliation, and bad-mouthing [13, 274]. Moreover, online marketplaces often try to verify users' identities to prevent fraud. Paxful and LCS allow users to verify identities, phone numbers, email, and physical addresses. These verifications are then shown in user profiles, and show be indicative of a lesser risk of fraud.

To evaluate profile signals in predicting misbehavior, we collect data from two leading P2P marketplaces—Paxful and LocalCoinSwap (LCS)—over 12 months, and monitor user activity including profile changes, posted advertisements, feedback received, and account suspensions. We test seven machine learning (ML) models to predict account suspension. Besides reputation metrics, we combine different publicly available information such as user profiles, ads, and trade information, thereby obtaining a more precise representation of suspicious accounts. We perform the same experiment on LocalCoinSwap (LCS) and test the generalizability and transferability when using attributes common to both platforms. We then conduct a prospective cohort study to evaluate the practical usefulness of our model. That is, we pre-select three groups of accounts: 1) users with a high likelihood of suspension from our ML model, 2) users with the lowest reputation scores, as well as 3) a baseline consisting of a random user sample, and follow them over a month. We then compare suspension rates across these three groups. In summary, we make the following contributions:

1. Our study evaluates online safety and trust in cryptocurrency P2P markets by creating year-long datasets and formalizing the methodology of data collection.
2. We empirically show the limitations of feedback-based reputation by manually investigating review quality and finding evidence of user collusion and automa-

tion.

3. We develop a mechanism to identify account suspension using only public signals, and achieve a 0.86 F1-score and 0.93 AUC using tree-based ensemble approaches in one of the largest cryptocurrency P2P markets.
4. While our model itself has limited transferability across P2P platforms, we distill features critical to both platforms.
5. Our online evaluation illustrates that our model is significantly better at proactively identifying risky accounts than existing reputation systems.

We find that our model is significantly better at identifying future suspensions than the reputation-based baseline. Surprisingly, the most prominent and theoretically relevant interface signals——ratings and verification badges——are weak predictors of actual risk. Instead, the features driving model performance are less emphasized in the user interface. Instead of feedback scores or verification status, it is behavioral signals like trade frequency, trade partner diversity, and price premiums that are most predictive. These findings echo earlier chapters in this thesis by reinforcing the disconnect between which signals are most visible and which signals are most informative.

Beyond demonstrating predictive accuracy, this chapter contributes a framework for auditing profile signals and informing interface redesign. Rather than advocating for immediate account bans based on model output, we propose using predictive models to prioritize moderation efforts, focusing scarce human review on accounts with the highest likelihood of harmful behavior. We also show how platform administrators can evaluate the effectiveness of their current trust signals by benchmarking them against empirical risk.

These findings advance the thesis in three ways. First, they demonstrate the applicability of our modeling approach to a new and increasingly relevant domain: decentralized financial platforms. Second, they strengthen the case for continuous, data-driven auditing of profile signals, especially in high-risk environments. Third, they provide practical tools and recommendations for platform operators to surface more meaningful signals, mitigate abuse, and improve user safety.

Our method can help platforms design safe and more secure environments and could help moderate suspicious activity potentially more efficiently. Our findings could also improve user experience by allowing users to more accurately identify (un)trustworthy vendors. Last, given the overall scarcity of empirical research on reputation systems due to data availability, our work benefits not only cryptocurrency P2P exchanges but also other online marketplaces to design more informative reputation systems, as a complement to existing feedback-based systems.

6.2 Background

This section overviews cryptocurrency P2P marketplaces, describes transaction mechanisms, and delves into the role of the reputation system, its vulnerabilities, and possible attacks. In this study, we consider Paxful and LocalCoinSwap. A sample profile for each platform, respectively, is provided in Figures 6.1 and 6.2.

6.2.1 P2P Cryptocurrency Marketplaces

By providing lower friction than alternatives, Bitcoin [275] and other cryptocurrencies have been used for international remittances [276]. Also, despite being far from anonymous (with a few exceptions like Monero or Zcash), modern cryptocurrencies provide stronger privacy than most other electronic payments, and have been used in online anonymous marketplaces [33, 34], for malware and extortion payments [277], or even financial scams [278]. Additionally, due to their high volatility [279], related financial products, e.g., derivatives [280], have become increasingly popular as a speculative instrument.

While most people trade cash for cryptocurrencies through large centralized exchanges (brokerage or order-book style) such as Coinbase or Binance, cryptocurrency peer-to-peer (P2P) exchanges became popular by improving privacy through disintermediation. Anecdotally, those exchanges attract customers from emerging countries in Africa (Kenya, Nigeria, Ghana), Asia (China, India, Pakistan, Philippines, and Vietnam), and South America (Argentina, Colombia) where economic and/or political circumstances may limit available financial operations [281]. For instance, in Paxful, gift cards appear to be often used for remittances from the USA to Nigeria [282]. As such, P2P cryptocurrency exchanges are a plausible alternative for those with limited access to financial services.

P2P exchange mechanisms differ from centralized exchanges and are akin to other online marketplaces such as eBay, Craigslist, or Facebook marketplace. *Vendors* set offer prices for cryptocurrency (Bitcoin, Tether, etc) and post advertisements, indicating whether they want to buy or sell. Advertisements include payment type (e.g., bank transfer, mobile payment, gift cards), fiat currencies (e.g., USD, EUR, KES), and possible ID requirements for customers. *Customers* visit the exchange website, search for ads, and initiate transactions while communicating with vendors. P2P exchanges originally focused on face-to-face transactions; but a small portion of exchanges (e.g., LocalMonero) still offer this option, and face-to-face transactions represent a low percentage of all activities We thus only focus on online transactions.

We distinguish between custodial and non-custodial P2P exchanges. In *custodial* exchanges—such as Paxful—to initiate a transaction, users need to first send their

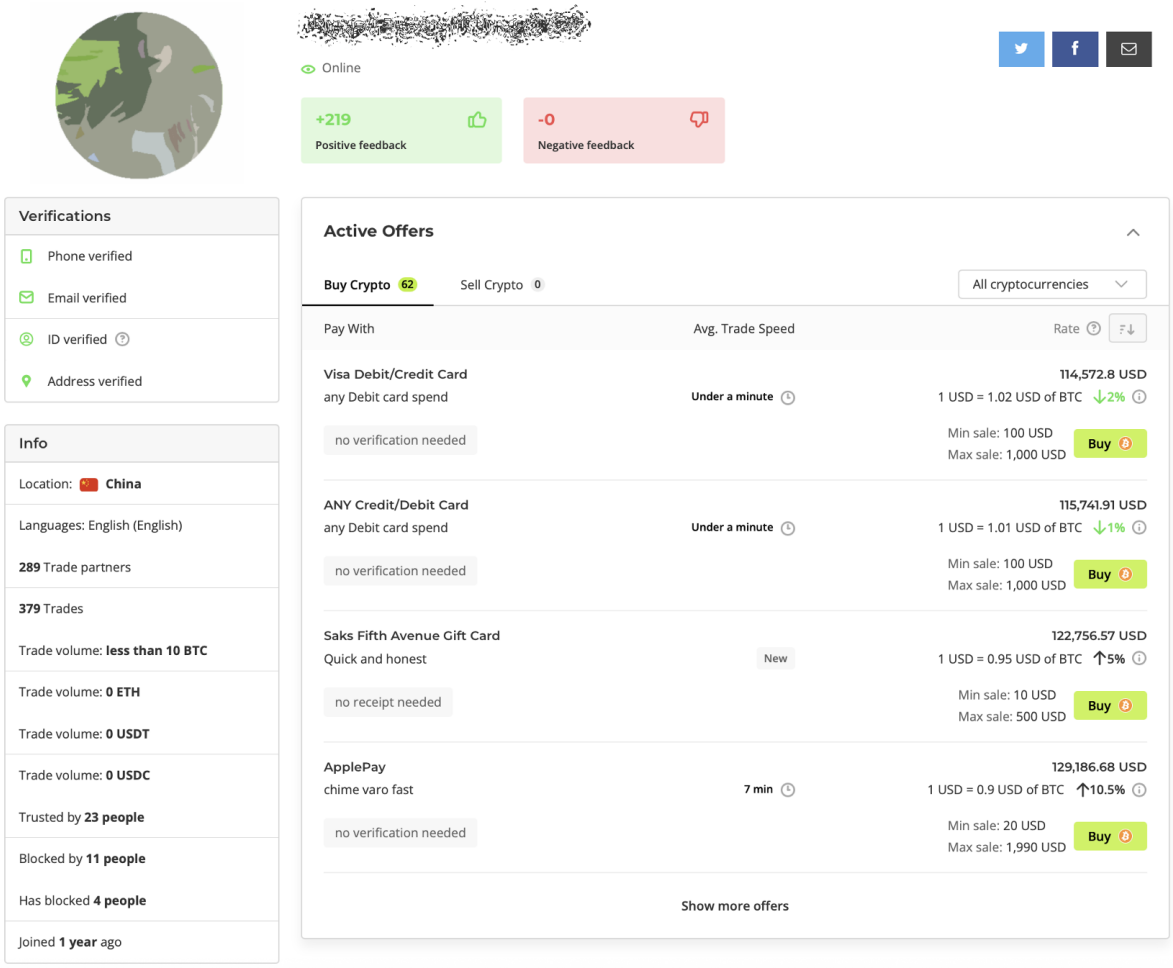


Figure 6.1: Example user profile on Paxful. The profile shows the user's reputation score, number of trades, and various forms of verification, among other signals.

cryptocurrency to the exchanges' wallet. *Non-custodial* exchanges like LocalCoinSwap (LCS) allow users to keep full control over their funds, and to directly exchange cryptocurrency between user wallets. In both cases, the platform acts as an escrow agent and moderates user disputes.

Figure 6.3 highlights the process for a transaction between a vendor (seller, here) and a customer (buyer, here): ❶ The seller sends or locks cryptocurrency (e.g., Bitcoin) to an escrow account from their wallet (either self-hosted or on the platform). ❷ The buyer pays the seller using a bank transfer, gift card, or other form of payment. ❸ The seller confirms the payment and notifies the platform. ❹ The platform releases the cryptocurrency to the buyer.

In addition, Know Your Customer (KYC) requirements may exist depending on the exchange and circumstances. For example, Paxful asks users to immediately com-

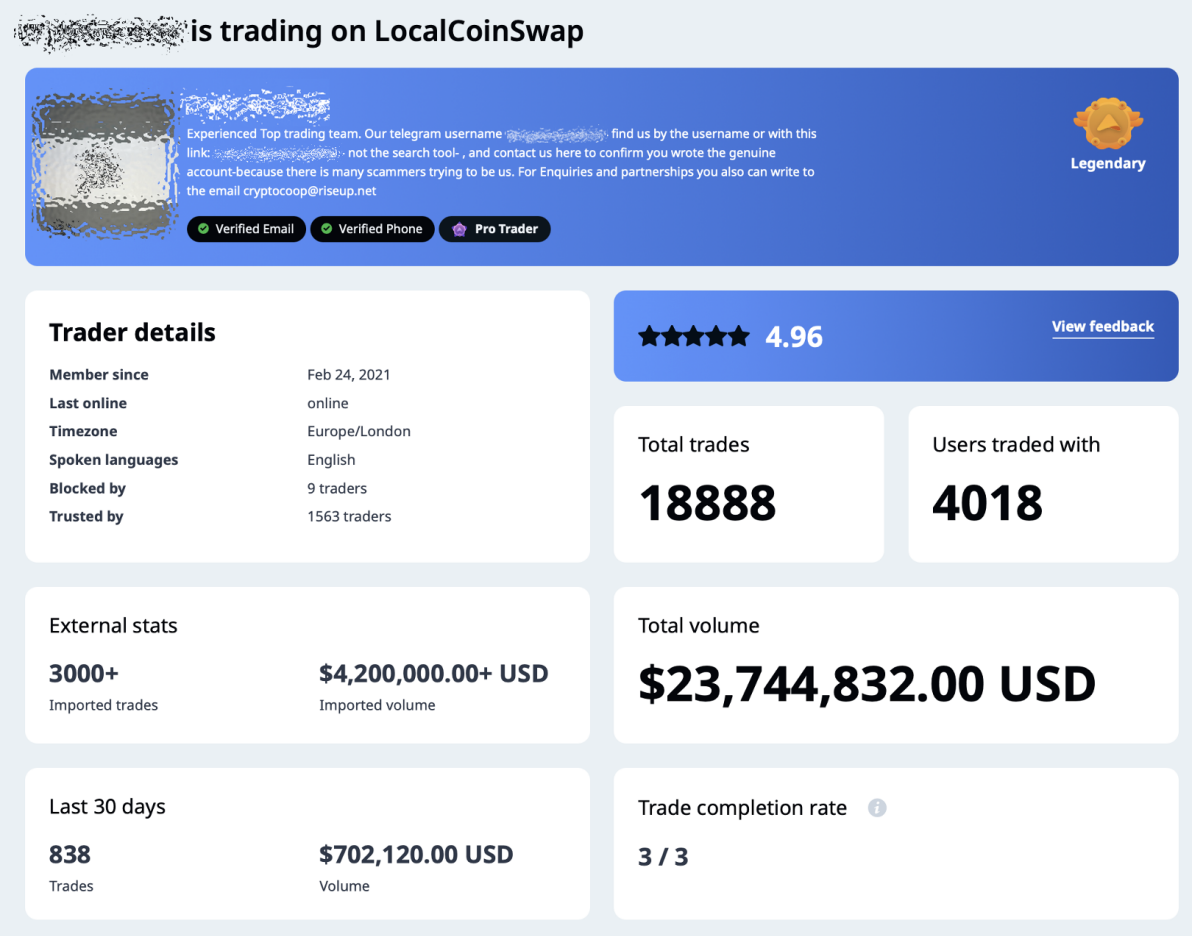


Figure 6.2: Example user profile on LocalCoinSwap. The profile shows the user's reputation score, number of trades, and various forms of verification, among other signals.

plete identity verification if they are in a listed country [283]; otherwise, identification is required when transaction volumes exceed a certain threshold, e.g., 1 000 USD. On the other hand, in LCS, ID verification is optional.

6.2.2 Reputation Systems: Benefits and Challenges

Since the dawn of the internet, online marketplaces have become the *de facto* place to exchange goods and services and help reduce inventories [1]. Without face-to-face communication, however, users face the risks of not seeing actual products, being cheated, or dealing with malicious vendors. Most exchanges build a reputation system to advise users on vendor credibility [284]. These systems are reportedly more accurate than word-of-mouth [2], and more effective at disseminating information. A

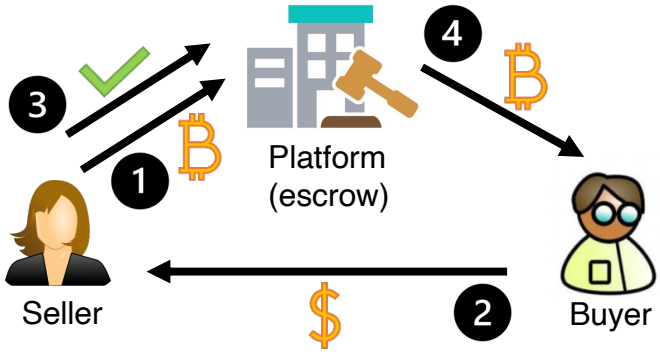


Figure 6.3: Transaction flow in cryptocurrency P2P markets.

number of studies discuss the role of online reputation and how it leads to safer and more efficient online communication, e.g., by looking at reputation system design [1, 21] or reputation impact on product price [2, 285].

Despite these benefits, reputation is vulnerable to manipulation such as whitewashing (re-entering the market under a different identity after having engaged in questionable transactions), Sybil attacks (fake accounts operated by a unique entity), slander, retaliation, and bad-mouthing [13, 274, 286]. In this chapter, we focus on *self-promoting attacks*, which Hoffman et al. [13] defines as "attackers seek[ing] to falsely augment their own reputation," by submitting fake positive feedback about themselves through their *own* Sybil accounts. Platforms that do not require user authentication or proof of interaction (e.g., payment) for feedback are particularly vulnerable. Self-promoting attacks can be conducted by a single entity or by colluding entities. We observe evidence of such attacks in our data, as we discuss in §6.4.2. Unsurprisingly, empirical evidence suggests the existence of SRE (seller-reputation-escalation) services to perform self-promoting attacks in online marketplaces [113].

6.3 Data

This section describes data collection and account suspension.

6.3.1 Collection

We collect data through Paxful's publicly available APIs from June 8, 2022 through June 26, 2023. On April 4, 2023, Paxful announced it suspended operations. Although operations eventually resumed a month later, data posterior to April 4, 2023 present oddities, including a large-scale account ban, so we choose to exclude them from

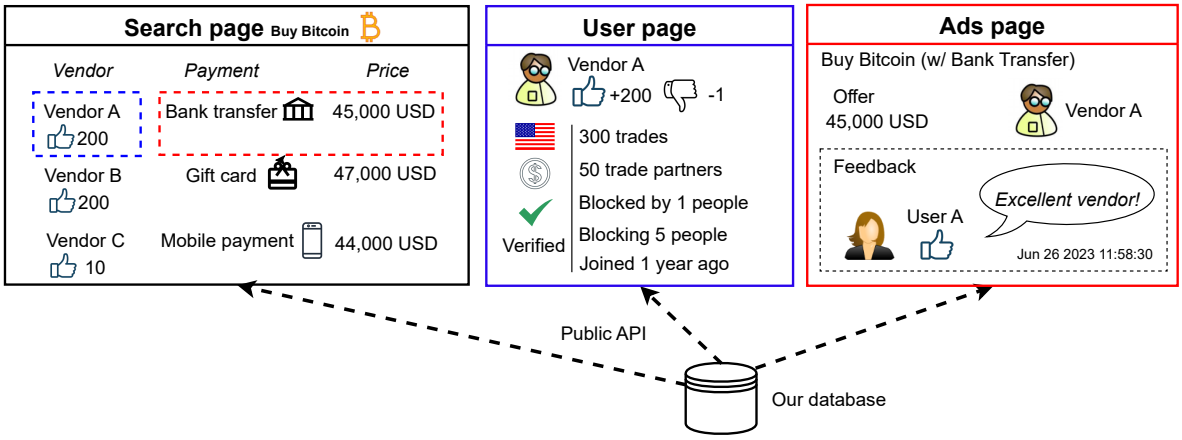


Figure 6.4: Data collection through Paxful APIs

account suspension prediction.

We query listed ads approximately every 100 seconds. The Paxful API requires us to specify trade types (SELL/BUY), types of cryptocurrencies (BTC, USDT, USDC, or ETH), and a list of countries. Because measuring every single country would be impractical, based on the number of transactions we historically observed on LocalBitcoins, we choose to limit ourselves to 10 countries (Russia, US, UK, Nigeria, Colombia, Germany, India, Peru, Kenya, and China) and the "Worldwide" option. Thus, we may miss ads *only* posted in other countries. (Most of the ads are cross-listed in multiple countries.) Ads include information on the type of cryptocurrencies and fiat currencies sought or offered, payment methods, price (and its deviation from the market price), and customer ID requirements. We visit all vendors with active ads at least once a day to longitudinally record their profile and activity statistics, which is used to evaluate our risk profiling methods in §6.7. We also collect historical feedback left by a customer associated with each ad as a one-time data collection. Feedback includes a textual review, rating, creation time, and the handle of the user giving feedback. Based on feedback data, we construct a "feedback graph" where each node represents a user and edges denote feedback from customers to vendors. Figure 6.4 visualizes the data collection scheme.

In total, we collected approximately 396 000 ads, 26 million longitudinal observations for 67 000 vendors with 4.7 million historical reviews, and information on the more than 664 000 users that left that feedback, up to June 26, 2023. In Paxful, only 0.27% of all feedback is negative—comparable to 0.39% in the eBay US market[21].

We also collect data from LocalCoinSwap (LCS) from May 27, 2022, to June 26, 2023, using their public APIs. LCS API gives us all the posted ads, so that we have perfect coverage. In addition to ad data posted on the platform, LCS API allows us to query all the historical feedback data so we can get information on all the users

who have given or received feedback at least once. In total, we collected over 52 000 ads, 14 000 users, and 146 000 feedback. Feedback is not binary, but on a 5-point scale; 1.7% are *below* 5.

Our user data corroborates anecdotal evidence that Paxful seems to attract a large proportion of customers from developing countries while LCS appears to attract more customers from western countries such as Europe and Australia (see Appendix C.1).

6.3.2 User Suspension

To maintain safety, both platforms restrict or suspend users who violate their terms of service (ToS). For Paxful, light violations (e.g., canceling a trade after its completion or using an outside app such as Telegram to conduct a trade without escrow) lead to restrictions being placed on the accounts. More serious transgressions lead to an immediate, permanent, and irreversible ban. Paxful lists four examples of such transgressions [287]: 1) using multiple accounts, 2) fake identities, 3) accessing from OFAC-banned countries [288], or 4) using unauthorized gift cards, reversing payment, and defrauding users. LCS ToS [289] strictly prohibits "spoofing trades" – i.e., self-promoting attacks – to protect the credibility of the reputation system. From each user page, we identify whether a user is suspended based on API responses. (See additional details in Appendix C.2). Surprisingly to us, as many as 46% of all Paxful vendors in our corpus who posted ads are suspended (24 562 users out of 53 224 until March 1, 2023). Throughout this study, we consider suspended accounts as "riskier" accounts (i.e., which have committed one of the heavy violations described above). Since we rely on the platform to label the risky accounts we will use in our machine learning model (§6.5 and §6.6), we perform several additional validations of label quality. We check that 1) account suspension is at least partially handled by humans (and not through a purely automated process) and 2) most suspensions are permanent bans. Appendix C.3 contains details. To evaluate the level of current moderation effort, we estimate how long the platform takes to find malicious accounts after the creation of accounts; 18% are suspended within a week, 48% within a month, and 83% within a year. We also measure how long it takes to unban accounts that turn out to be benign; 32% are unbanned within a week, and 68% within a month. Full details are in Appendix C.4.

6.3.3 Ethics and Legality

We collected data through publicly accessible APIs, abiding by both platforms' terms of service as of the end of data collection. In particular, we did not scrape websites. The same ToS prevents us from redistributing the data we collected to the public at

large, but we will consider requests for use of our data from academic researchers on a case-by-case basis. In any case, this chapter should provide enough information about our collection methods for interested parties to reproduce our work. Our data do not contain personally identifiable information, so our IRB does not consider this study human-subject research. Finally, following Martin and Christin [235]'s ethics guidelines for crypto-markets, our research does not pose any measurable risks to the researchers or any party using the platform.

6.4 Evaluation on Existing Reputation System

In this section, we evaluate the current feedback-based reputation system used by both Paxful and LCS. We center our analysis on two guiding questions, derived from prior work: (1) *Does feedback convey enough information for customers to recognize risky vendors?* (2) *Is the reputation system trustworthy or is it susceptible to manipulation, such as self-promoting attacks?* [13, 21].

To address these questions, we conduct two empirical evaluations. First, we demonstrate that numeric (i.e., scores) and textual (i.e., reviews) feedback left about vendors is noisy and does not convey sufficient signal to properly assess vendor quality. Second, we identify the instances of self-promoting attacks and distill public signals that significantly differ between suspended and non-suspended accounts across both markets. We leverage these findings to inform the development of our prediction model in §6.5.

6.4.1 Feedback Signals

We test whether the numeric and textual feedback conveys enough information for customers to discern potentially malicious accounts. Paxful shows the number of positive/negative feedback at the top of each user page. LCS displays the average feedback (on a five-point scale) on overall transactions for each user. To facilitate comparisons, we map these quantities to the $[0, 1]$ range.

First, feedback is skewed towards perfect scores, which makes it harder for customers to distinguish between good and bad vendors. 96.43% of Paxful (resp. 95.48% of LCS) users have a feedback score greater than 0.95; and 90.89% of Paxful (resp. 84.67% of LCS) users have scores greater than 0.99. In other words, getting *one* negative feedback out of 20 transactions suffices to drop a vendor to the bottom 5%. This is not unique to cryptocurrency marketplaces: 96.5% of transactions were rated 5/5 in the Silk Road anonymous marketplace [33], and 90% of vendors have 98% feedback scores or higher in eBay [21]. To mitigate this skewness, Nosko and Tadelis [21]

suggest EPP (Effective Positive Percentage), defined as the number of positive feedback divided by the number of total feedback. However, in Paxful, customers may conduct multiple transactions within a single listing, for which they can only leave one piece of feedback. Thus, EPP calculated on this platform is not comparable to that in previous literature.

We analyze the feedback text (i.e., reviews) next. We use the Google Translate API to translate into English approximately 8% of non-English reviews in languages other than English. Through manual inspection, we identify six categories of negative feedback:

1. **Scam accusations**: users sometimes explain fraud details, or simply call the vendor a scammer (e.g., *"He tried to rip me. Stay away from him," "Fake payment for payoneer invoice kindly don't trade this person. Return my amount 500 usd"*).
2. **Complaints about speed**: being slow or unresponsive also leads to major complaints (e.g., *"Not fast," "I regret trade with him. 6hrs??"*).
3. **Slander**: reviews that insult vendors without further details (*"Bad vendor," "Stupidity"*).
4. **False negatives**: positive reviews registered as negative (e.g., *"Goodd," "Positive," "++++++++"*).
5. **Quid-pro-quo**: ask/threaten trade partners to leave feedback in exchange for positive feedback (*"When you leave positive feedback I'll update mine," "selfish fello who doesn't leave a feedback after trade"*).
6. **Unclear/other**.

To quantify the ratio between those categories, two of the authors independently manually labeled categories for 500 randomly selected negative reviews. For Coder 1 (Coder 2), 55.4% (46%) are scam accusations, 12.6% (10.4%) are about speed, 14.6% (22.4%) are slandering, 5.2% (5.0%) are apparent false negatives, 5.2% (3.2%) are quid-pro-quo, and 6.6% (13.0%) are others, respectively. The Cohen Kappa statistic [290], the agreement between two coders, is 0.706, which is considered "substantial agreement." Interestingly, even when manually annotating the data, extracting a clear signal from the text (or verify the credibility of reviews) is difficult, as observed by the disagreement between coders. In particular, coders had the most disagreements judging scam- and slander-related feedback. Furthermore, negative feedback tends to attract replies that rebut the reviews (e.g., *"As if it was very difficult to do what you did, you are very smart to make other people look bad"*). Indeed, 19.22% of negative reviews get a reply (compared to 0.71% in all comments), which implies that some negative reviews may be a form of retaliation or attempts to taint the reputation of competitors through a "badmouthing attack" [274] (e.g., *"stay away from him he will destroy your reputation he will mess you up after successful trade"*). As noted by the high skew

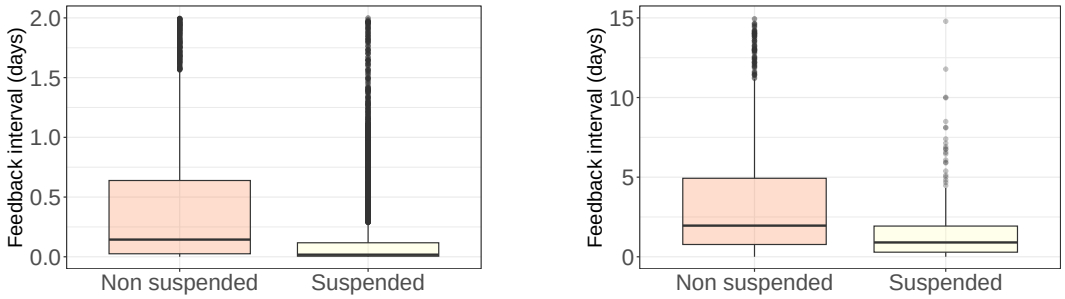
of reputation scores, such retaliation attempts may be particularly effective against otherwise reputable vendors. Our observations suggest that obtaining a clear signal on the quality of a vendor either through their numeric reputation score and/or the reviews is difficult. This is exacerbated by the fact that customers may often prefer to leave feedback outside of the platform due to retaliation fears [1] and employ external avenues, such as forums [38], or not even leave feedback at all due to the lack of economic incentives [284]. In our study, we observed users posting reviews on Reddit (e.g., `/r/Paxful`), Telegram (e.g., LCS Telegram channel), or even leaving negative reviews in app stores.

6.4.2 User Collusion and Automation

Our manual investigation reveals a set of accounts that exhibits the following traits. 1) More than hundreds of accounts are giving feedback together repeatedly. 2) Many *positive* feedback messages are submitted in a short range of time. 3) They reuse a similar set of simple feedback messages (e.g., *"Excellent trader very fast.," "Good and quick"*). 4) They appear to arbitrarily pick rare payment methods, that are not currently in use in the account's origin country. 5) Many accounts share the exact same number of trade counts. Appendix C.6 describes the details, but this analysis is inspired by Fusaro et al. [291] that illustrates the unnatural distribution of trade volume as a sign of "wash trading" (creating fake trades by selling items to oneself to give the appearance of larger volumes) in centralized cryptocurrency exchanges. Based on those characteristics, we believe that these accounts engage in self-promoting attacks. Not only do these patterns suggest that existing reputation systems may be easily manipulated, they also hint at features that may be indicative of risky accounts. We next look at features that suggest manipulative behaviors such as collusion and automation.

(1) **Interaction networks between accounts**: Consistent with other studies of online financial communication platforms [14], our data show that suspended accounts comparatively interact more frequently with other suspended accounts. Only 16.26% of the feedback from *non-suspended* accounts is directed towards suspended accounts, whereas 24.82% of the feedback from *suspended* accounts goes to other suspended accounts. Around 300 000 reviews – or 6.3% of all feedback observed – are generated between suspended accounts. This result motivates us to incorporate the information from neighboring users based on feedback interaction.

(2) **Feedback to transaction ratio**: suspended and non-suspended accounts also have unique differences in their feedback rates. Benign accounts often receive little feedback compared to the number of transactions they conduct—a predominant phenomenon across online marketplaces [1]. However, suspended accounts boast an un-



(a) Paxful (b) LCS

Figure 6.5: Feedback interval (days) boxplots. We restrict the y-axis ranges for better visualization.

naturally higher feedback rate. For instance, among accounts with 10 trades, 50.4% of suspended accounts received feedback for every transaction (i.e., they got 10 reviews); only 8.4% of non-suspended accounts were in that position. For 200 transactions, 6.7% of suspended accounts still get feedback for every transaction, whereas the number drops to 0.36% among non-suspended accounts. Thus, high feedback rates suggest possible user collusion.

(3) **Feedback interval**: Previous literature has shown that bots have a very different posting behavior from legitimate users [292], We investigate how frequently each user receives feedback. We define *feedback interval* as the median time between two consecutive reviews. We exclude accounts with less than 10 feedbacks due to noise. Figure 6.5 shows that suspended accounts received feedback far more frequently than non-suspended accounts. We confirm the statistical difference ($p \ll 0.01$) between the two groups for both markets by the Mann-Whitney U test (robust to outliers in our data). As an extreme example, one Paxful user received feedback every 4–5 seconds, which raises strong suspicions of automation.

6.5 Predicting Account Suspension on Paxful

The above results answer the two questions posed earlier: existing reputations convey insufficient signals to determine the quality of accounts, and are manipulated by user collusion and automation. Furthermore, there are significant differences in features besides feedback scores between suspended and non-suspended accounts. This suggests that other public signals, not captured by current reputation systems, can characterize problematic accounts. We next rely on these features to design a classifier, which can predict which vendors are suspended on Paxful. (We will discuss LCS in the next section.)

6.5.1 User Features

We derive user features from four sources: user profiles/statistics, ads, and feedback. Feature selection is informed by our exploratory analyses in the previous section and by related work [14, 292, 293]. User profile and statistics include the number of users blocked by/blocking/trusted, registration time (Appendix C.4), registration country (given by IP address), number of trades, trade volume for each currency (BTC, USDT, ETH), number of trade partners, number of positive/negative feedback. We also keep track of users who access the platform from countries different from where they initially registered (Appendix C.1).

For listings, we aggregate all the collected ads at the user level (e.g., posting 60% of ads in USD makes "ratio of USD in ads" variable equal to 0.6). An important feature derived from user ads is the price premium, defined as the difference between the advertised price and the market price, i.e., Price premium $= \frac{\text{Proposed price} - \text{Market price}}{\text{Market price}}$ Prior work on Craigslist has found that scammers often set unreasonably low-price premiums [76, 293, 294], which motivates using it as a feature. Other ad data include timezones (based on the city listed in the ad), payment method (e.g., bank transfer, PayPal, Amazon gift cards), types of fiat currencies (e.g., USD, EUR, KES), cryptocurrencies (e.g., BTC, USDT), any customer verification requirements, and whether users are marked as "verified" by the platform [295]. We further compute feedback interval (§6.4.2), and incorporate the negative feedback content identified by keyword search (see §C.5).

Finally, we rely on the feedback graph (where each node is a user and a directed edge $A \rightarrow B$ is feedback from user A to user B) to include neighbor information. Since feedback is not mandatory, the feedback graph is a strict subset of the entire trade graph. From this graph, we derive network metrics such as ego density and some centrality measures to incorporate how they interact with others, and how influential they are. Importantly, users are allowed to change their username *only once* on Paxful. We keep track of those changes and reflect them when we aggregate all the features. We normalize all features (mean 0, std. dev. 1) to stabilize model training, except for binary variables and features already in $[0, 1]$.

6.5.2 Machine Learning Models

Using the labels described in §6.3.2, we build a machine learning model to classify suspended accounts between suspended (24 562) and not-suspended (28 662). Our model construction is inspired by prior bot detection work (e.g., Davis et al. [292]). We implement seven machine learning models: Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost [296], LightGBM [297], and Neural Network, using Python *scikit-learn* to compare their performance. We choose those

Table 6.1: Prediction results: seven models with CI (2.5%, 97.5%).

	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.769 (0.758, 0.781)	0.768 (0.757, 0.78)	0.767 (0.755, 0.778)	0.767 (0.756, 0.779)	0.849 (0.838, 0.859)
K-Nearest Neighbors	0.775 (0.764, 0.786)	0.779 (0.768, 0.79)	0.779 (0.768, 0.79)	0.775 (0.764, 0.786)	0.86 (0.85, 0.87)
Decision Tree	0.818 (0.808, 0.829)	0.817 (0.807, 0.828)	0.818 (0.808, 0.828)	0.818 (0.807, 0.828)	0.879 (0.87, 0.889)
Random Forest	0.856 (0.847, 0.866)	0.859 (0.849, 0.868)	0.853 (0.844, 0.863)	0.855 (0.845, 0.865)	0.931 (0.924, 0.937)
XGBoost	0.862 (0.853, 0.871)	0.862 (0.853, 0.871)	0.861 (0.851, 0.87)	0.861 (0.852, 0.87)	0.935 (0.929, 0.941)
LightGBM	0.861 (0.852, 0.87)	0.862 (0.852, 0.871)	0.859 (0.85, 0.868)	0.86 (0.851, 0.869)	0.932 (0.925, 0.938)
Neural Network	0.825 (0.815, 0.835)	0.824 (0.814, 0.834)	0.826 (0.816, 0.836)	0.825 (0.814, 0.835)	0.903 (0.895, 0.911)

seven models because they are standard interpretable models (explainability is very important in this experiment). We use grid search with 5-time cross-validation to tune model hyper-parameters/architectures (e.g., the level of regularization, the depth/number of trees, and the number/dimensions of neural network layers). We divide the entire dataset into 80% training/validation set and 20% test set, and use the test set to conduct out-of-sample prediction and compare performance. Table 6.1 summarizes the results of each model for accuracy, precision (macro), recall (macro), F1-score (macro), AUC (area under the curve), with a threshold of 0.5. Ensemble-based tree algorithms (Random Forest, XGBoost, and LightGBM) outperform other methods, achieving 0.86 F1 and 0.93 AUC. To draw statistical differences between models, we randomly pick 50% of test sets, bootstrap for 10 000 times and derive the (2.5%–97.5%) confidence intervals (CI) shown in parentheses in Table 6.1. For example, for Random Forest, the F1-score falls in the 0.842–0.862 range for 95% of bootstrapping. Based on it, we conclude that the three ensemble tree-based algorithms (Random Forest, XGBoots, LightGMB) perform equally well while significantly outweighing the others.

To delve into how our model identifies risky accounts, Table 6.2 (Paxful: first column) highlights the top-10 most important features for tree-based ensemble models. This is calculated based on how good the split is ("gain") when using each feature. The most important source of information is the number of accounts the user is blocking, which is a good proxy for how adversarial the account is. Models also seem to rely on various sources of data including user profiles (registration time), trade statistics (number of positive feedback, number of trade partners), ads information (price premium, currency), and network metrics from feedback graphs (ego density). Some

Table 6.2: Top 10 most important features for Paxful (§6.5) and LCS (§6.6) categorized by data source. Number in parentheses is the feature importance rank.

	Paxful	LCS
User profile	Number of user blocking (1) Registration time (2) Number of users trusted by (9)	Registration time (1) Number of users trusted by (8) Number of users blocked by (10)
Trade statistics	Number of trades (3) Ratio of positive feedback (5) Number of trade partners (6)	Number of trade partners (3) Number of trades (4) Average response time (7)
Ads	Price premium (4) Ratio of USD (10)	
Feedback	Ego density (7) Total degree (8)	Eigenvector centrality (2) Total degree (5) Ego density (6) Feedback receiving interval (9)

of the features, e.g., pricing strategies [298] and ego density [14] were found to be characteristic of suspicious accounts in previous literature. A number of trades, positive feedback, feedback interval (ranked in the top-15 features), and network metrics are frequently associated with user collusion and feedback automation (§6.4.2). "Verified" user badges, on the other hand, have little impact on our model's decision-making (not in the top-50 features); this echoes other studies [112, 299]. In short, integrating multiple sources of public information, rather than merely assessing reputation through feedback scores and/or badges, appears desirable.

6.5.3 Evasive Measures

Our machine learning model presents a few limitations that malicious participants could potentially exploit.

First, assuming that an attacker knows the detailed implementation of our machine learning model, they can control some parameters to avoid detection. For example, they can avoid using certain types of payments (e.g., PayPal, M-Pesa), or types of currencies or coins (e.g., USD, KES, BTC). An attacker could use a VPN to obfuscate their location (see Appendix Figure C.2) if they are aware that the model tends to pick more users from a certain country. Our model also fails to capture users who rely on new or unpopular types of payment, currencies, or locations. On the other hand, changing those would make it much harder for an attacker to attract legitimate customers. In other words, evasion, while possible, could come at a potentially hefty price to the attacker.

Second, some features (e.g., the number of users being blocked) may slowly evolve, and a malicious participant could exploit the time lag before they get flagged. However, this latency also applies to feedback-based reputation (feedback comes later), and our model is less susceptible to it since it combines multiple features.

Third, the model is vulnerable to whitewashing attacks [13]. If a scammer creates a new account to purge their entire history, the model will fail to identify them, at least initially. However, this too comes at a cost: reputation needs to be rebuilt from scratch.

6.6 Generalizing the Model Across Markets

To test the generalizability/transferability of our models across platforms, we repeat the previous experiment beyond Paxful, varying features/training sets (Paxful vs. LCS), and prediction targets (Paxful vs. LCS as well) to generate six different models (Model 1–6) for testing. Model 1 is the model described in the previous section as our baseline. For simplicity, we limit our use to Random Forest (one of the best-performing models in Table 6.1) in this section. Table 6.3 summarizes our results for these six models as described below.

From historical feedback LCS data, we extract 11 657 accounts. For those, we check the user page status and find 1 547 (13.27%) suspended accounts. In LCS, account information becomes unavailable after users get suspended. As a result, we can only collect user profiles for 167 suspended accounts. To account for this data loss, we downsample the non-suspended accounts to keep the suspended and non-suspended ratio identical (13:87) to the original population. We repeat the same procedure described in §6.5 for LCS, and use the data prior to March 1st, 2023 to temporally align with our Paxful experiment. Since available user information differs from what Paxful provides, we use different features in LCS such as the average response time and primary currency/language. Model 2 is identical to Model 1, but independently trained and predicted on LCS. Model 2 does not achieve the same performance level as Paxful (Model 1). This is probably due to the smaller number of data samples, fewer features, and imbalanced label distributions. To test model generalizability, we then only use features common to both platforms. These include feedback interval, trade counts, negative feedback ratio, the number of trade partners, and network metrics such as ego density on the feedback network. We do not normalize features. First, we re-train the model with those common features on Paxful data (Model 3) and on LCS data (Model 4). Model 3 does not quite manage to match the performance of Model 1; on the other hand Model 4's performance is roughly the same as Model 2's, depsite the smaller number of features. This indicates some features only (publicly) available in Paxful, such as the number of users being blocked by the user, are crucial

Table 6.3: Performance results for two markets (Susp. = num. of suspended accounts, Acc. = Accuracy).

Model	Data					Performance				
	Training	*Features*	*Prediction*	*Test size*	*Susp.*	*Acc.*	*Precision*	*Recall*	*F1*	*AUC*
1	Paxful	All	Paxful	10645	4935	0.858	0.860	0.855	0.857	0.931
2	LCS	All	LCS	260	38	0.869	0.773	0.596	0.624	0.684
3	Paxful	Common	Paxful	10645	4935	0.723	0.723	0.719	0.719	0.791
4	LCS	Common	LCS	260	38	0.858	0.712	0.557	0.567	0.638
5	Paxful	Common	LCS	1300	169	0.840	0.600	0.566	0.576	0.659
6	Paxful	Common	LCS	367	169	0.594	0.647	0.565	0.510	0.632

to performance.

Finally, to test model transferability, we first train the model using Paxful data, freeze the model weights, and make predictions for all the users on LCS (i.e., using both train and test data on LCS) (Model 5). Since Paxful has a larger number of samples than LCS, we should observe a performance increase in LCS if we can successfully transfer some knowledge from Paxful. Unfortunately, the performance does not significantly improve from simply training on LCS independently, which means that, within our dataset, the model does not appear to be directly transferable from Paxful to LCS. To explain why, we consider three factors. First, the proportion of suspended accounts is 46% in Paxful but 13% in LCS, so the model might have been confused. To test this conjecture, we downsample the non-suspended accounts to keep the ratio identical to Paxful's (Model 6), but do not observe any increase in performance. Second, the user base is markedly different (see §6.2.1, §6.3.1, and §C.1). Third, both platforms operate at different scales. Paxful has *at least* 4.7 million reviews whereas LCS has only about 146 000 reviews; however, feature normalization does not alleviate this issue. On a more positive note, we find, in Table 6.2, that some features, indicative of risky accounts, are important to both platforms, such as network metrics (ego density) and trade statistics (number of trade partners, trade counts, feedback interval).

6.7 Prospective Cohort Study

Our model can be used in a variety of ways. It can be used to flag risky users and dedicate—often scarce—moderation resources to them. It can help audit user interfaces and inform their redesign. It can also power a browsing addon that alerts users about risky accounts, given that it is trained on public information. To demonstrate the utility of our model, we conduct an online evaluation on Paxful, where we monitor users for 30 days after the model prediction. This study is known as a prospective

cohort study, where we observe users that are not suspended at the beginning of the study, and check whether they are suspended later. We compare the model's performance against two baselines: 1) a reputation-based rule, where we select users with the lowest reputation scores, and 2) a random selection of users. We expect our model to outperform both baselines, as it uses richer features than reputation scores alone.

6.7.1 Experimental Setting

We perform an online evaluation from March 1–30, 2023. From the active users as of March 1st, we create three sets of 500 users each: 1) the "riskiest" users based on our machine-learning model prediction "ML," 2) users with the lowest reputation, "REPUTATION (REP)," and 3) randomly chosen users, "RANDOM (RND)." For ML, we define the "riskiest" users as those that have not been suspended yet, but our model predicts will be suspended with the highest probability. We train our model using data until February 28th, 2023. For REP, we choose the users with the highest ratio of negative feedback with at least 10 reviews. We check each user at least once every day for suspension and trade count. There is some overlap in users between each group, so, to keep independence assumptions, we exclude these common users when performing statistical tests. We set the p-value statistical threshold to 5%, and apply the Bonferroni correction to account for multiple hypothesis testing.

6.7.2 Results and Implications

Around 20% of ML-selected users (95/500) were suspended within 30 days, compared with 46/500 in REP and 30/500 in RND. Pairwise χ^2 tests show ML flags significantly more suspensions than both baselines (ML-REP χ^2 =20.14, p¡.001; ML-RND χ^2 =40.98, p¡.001), while REP and RND do not differ after Bonferroni correction (χ^2 =3.21, p=.073). In short, a feedback-only reputation rule performs little better than random, whereas a classifier using richer features provides a far more accurate early warning signal.

Next, we discuss the timing of user suspension. Figure 6.6 shows the number of suspensions over time for each group, that is the number of users initially active on March 1, 2023, that are later suspended. We then define the suspension of an account as a "death" event and compute the survivability curves. Log-rank tests confirm ML differs significantly from both REP (χ^2 =21.99, p¡.001) and RND (χ^2 =43.44, p¡.001), while REP and RND are indistinguishable after correction (χ^2 =3.63, p=.057). In other words, "risky" users according to the ML prediction are much more likely to be suspended soon. This result suggests that our model is able to identify risky accounts that have not yet been flagged by the platform (i.e., false negatives) earlier.

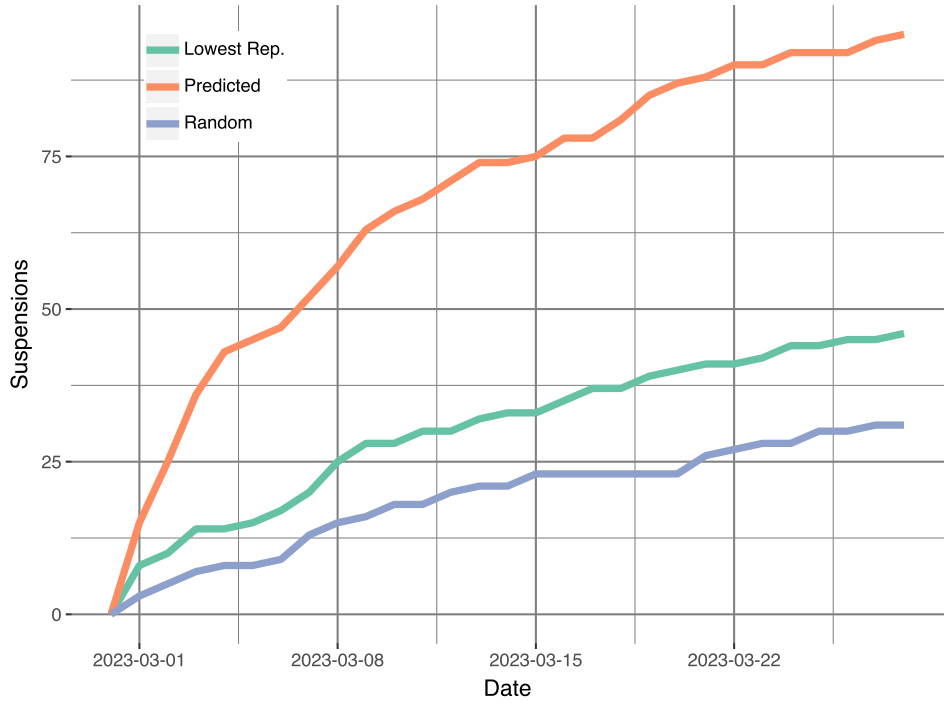


Figure 6.6: Suspensions on Paxful over a one-month period for three user groups (each $n = 500$): those flagged by our classifier as most likely to be suspended, those with the lowest reputation scores, and a random sample. The classifier-based group shows substantially higher suspension counts.

Besides suspensions, we measure the number of trade completions, which is a good proxy of how successful and active users are. We conjecture that there is a negative impact from a low reputation score on the amount of trade. To account for the fact that some users get suspended in the middle of the observation period, we divide the total number of completed transactions during this experiment by the number of active days over the month. Using a t-test, we find a significant difference between RND and the other two groups, indicating that risky users from the ML model and the low reputation group complete fewer transactions. RND users averaged 38.83 trades/day, significantly above both ML (16.13; ML-RND t=-2.76, p=.006) and REP (9.15; REP-RND t=-4.29, p¡.001), while ML modestly exceeded REP (t=2.04, p=.041). In other words, Although our machine learning model is optimized to predict account suspension, it can, to some extent, identify unsuccessful vendors. Unsurprisingly, users with poor feedback tend to be less successful on the market too.

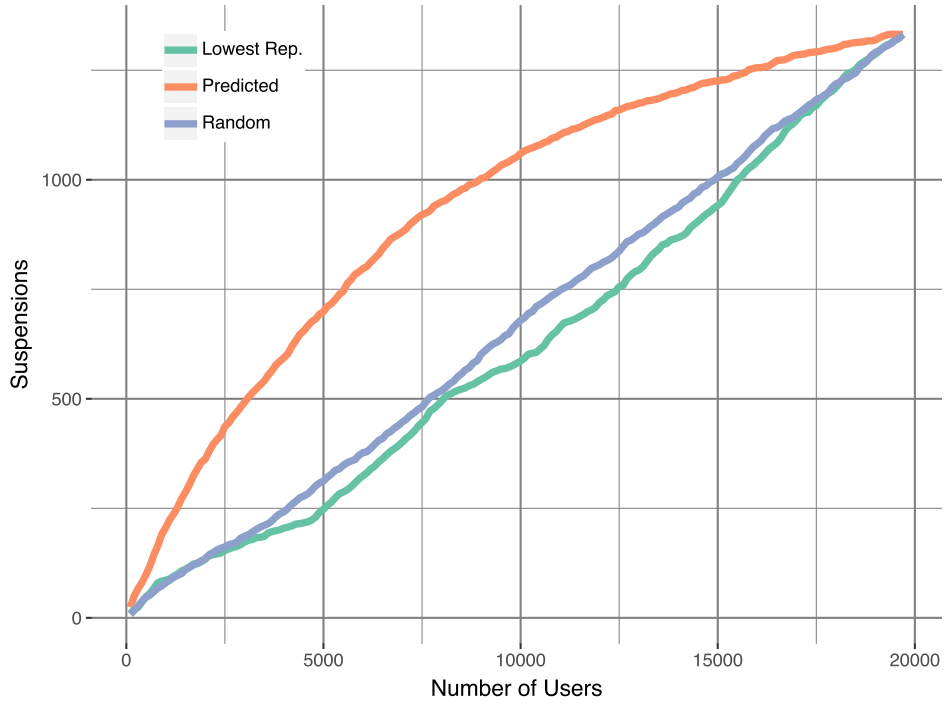


Figure 6.7: Suspension rates on Paxful when varying group size n. The figure compares the three user groups described in Figure 6.6, demonstrating that classifier-selected users consistently show higher suspension rates across group sizes.

6.7.3 Robustness

To validate the robustness of our method, we perform three additional experiments to check 1) how much variance exists when randomly picking users (RND group), 2) whether the result changes using a different timeframe, and 3) the optimal number of users to pick (i.e., not fixing it to 500 users). All the experiments confirm the superiority of our ML method.

In §6.7, we randomly pick 500 accounts out of over 28 000 active accounts for the RND method, but we do not know how the results vary depending on the users we pick. To address this, we randomly draw 500 accounts, count the number of suspensions in a month, and repeat the process 10 000 times to check for any deviations in the results. More than 95% of the time, 18–39 out of 500 RANDOM accounts end up being suspended; in other words, our results in Section 6.7 about the significantly superior performance of ML/REP holds across many RANDOM samples.

Second, we perform the same online evaluation on a different time period: 30 days from February 1st, 2023. Our ML model is trained using only data before January 31st, 2023. Our analysis is consistent with the result presented in the main body —for a time interval starting on March 1st, 2023. The ML group had the highest number of

Table 6.4: Statistical test results comparing group pairs. Entries show test statistic and p-value.

Group Pair	Susp.	Survival	Trade
Predicted vs. Lowest Reputation	20.14 (0.000)	21.99 (0.000)	2.04 (0.041)
Predicted vs. Random Baseline	40.98 (0.000)	43.44 (0.000)	–2.76 (0.006)
Lowest Reputation vs. Random Baseline	3.21 (0.073)	3.63 (0.057)	–4.29 (0.000)

suspensions (104) and moderate trade activity (10.78 trades per day). The RP group saw fewer suspensions (49) and lower trade activity (6.09), while the RN group had the fewest suspensions (26) but exhibited extremely high trade activity (159.99), likely due to outliers. Statistical tests show that all pairwise group comparisons are significant ($p < 0.05$) across suspension count, trade volume, and survival outcomes. The ML method consistently identifies significantly more suspended users than the other two groups, targets more active traders than RP (but fewer than RN), and demonstrates strong separation in survival analysis. Overall, the results confirm that the ML method continues to outperform baseline approaches in identifying high-risk, active users.

Third, in our online evaluation in §6.7, each group is the 500 riskiest/low reputable/random users. Here, we calibrate the number from 100 to 28 000 users. In other words, we monitor x riskiest/low-reputation accounts and vary x instead of fixing $x = 500$, and quantify the impact of x on the number of suspensions. Figure 6.7 shows the number of total suspended accounts in a month (y-axis) based on the number of users monitored (x-axis). Obviously, if each method selects all active 28 000 vendors, all (ML, REP, and RND) methods have the exact same number of suspensions (i.e., the right top of the figure). However, the figure clearly illustrates that ML outweighs other methods regardless of the number of users monitored. It works best until around 10 000 users. Depending on the number of moderators the platform employs, they can adjust the number of users being monitored. The advantage of using our ML method marginally decreases when a large number of users are monitored.

6.8 Related Work

This section relates our work with previous efforts on 1) cryptocurrency P2P exchanges and 2) online misbehavior in other platforms, and highlights the novelty of our research.

The cryptocurrency P2P marketplace landscape largely remains understudied. In LocalBitcoin/Paxful, Von Luckner et al. [276] identified many transactions as re-

mittances from the US to developing countries. Andreianova et al.'s survey [300] further clarified that many users from Latin America use P2P platforms for remittances, whereas users in Africa use the platform for trading/profit generation. Van de Laarschot and van Wegberg [199] connect online anonymous market vendors to major P2P cryptocurrency exchanges. However, despite their relevance, no prior work has evaluated the online safety of cryptocurrency P2P markets.

On the other hand, some empirical studies look at scams in other marketplaces such as Craigslist [76, 293, 294, 298, 301]. A common detection approach is to use the platform-provided labels [76, 301] and complement them by unrealistically low price premiums [293, 298], or directly interacting with suspicious accounts [294]. We choose the first approach to discover suspicious accounts and perform some validations to confirm label quality (i.e., a low number of false positives), and additionally extend our analysis to multiple markets for generalization. We further develop a platform monitoring scheme to prove the practicability of our method as well.

Another related line of work revolves around user misbehavior on social media platforms, particularly social bot detection [53, 98, 292, 302] using machine learning on large-scale data [53]. In particular, our work adopts a similar methodology to Davis et al.'s work [292] on feature selection and algorithms. Others have studied account suspension [303, 304], and shown that fake/suspended accounts form closely knit communities [14, 305–307], which our study confirms.

6.9 Design Insights

Our findings suggest that traditional reputation and trust indicators—such as ratings and verification badges contribute little to predicting vendor suspensions. Instead, behavioral and structural profile signals provide more reliable indicators of misconduct risk. This has implications for platform design and primarily draws from the literature on the design and presentation of security indicators [308, 309]. First, we recommend that platforms use statistical models to audit user interfaces and identify which features are most predictive of misbehavior. This can help platforms prioritize which signals to surface to users, rather than relying on legacy reputation widgets that may not be effective. Another option is that, rather than centering legacy reputation widgets, platforms could prioritize predictive and decision-relevant cues, surfacing them at the point of decision (e.g., search results, checkout, profile visit) to better support consumers in evaluating counterparties.

These insights rely on careful risk communication. The literature on security indicators has found security indicators to be useful to warn users about potential risks when visiting websites or installing software [308–310]. This body of work emphasizes that interventions are most effective when they are rare, specific, and action-

able [308, 309]. Certainly, integrating model-driven signals directly into workflow defaults——such as pre-sorting or filtering out high-risk vendors——can reduce reliance on warnings altogether. However, users and attackers in these marketplaces are constantly evolving. Users develop new heuristics to identify risky vendors, and attackers adapt to evade detection and attract users. In light of this dynamic, we recommend that platforms incorporate lightweight, adaptive risk indicators that are updated over time as new modeling takes place.

Additionally, the literature on explainable AI underscores that explanations should help calibrate, rather than maximize, user trust [311]. Explanations that are contrastive or example-based can make model judgments intelligible without overwhelming users or exposing precise thresholds that could be gamed. Platforms could adopt simplified risk categories with accompanying confidence indicators, paired with plain-language micro-explanations about why a vendor may pose greater risk [311].

In summary, we see this statistical modeling approach as a step towards identifying signals that would be useful to highlight and integrate in online platforms' user interfaces. However, this procedure does not replace usability testing; rather, it enhances it. That is, given a signal that is predictive of misbehavior, we should investigate how to best present it to users and evaluate it in practice. Through this process, our modeling tool can reduce the set of possible signals and also identify signals which have experienced decay in predictive power over time.

6.10 Takeaways

This chapter investigates online misbehavior in cryptocurrency P2P marketplaces. We outline the limitations of solely relying on feedback-based reputations and attempt to build a better system for uncovering risky vendors. Using only publicly available data, our model achieves 0.93 AUC in identifying account suspension in Paxful, one of the most active cryptocurrency P2P marketplaces. We expect the performance would increase with access to private information such as IP addresses, especially on a smaller platform like LCS. We could not replicate our experiments on other platforms such as Binance P2P, which do not provide indicators of account suspension. However, in practice, any marketplace can follow the same procedure and incorporate features we identified as important. We further provide a framework to improve platform moderation. Instead of directly banning the accounts the model identifies, we suggest selecting the set of accounts with the highest likelihood of suspension and prioritizing them for monitoring.

Our results also benefit users. Our study shows users should review various types of features besides feedback, such as price premiums, and who is giving feedback.

Note that our model does not protect users after they initiate transactions (i.e., only help identify risky vendors as a precautionary measure). After starting a trade, platforms recommend users verify payments in addition to the receipt sent by counterparties, take screenshots frequently to gather evidence, and avoid outside channels to communicate [312].

More generally, our work helps broader research on other online marketplaces that remain understudied (e.g., gift cards, NFT, online loans). Those platforms rely on reputation systems and face issues similar to what we observe. Another research area lies in reputation system design (i.e., how to convey the risks associated with vendors) since the way the platform aggregates/presents reputation scores significantly affects user behavior [284].

Chapter 7

Conclusion

This dissertation has examined the role of profile signals in shaping trust, behavior, and security across online platforms. Through a series of empirical studies, I showed that while profile signals are central to user experience and platform governance, their informativeness varies widely. In many instances, reputation signals and trust indicators failed to achieve their theoretical goals and proved to be poor predictors of risk or quality. In contrast, a statistical modeling approach uncovered behavioral and structural signals with stronger predictive power, allowing us to model how users adapt their decision-making strategies based on available cues—and how platforms must likewise adapt to improve both user experience and security.

The central contribution of this work is the development of a statistical modeling framework to evaluate and audit profile signals. The persistence of legacy reputation widgets, such as the star rating systems introduced decades ago, illustrates the need to revisit the design of reputation interfaces. Rather than treating interface elements as static artifacts, I demonstrate how they can be evaluated for predictive validity against outcomes such as financial performance, vendor exit, and account suspension. This approach bridges traditional usability testing with computational evaluation, offering a scalable way to identify which signals remain informative, which have degraded over time, and which deserve greater prominence in user interfaces. Platforms should incorporate statistical auditing into the lifecycle of interface development, deprioritizing cues with low predictive validity and surfacing new signals for usability testing. Effective indicators must be informative, transparent, and contextually relevant, drawing from research on security warnings and explainable AI to ensure that users can interpret and act on them.

At the same time, this work shows that profile signals can both mitigate and enable security risks. On the one hand, they can help curb abuse and encourage positive behavior. On the other, they can mislead users, exposing them to fraudulent or risky actors, as observed in darknet marketplaces and peer-to-peer cryptocurrency

platforms. In both settings, reputation scores and trust badges were prominently displayed but had little correlation with vendor quality, longevity, or misconduct. Worse, dishonest users can manipulate these signals to their advantage, as in the case of YouTube, where accounts with misleadingly attractive signals were repurposed to spread harmful content. As we demonstrated, emerging secondary markets that facilitate trade of such accounts only increase these risks.

This dissertation offers several practical recommendations for the stakeholders of each of the projects studied. For YouTube, we demonstrate how account repurposing occurs, how prevalent it is, and its impact. For law enforcement agencies and policymakers, we offer various methodological advances to more accurately measure darkweb marketplaces and better prioritize targets for intervention. For peer-to-peer cryptocurrency platforms, we identify key behavioral signals that can help distinguish high-risk from low-risk users and the various techniques that dishonest users may employ to manipulate reputation systems in their favor. However, these findings are not limited to the platforms studied here. The insights can be applied to other social media platforms where account repurposing is possible and for which there are accounts for sale. Our measurement techniques also extend to e-commerce platforms that can be publicly scraped. And, the modeling framework can be applied to any platform with user profiles and outcomes of interest. Taken together, all these recommendations can help a variety of platforms improve user experience and security.

Beyond platform design, this work carries implications for policy and consumer protection. We need greater transparency for accounts with substantial reach, particularly on social media platforms where accounts with misleading signals can distort public discourse and amplify harmful content. This concern should matter to both regulators and platform operators, as access to accurate, decision-relevant information is a core consumer protection right. By introducing new signals grounded signals in statistical modeling and making them more transparent, policymakers and platforms can reduce users exposure to fraud, harassment, and disinformation. For platforms unwilling to act, the methods developed in this dissertation can also support motivated third parties in building independent tools and services that serve the public interest.

Ultimately, this dissertation advocates a paradigm shift: profile signals should not be static but dynamic, continuously evaluated indicators. As malicious actors adapt and new risks emerge, platforms must evolve not only their backend detection algorithms but also the interfaces that users rely on. By embedding continuous auditing and model-driven evaluation into the design of reputation systems, online ecosystems can become more resilient and trustworthy. In doing this, we can align the signals users rely on with the integrity and safety that platforms aspire to uphold.

Appendix A

Chapter 3 Prompts and Qualitative Coding Materials

A.1 LLM Prompts

We used the following prompts to annotate channel repurposing (Figure A.1) and topic categorization (Figure A.2). Note, for the topic categorization prompt, we used a few-shot prompting strategy, where we provided the LLM with examples of each topic to help it understand the classification task. For presentation purposes, we provide the few-shot examples in Figure A.3. However, in practice, each of these examples was provided in the prompt as an example, where indicated

A.2 Qualitative Coding Guide and Template

To derive the categories for channel categorization, we used a qualitative coding approach. We developed a coding guide (Table A.1) that outlines the categories and definitions used for channel categorization. This codebook was created by human coders by labeling a subset of channels, as described in Chapter 3. To ensure that human coders reviewed YouTube channels in a consistent way, with access to the same information, we generated a markdown document for each channel, which included the channel's handle changes, a timeline of the channel's profile information over time, and all videos observed for the channel (Figure A.4). The coders were instructed to read through this document and use the coding guide to categorize the channel.These categories were then used to make our LLM categorization prompt, shown in Figure A.2 and to identify few-shot examples for the prompt, shown in Figure A.3.

Table A.1: Coding guide for channel categorization.

Did the channel contain:	**The channel discusses or contains videos:**
Politically-related content?	Describing contemporary political subjects, events, or figures.
News-related content?	Describing news reports and/or world events emulating traditional newscast.
Health-related content?	Describing medical or health-related subjects, events, or figures.
Gambling-related content?	Describing gambling and/or betting subjects, events, or figures, including sports gambling and online casinos.
Content that may infringe copyright?	Describing software, shows, movies, or music that they do not own, as well as advertising links or websites that contain these materials.
Content geared towards children?	Oriented towards very young audiences. For example, they contain videos about nursery rhymes and kids shows.
Alternative forms of monetization?	Containing information on how to purchase products/services from an external site, business inquiry emails and phone numbers, links to WhatsApp or Telegram groups advertised to make money.
Cryptocurrency-related content?	Describing cryptocurrency subjects, events, or figures.
Money-making content?	Describing investing and/or financial subjects, events, or figures, including trading, digital marketing, and e-commerce.
Religious content?	Describing religious subjects, events, or figures.
AI-generated content?	Showcasing content created with generative AI tools, e.g., Midjourney, Sora, Runway, etc.
Manosphere content?	Discussing manosphere or redpill topics, events, or figures.
Extremist content?	Showcasing toxic content or content from extremist groups, such as white supremacist, jihadist, neo-nazi, alt-right, etc.

Prompt: Channel Repurpose Annotation

You are an AI that analyzes pairs of YouTube handles and titles to determine if they belong to the same entity. Respond with only a JSON object containing one number: `"is_same_entity"` (0 or 1).

Rules for classification:

- If the handles are significantly similar, set `is_same_entity` to 1 and `reasoning` to `"handle"`.
- If the titles are significantly similar, set `is_same_entity` to 1 and `reasoning` to `"title"`.
- If the descriptions are significantly similar, or if the `new_description` references the same entities (e.g., URL, email, etc.) as the `old_description`, set `is_same_entity` to 1 and `reasoning` to `"description"`.
- Otherwise, set `is_same_entity` to 0 and `reasoning` to `"none"`.

Inputs:

- `old_handle`: The old YouTube handle.
- `new_handle`: The new YouTube handle.
- `old_title`: The old channel title.
- `new_title`: The new channel title.
- `old_description`: The old channel description.
- `new_description`: The new channel description.

Output format:

```
{
    "is_same_entity": 1,
    "reasoning": "title"
}
```

Figure A.1: Prompt used for channel repurposing annotation.

Prompt: Topic Annotation

You are analyzing YouTube channels and their video content to classify them based on specific topics.
Given a dataset containing a YouTube channel's title, description, and a list of video titles and descriptions, extract whether the channel title, channel description, or its videos discuss any of the following topics:

1. **Non-YouTube Monetization:** The channel provides links to external sites for purchasing products/services. The channel contais email addresses or phone numbers for business inquiries. The channel mentions WhatsApp or Telegram groups advertised for making money. Exclude links to YouTube, Twitch, Tiktok, Facebook, and Insgtagram. *Few-shot Example 1*.
2. **AI-Generated Videos:** The channel contains AI-generated videos as indicated in video or channel metadata, such as `#AI`, `#AIChannel`, `#AIGenerated`, `#DALLE`, `#Midjourney`, `#StableDiffusion`, etc. *Few-shot Example 2*.
3. **Political Content:** The channel mentions political subjects, events, or figures in the channel title, URL, description, or video content. Content about any national military should be marked as political. *Few-shot Example 3*.
4. **Religious Content:** The channel discusses religious subjects, events, prayers, or figures in the channel or videos. *Few-shot Example 4*.
5. **News Content:** The channel emulates traditional newscasts or reports on world events. Examples: The channel discusses news, current events, or other news-related topics. *See Few-shot Example 5*.
6. **Medical/Health Content:** The channel discusses medical or health-related topics, events (e.g., COVID-19, vaccines, etc.), or figures (e.g., RFK Jr., Dr. Fauci, etc.).
7. **Cryptocurrency Content:** The channel mentions cryptocurrency topics, events, or figures (e.g. Sam Bankman-Fried), companies (e.g., MicroStrategy, Coinbase, Binance), or other cryptocurrency-related topics. Examples: The channel discusses cryptocurrency, USDT, Bitcoin, Ethereum, or other cryptocurrency topics.
8. **Gambling Content:** The channel discusses gambling or betting topics, events, or figures. *Few-shot Example 6*.
9. **Money-Making Content:** The channel discusses stocks, market topics, investment advice, options trading, or other stock market-related topics. The channel also covers e-commerce, marketing, and general strategies for making money online or offline. *See Few-shot Example 7*.
10. **Kids Content:** The channel is oriented towards young audiences, such as nursery rhymes or kids' shows like CocoMelon.
11. **Potential Copyright Infringement:** The channel discusses or includes content from shows, movies, software, or music that they do not own. *Few-shot Example 8*.
12. **Manosphere/Redpill Content:** The channel discusses manosphere or redpill topics, events, or figures, alpha males, pickup artists, Jordan Peterson and Andrew Tate. *Few-shot Example 9*.
13. **Extremist Content:** The channel contains toxic content or content from extremist groups, such as white supremacist, jihadist, neo-nazi, alt-right, etc.

Instructions: For each text document, you will return a JSON object with the topics as top-level fields and the ids associated to each document used in the classification, as well as a string explaining the reasoning for the classification.

You must only output a valid JSON object matching this schema:

```
{"non_youtube_monetization": {"ids": ["list of ids"], "reasoning": "string"},
  "ai_generated_videos": {"ids": ["list of ids"], "reasoning": "string"},
  "political_content": {"ids": ["list of ids"], "reasoning": "string"},
  "religious_content": {"ids": ["list of ids"], "reasoning": "string"},
  "news_content": {"ids": ["list of ids"], "reasoning": "string"},
  "medical_health_content": {"ids": ["list of ids"], "reasoning": "string"},
  "cryptocurrency_content": {"ids": ["list of ids"], "reasoning": "string"},
  "gambling_content": {"ids": ["list of ids"], "reasoning": "string"},
  "financial_content": {"ids": ["list of ids"], "reasoning": "string"},
  "kids_content": {"ids": ["list of ids"], "reasoning": "string"},
  "potential_copyright_infringement": {"ids": ["list of ids"], "reasoning": "
      string"},
  "manosphere_redpill_content": {"ids": ["list of ids"], "reasoning": "string"},
  "hateful_extremist_content": {"ids": ["list of ids"], "reasoning": "string"}}
```

Figure A.2: Topic annotation prompt.

Few-Shot Examples: Topic Annotation

Example 1 – Non-YouTube Monetization:

```
{"id": 1, "text": "Hey, There Welcome To @UsdtMoneyMake899 Here You'll Get A New
    Earning Video And If You Earn To Watch USDT/TRX Video Then Don't Wait Just
    Subscribe Now We Are Here For This! Verified YouTube Channel. FOR BUSINESS.
    TELEGRAM id - @UsdtMoneyMake899 Telegram id Link: https://t.me/<censored>"}
{"id": 2, "text": "NUMBERS ON TELEGRAM"}
```

Note: Original prompt included example above (id 2) in Russian. For ease of presentation, the excerpt above was translated).

Example 2 – AI-Generated Videos:

```
{"id": 2, "text": "#AI #AIChannel #AIGenerated #DALLE #Midjourney #StableDiffusion
    "}
```

Example 3 – Political Content:

```
{"id": 3, "text": "Donald Trump will make America great again. Do you agree?"}
{"id": 4, "text": "#army #armyofficre #indianarmy #armylover #commando #bsf"}
```

Example 4 – Religious Content:

```
{"id": 5, "text": "Who would win in a battle, Jesus  or The Devil"}
```

Example 5 – News Content:

```
{"id": 6, "text": "Police Ordered Her to Stop...But She Kept Driving!"}
```

Example 6 – Gambling Content:

```
{"id": 7, "text": "Slot Games From You WISDOM OF ATHENA 1000X | GOLDEN FISH
    RAINING FROM THE SKY"}.
{"id": 8, "text": "This channel is made for entertainment purposes. It is operated
     for the purpose of providing gaming information, news, strategies and rules.
    Disclaimer* Gamble responsibly. My channel is here only to upload my gambling
    experiences and entertain you. Gambling is a very good way to lose money, so I
     recommend that you do not gamble at all. Please do your own research on the
    validity of these suggestions before playing or buying. Please note that
    gambling always results in losses in the long run. You can't beat the casino
    and casinos should be considered as a form of entertainment only."}
```

Note: Original prompt included these examples (id 7 and 8) in Turkish and Korean. respectively. For ease of presentation, the excerpts above were translated.

Example 7 – Stock Market Content:

```
{"id": 9, "text": "Arbitrage School"}
{"id": 10, "text": "Global Leader in Grid Strategy | Connecting the World to
    Unleash Infinite Flexible Capital. We specialize in innovative grid strategies
     that link global market"}
{"id": 11, "text": "How to Make Money with AI Tools in 2024 Easy & Fast Methods"}
```

Example 8 – Copyright Infringement:

```
{"id": 12, "text": "download coreldraw graphics suite 2024 crack free"}
```

Example 9 – Manosphere/Redpill Content:

```
{"id": 13, "text": "We Don't Need MEN Anymore | Jordan Peterson Edit"}
```

Figure A.3: Few-shot examples for topic annotation prompt.

Custom URL Changes

- Changed on 2024-12-19 01:53:11
  - From: `@uzbek24`
  - To: `@mikecryptonews`

Channel Timeline

Time	Title	Custom URL	Subscribers	Views	Videos	Country	Topics	For Kids?
2024-12-19 01:53:11	Crypto Mike	@mikecryptonews	393,000	88,657,354	486	United States	Society, Politics	No

Channel Description

Time	Title	Custom URL	Subscribers	Views	Videos	Country	Topics	For Kids?
2024-12-14 05:40:05	UZBEK 24	@uzbek24	393,000	88,529,021	483	Germany	Politics, Society	No

Channel Description
tg <Censored ID>

Time	Title	Custom URL	Subscribers	Views	Videos	Country	Topics	For Kids?
2024-12-07 02:12:30	UZBEK 24	@uzbek24	392,000	88,193,479	483	Germany	Politics, Society	No

Channel Description
Ассалом алейкум каналимизга хуш келибсиз, каналимизда барча хабар ва янгиликлар тасдикланган. Янгиликларимзга кушимча талаб ва таклифингизни албатта коментарияда колдиринг барчасини инобатка оламиз.

• • •

[Cropped for brevity]

Videos

Thumbnail	Title	Published	Views	Likes	Comments
	The best Solana arbitrage strategy 2025 I SOL *Crypto Arbitrage* Scheme I Big Cryptocurrency News	2025-01-02 23:54:48	824	3	3
	БУЮК ГАРБ СИНДИ АМЕРИКА ДАВЛАТИ РОССИЯГА РУБЛЬДА ТУЛОВНИ БОШЛАДИ	2024-09-18 14:42:13	104,799	3,433	165
	ГАРБ РОССИЯ ХУДУДЛАРИГА ХУЖУМЧИ ДРОНЛАР ОРКАЛИ ОММАВИЙ ЗАРБАЛАР БЕРДИ	2024-09-10 11:28:10	21,929	941	27
	ХИТОЙ УЗИДАН БУЮК ЯСАБ ОЛГАН АМЕРИКАНИ ЖИДДИЙ ОКИБАТЛАР БИЛАН ОГОХЛАНТИРДИ	2023-03-23 15:22:43	206,956	3,143	155

• • •

[Cropped for brevity]

Figure A.4: Markdown document generated for each channel and provided to coders during qualitative analyses. The document provides the handle changes, a channel timeline which includes snapshots of profile information over time, and all videos observed for the channel. The channel above has since been deleted.

Appendix B

Chapter 4 Data Schemas and Extended Analyses

B.1 Object features in the Hansa database

The objects listings, reviews, users, orders and transactions have many features in the database. Tables B.1, B.2, B.3, B.4 and B.5 show the features used for analysis. The features above the dashed line were available in the original table in the Hansa back-end database. The features below the dashed line were either computed from other features, or merged from different tables.

B.1.1 Bias analysis

B.2 Exploratory Factor Analysis

We begin by constructing a $n \times k$ data matrix, with n corresponding to the number of listings ($n = 123,133$) and k to the number of features ($k = 9$) for each listing. Since our variables are a mix of numeric and binary types, we calculate *polychronic* and *Pearson* correlations between our variables from the $n \times k$ data matrix and use the resulting $k \times k$ heterogeneous correlation matrix as input for our exploratory factor analysis. We tested the suitability of our data for factor analysis by performing the KMO and Bartlett's tests. The results in Table B.8 show already that there is a very low degree of information overlap among the variables.
Factor analysis generates a set of i latent factors, each labeled as MR_i, from our correlation matrix. We first use scree-plot analysis [313] and Horn's parallel analysis [314] to determine a suitable i, the number of latent factors to look for ($i = 4$ in our case). Given the $k \times k$ correlation matrix, we then look for three underlying latent factors using a so-called "*minres*" factor analysis method. Moreover, we also apply a so-called

Table B.1: Listing features

Feature	Description	Type
ID	Unique identifier of the listing	Integer
User ID	ID of the vendor	Integer
Time created	Timestamp of creation	Timestamp
Time updated	Timestamp of last edit	Timestamp
Title	Title of the listing	String
Description	Description of the listing	String
Class	Digital (d) of Physical (p) listing	String
Price	Price of the listing in the chosen currency	Double
Currency	Currency chosen by the vendor	String
BTC price	BTC price of the listing on 20-7-2017	Float
Ships from	Country listing is shipped from	String
Ships to	List of countries listing can be shipped to	String
Is hidden	Whether the listing is hidden (1) or not (0)	Binary
Is deleted	Whether the listing is deleted (1) or not (0)	Binary
Views	Number of views on the listing since 2015/8/12	Integer
Category	Vendor-provided category of the listing	String
# Reviews	Total number of reviews on the listing	Integer
# Orders	Total number of orders of the listing	Integer
Age listing	Amount of days between creation and 2017/7/20	Integer
USD price	BTC price converted to USD price on 2017/7/20	Float
soldNoReview	Proxy for custom listings (sold without review)	Binary
Scraped	Whether the listing has been scraped (1) or not (0)	Binary

Table B.2: Review features

Feature	Description	Type
ID	Unique identifier of the review	Integer
User ID	ID of the buyer	Integer
Vendor ID	ID of the vendor	Integer
Order ID	ID of the order the review belongs to	Integer
Listing ID	ID of the listing the review belongs to	Integer
Review	Review message written by the buyer	String
btcValue	BTC price of the listing at the time of the order	Float
Time review	Timestamp of when the review was written or updated	Timestamp
Is edited	Boolean of whether the review is edited (1) or not (0)	Binary
Is purged	Boolean of whether the review is purged (1) or not (0)	Binary
Listing title	Title of the listing when the review was written	String
Deleted listing	Whether the listing the review belongs to was deleted	Binary
Hidden listing	Whether the listing the review belongs to was hidden	Binary
Scraped	Whether the review has been scraped (1) or not (0)	Binary
Scraped listing	Whether the listing the review belongs to was scraped	Binary
Scraped vendor	Whether the vendor of the listing the review belongs to was scraped	Binary

"*oblimin*" rotation to the resulting set of factors since we expect the resulting factors to be correlated.

The resulting four factors, their so-called "*loadings*," in addition to several other quan-

Table B.3: User features

Feature	Description	Type
ID	Unique identifier of the user	Integer
Username	Username chosen by the user	String
Is vendor	Whether the user is a vendor (1) or not (0)	Binary
Time registered	Timestamp of when the user registered	Timestamp
# Listings	The number of listings the user created	Integer
# Reviews	The number of reviews the user received on its listings	Integer

Table B.4: Order features

Feature	Description	Type
ID	Unique identifier of the order	Integer
User ID	ID of the buyer	Integer
Listing ID	ID of the listing that is bought	Integer
Quantity	Amount of products that are bought	Integer
Payment address	BTC payment address for the buyer	String
BTC items	BTC amount of the items (btc price * quantity)	Float
BTC shipping	BTC amount for shipping costs	Float
BTC received	Amount of BTC received on payment address	Float
Fee	Fee paid to the Hansa market	Float
Message	Message of the buyer to the vendor	String
Message vendor	Message of the vendor to the buyer	String
Time purchase	Timestamp of when the buyer pressed 'buy'	Timestamp
Time accepted	Timestamp of when the vendor accepted the order	Timestamp
Time payment	Timestamp of when the order was paid by the buyer	Timestamp
Time shipped	Timestamp of when the order was shipped by the vendor	Timestamp
Time dispute	Timestamp of when a dispute was started between buyer and vendor	Timestamp
Is refunded	Boolean of whether the item was refunded (1) or not (1)	Binary
USD price	BTC received transformed to USD[1]	Float
Vendor ID	ID of the vendor the listing that is bought	Integer

Table B.5: Transaction features

Feature	Description	Type
ID	Unique identifier	Integer
Transaction ID	BTC transaction ID	String
Address	Order payment address	String
BTC	Amount of bitcoin paid	Float
Time tx	Timestamp of the transaction	Timestamp
Fee tx	Fee of the transaction	Float
Cluster	Name of the internal cluster used	String

tities of interest in factor analysis are illustrated in Table B.9. Factor loadings in Table B.9 (the values reported under each MR_i column), express how much a factor can explain a corresponding variable as a number ranging from -1 to 1. Crudely put, a

	Tests			Scraped reviews $n = 139,271$				Not scraped reviews $n = 118,913$			
Variable	Test	Statistic	p-value	M	μ	σ	min-max	M	μ	σ	min-max
listingScraped	χ^2 test	137,129.00	0.00	1.00	1.00	0.00	1–1	0.00	0.32	0.48	0–1
vendorScraped	χ^2 test	24,324.84	0.00	1.00	1.00	0.04	0–1	1.00	0.83	0.37	0–1
isEdited	χ^2 test	96.13	0.00	0.00	0.02	0.14	0–1	0.00	0.02	0.12	0–1
isPurged	χ^2 test	2,642.36	0.00	0.00	0.01	0.09	0–1	0.00	0.04	0.19	0–1

Table B.6: Results of the Mann-Whitney U and χ^2 tests between scraped and not scraped reviews

	Tests			Scraped vendors $n = 1,929$				Not scraped vendors $n = 1,696$			
Variable	Test	Statistic	p-value	M	μ	σ	min-max	M	μ	σ	min-max
hasListing	χ^2 test	1,277.90	0.00	1.00	1.00	0.02	0–1	0.00	0.49	0.50	0–1
listingScraped	χ^2 test	3,517.78	0.00	1.00	0.99	0.11	0–1	0.00	0.00	0.05	0–1

Table B.7: Results of the χ^2 test between scraped and not scraped vendors

Table B.8: Results of the KMO and Bartlett's tests

Test	**Test statistic**	p**-value**
KMO	0.546	
Bartlett	210072.289	0.0

Table B.9: Factor Analysis Output

loading expresses association strength between the latent factor and the original variable. A loading value close to 1 or -1 indicates that a factor "loads" highly onto a variable – i.e., is strongly associated with and explains the observed variance of that variable, while a value close to 0 expresses weak association. For each factor we apply a cut-off point value of 0.4 to its set of loadings, a common threshold used in the literature, to determine the most prominent associations [315]. These are reported in **bold** font, and indicate variables strongly associated with latent factors.

In general, the four latent factors (or three, if we exclude MR_4 based on no variable surpassing the loading threshold of 0.4) only capture 0.35% of the variance. Here, we also observe that of our nine variables only two seem to be associated with the same underlying latent factor, namely `numReviews` and `numOrders`. However, we reason that this is an artifact of the market policy that forcibly associates reviews with actual orders. Thus, for our analysis of testing whether any significant differences exist between scraped and not-scraped listings, we include all variables individually.

B.3 Abundance Estimation Algorithms

We summarize here the three abundance estimation algorithms we employ.

Lincoln-Petersen (LP) The Lincoln-Petersen method estimates N, the population, as

$$\hat{N} = \frac{Kn}{k}\,, \tag{B.1}$$

where n is the number of units marked on the first sampling, K is the number of units marked in the second sampling, and k the number of recaptured units that were marked [316].

Schnabel The Schnabel method extends the LP method for situations where we have various samples:

$$\hat{N} = \frac{\sum_t (C_t M_t)}{\sum_t R_t + 1}\,, \tag{B.2}$$

where C_t are the total number of units caught at time t, R_t are the number of units already marked at time t, and M_t is the number of marked units at time $t-1$ [316]. Both the Schnabel and LP methods, however, assume that the populations are *closed*, that is, no units appear (births) nor disappear (deaths). To relax these assumptions, "open-population" models which model recruitment and survival were introduced. In this paper, we use the Jolly-Seber (JS) estimator [317].

We used the POPAN formulation [317]. We estimate $\hat{p}_t$ the probability of capture, $\hat{\phi}_t$ the probability of survival between periods, and $\hat{b}_t$ the probability of entering the population. These parameters are estimated using a Maximum Likelihood Estimation (MLE) procedure on a multinomial distribution, where each *encounter history* is a possible outcome. An encounter history is a series of observations of the studied object, encoded as a string of 0s for sampling dates when the object was not observed and 1s when it was observed. The total population N is estimated at each time t by:

$$\hat{N}_t = \hat{N}_{t-1}\hat{\phi}_{t-1} + B_{t-1}\,, \tag{B.3}$$

where B_t is the number of new entrants to the population.

B.4 Scrape Dates

We obtained scrapes from the Hansa marketplace taken on: October 8$^{\text{th}}$ 2015, October 11$^{\text{th}}$ 2015, October 16$^{\text{th}}$ 2015, October 23$^{\text{th}}$ 2015, October 25$^{\text{th}}$ 2015, November 2$^{\text{nd}}$ 2015, December 1$^{\text{st}}$ 2015, December 13$^{\text{th}}$ 2015, January 7$^{\text{th}}$ 2016, January 17$^{\text{th}}$ 2016, April 30$^{\text{th}}$ 2016, June 8$^{\text{th}}$ 2016, July 7$^{\text{th}}$ 2017 and July 14$^{\text{th}}$ 2017.

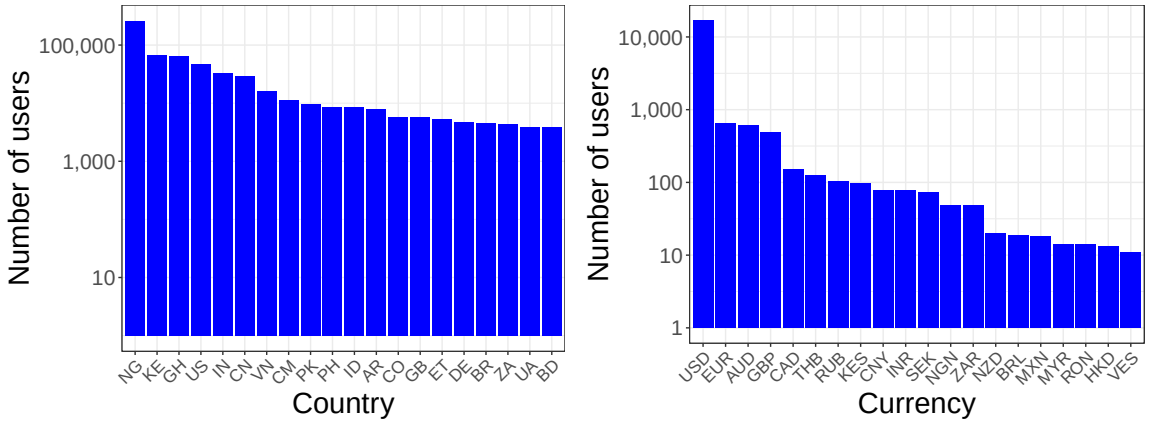
Appendix C

Chapter 6 Complementary Analyses

C.1 Geographical Considerations

User origin is another feature we consider for our experiments. For each platform, we aggregate the number of users by country of origin. In Paxful, the country seems to be determined by the IP address used when registering an account. In LCS, users self-disclose their local currency, so we employ this as a proxy for their location. The default is set as USD. Figure C.1 shows the number of users for each country: Paxful on the left and LCS on the right. The customer base seems to be significantly different between both platforms. Paxful features many users from Africa, such as Nigeria (NG), Kenya (KE), and Ghana (GH), while LCS attracts more users from Australia (AUD) and Europe (EUR).

At the time of data collection, the Paxful API returns both the country of registration and the country from which the user last accessed the platform, based on the user's IP address. Our long-term observations reveal that some vendors appear to log in from countries different from their country of registration. Figure C.2 is a heatmap that evidences these changes. The y-axis is the country of registration and the x-axis is the country of access (any point in our observation). For better readability, we only include pairs of countries with more than 30 distinct users, and normalize by the x-axis. We observe that many users route through the US, Kenya, and Nigeria. Given that these users registered in a different country, we hypothesize some of their traffic is over VPNs (or Tor) to obfuscate its true origin. Figure C.3 shows the ratio of users, per country, who access the site from a different country at least once. We only include countries that have more than 500 vendors. For example, more than 99% of vendors who registered in China later used IP addresses from a different country. Users in cryptocurrency-regulated countries such as China (CN), Bangladesh (BD), Indonesia (ID), Pakistan (PK), Vietnam (VN), and Cameroon (CM) [281] appear to connect to the site from alternate locations often. These users are incentivized to ob-



(a) Paxful (b) LCS

Figure C.1: Number of users for the top 20 countries in Paxful and currencies in LCS. y-axis is in log scale.

fuscate their location, but do not seem to maintain good operational security in the long run. Another possible motivation is to circumvent restrictions that Paxful has for certain payment methods in some jurisdictions. For example, as reported in the Paxful subreddit [318], Zelle is prohibited in Cameroon, China, Ghana, India, and Nigeria at the time of writing.

C.2 Identifying Suspension

We rely on values returned by the API(s) to distinguish between regular users, suspended users, and users who have changed their usernames. In Paxful, upon suspension, a user is marked as "not active" on the web page and the API call for their profile returns a JSON field "`is_active`" as False. According to a Paxful moderator on Reddit [319], this indicates either an account ban (non-reversible) or an account lock (reversible). In terms of account deletion, Paxful API does not appear to change.

Unlike Paxful, LCS does not explicitly mark accounts as suspended, a user page is taken down when the user changes to a different username, or when they get suspended by the platform. The page says "not found" when the username has changed, but redirects to the ads page if the user was suspended. Likewise, the API responds differently. We attempted to collect account suspension data from other platforms such as Binance P2P, one of the largest players in this space, but could not identify the signs of account suspension on those. Indeed, a Binance P2P user page does not seem to change even if the user deletes the account. We use account suspension as the main label (and prediction target) of our machine learning model in §6.5 and §6.6.

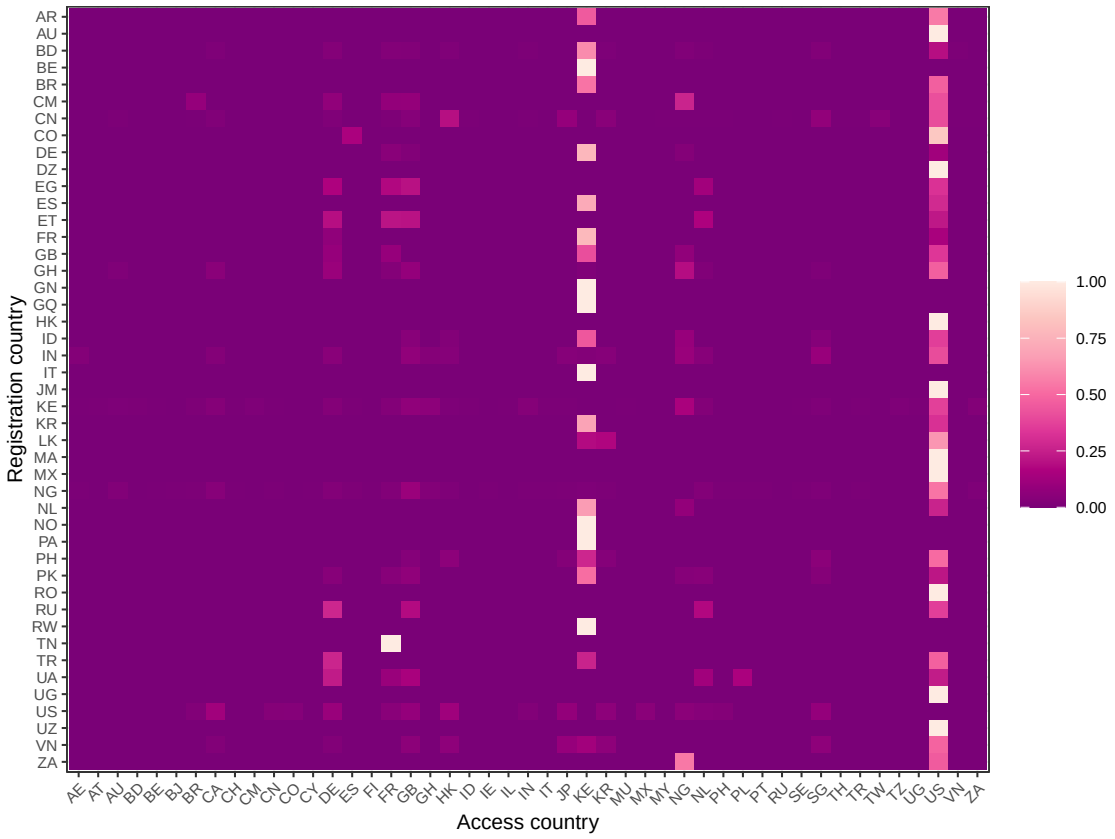


Figure C.2: Heatmap of country changes between registration and subsequent accesses (normalized by *x*-axis).

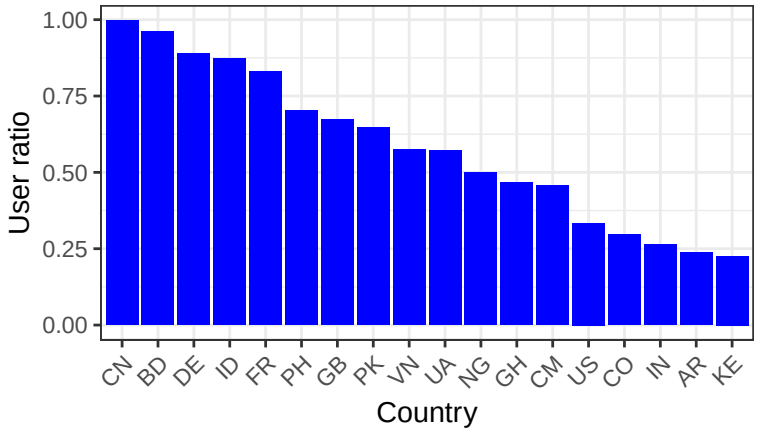


Figure C.3: Ratio of users for which the reported country of access is different at least once from the registration country (for countries with more than 500 users).

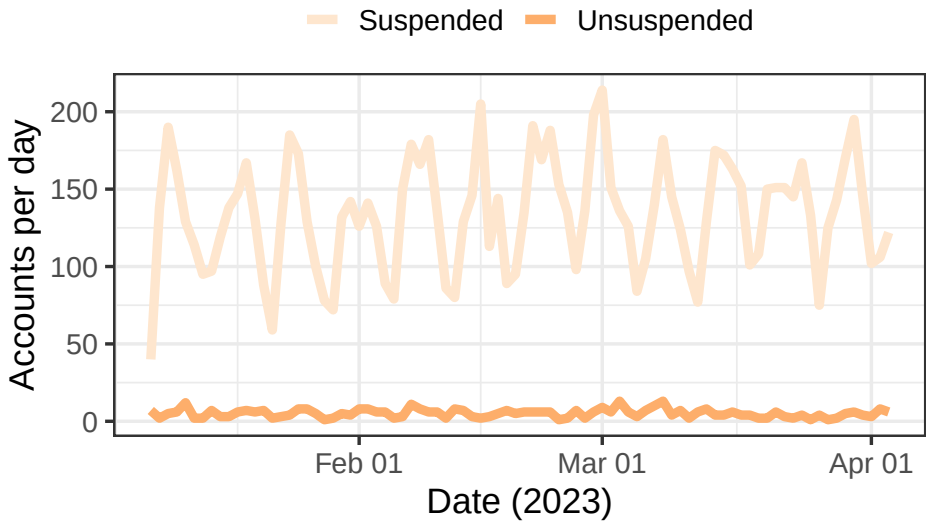


Figure C.4: Account (un)suspensions per day on Paxful, 01/08/23–03/31/23.

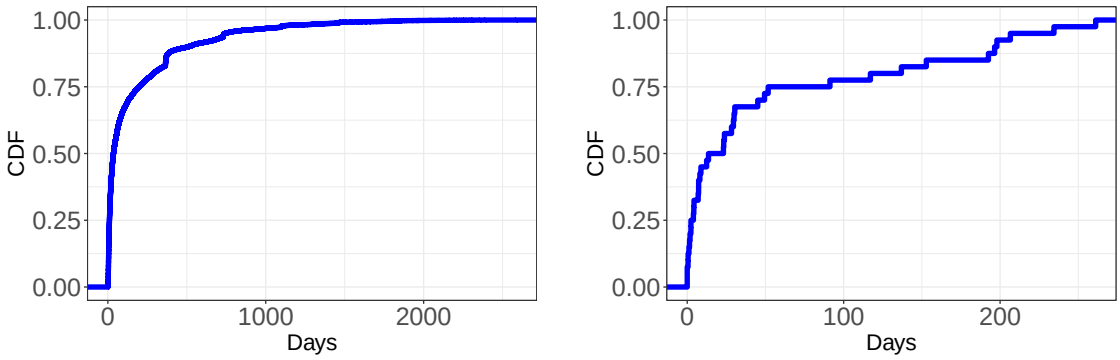
C.3 Suspension label validation (Paxful)

We performed several validation tests to ensure the quality of the suspension labels obtained from Paxful. Figure C.4 shows the number of suspensions and unsuspensions for each day between January 8th, 2023 and March 31st, 2023. We only include users for whom we can confidently determine the time of the suspension. More precisely, we pinpoint the time of suspension if the user ban status changed from false to true based on two consecutive observations within one day (86 400 seconds). We can confirm some weekly seasonality—there is a decrease in the number of suspensions on weekends—suggesting that platform moderation is not purely automated, and instead relies on human input to some extent. Second, the label seems to imply permanent suspension for most users. Longitudinal observations confirm that only a small portion of those are unsuspended (lower curve).

C.4 Platform Moderation Evaluation

To investigate the level of platform moderation, we evaluate how long the platform takes to find malicious accounts and how long it takes to lift suspensions on accounts that turned out to be benign.

We first calculate the number of active days for suspended accounts. Paxful API returns a rough estimate of registration time (e.g., “3 hours before” or “1 month before”). We monitor changes in that response across queries (e.g., “4 days before” to “5 days before”), and estimate the registration timestamp based on multiple data



(a) Days to suspension

(b) Days to unsuspension

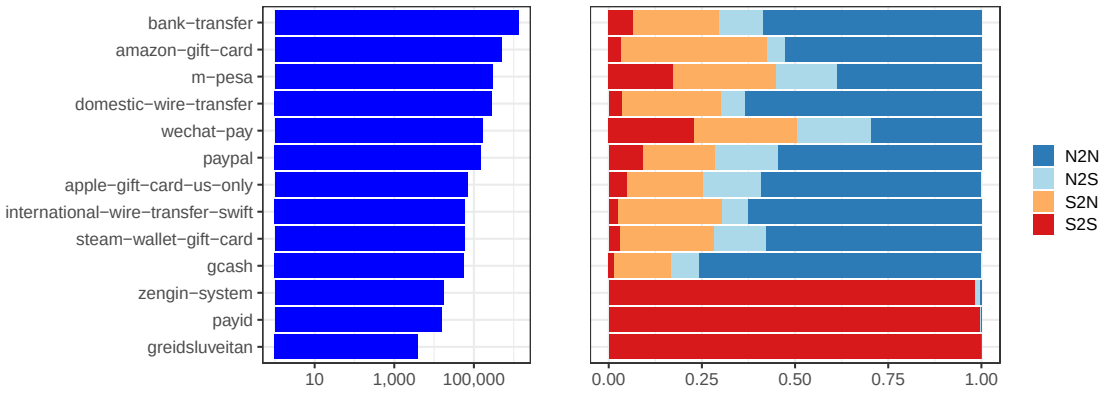
Figure C.5: CDFs of the number of days to suspension and to release.

points. Figure C.5a shows the Cumulative Distribution Function (CDF) for the ages of suspended accounts. 18% of suspended accounts are suspended within one week of registration, 48% within a month, and 83% within a year. The small spikes around 365 and 730 days are an artifact of the coarseness of our estimated registration time, which becomes less accurate for old accounts (e.g., the API returns "1 year ago" to "2 years ago").

We next derive the number of days the platform takes to lift suspensions on accounts that turned out to be benign. Minimizing the length of an erroneous suspension is critical to building trust with customers. To measure this, we select users who have been suspended once but were unsuspended later. We calculate the time span between the observation when they first get suspended and one observation before the status changed to "active" (i.e., a lower bound). Figure C.5b shows the CDF of the length of time before the status of a suspended account is restored. We only include accounts in which we can confidently identify the timing of the unsuspension ($n = 41$). Within a week of (erroneous) suspension, 32% of these accounts see their bans lifted; 68% are unsuspended within a month. We do not include accounts that have not been released at the end of our observation period.

C.5 Feedback Keyword Searches

To find scam-related feedback at scale and spot risky users from feedback comments, we use a list of keywords: "scam," "rip," "liar," "conman," "thief," "thieves," "crime," "criminal," "fraud," "steal," "stole," "cheat," "fake," "ghosted," "swindle," "chargeback," "reverse," "coin locker" and perform a keyword search (perfect match) to discover scam-related feedback for all the negative reviews we collected. We use the same procedure for slow vendors: "slow," "sluggish," "not fast," "not respon-



(a) Number of feedback (b) Frequency of interactions

Figure C.6: Top 10 payment methods + 3 selected payments. N: non-suspended accounts, S: suspended accounts (e.g., N2S: feedback from non-suspended to suspended accounts)

sive," and "delay" as well. We try to avoid false positives, i.e., flagging non-scam-related reviews as scams. To test the efficiency of our keyword-based approach, we run the keyword search on 500 reviews annotated by our first coder from §6.4.1 as validation. 41% of these reviews are captured as scam-related feedback with zero false negatives. The coder annotated 55.4% of these as "scam," meaning our automation failed to detect 14.4% of scam-related feedback. We thus regard the result of the keyword search as a lower bound.

Among all negative reviews, our automated classification flags 40% as scam-related feedback, and 9.45% of transactions were speed-related. By aggregating at a vendor level, 2 493 users have at least one scam-related feedback, and only 642 users (around 2.6% of total suspended accounts) received multiple scam-related feedback. Considering that a total of 24 562 (46%) vendors are suspended, solely looking at feedback data fails to spot many risky accounts. Nevertheless, we incorporate the number of scam/speed-related keywords as one of the features in our ML model – but realize it is not sufficient on its own.

C.6 Complementary Evidence of Self-promotion

This section complements the discussion about users that appear to engage in the self-promoting attacks described in §6.4.2.

First, those users pick rare payment methods, that do not appear to be used in their country of registration. Figure C.6a shows the number of reviews for 13 payment

methods: the top 10 payment methods (bank transfers, Amazon Gift card, M-Pesa, etc.), and three payment systems we choose to investigate: Zengin (Japan), PayID (Australia), and Greidsluveitan (Iceland). Figure C.6b further shows the split among these reviews for four interaction types: N2N (non-suspended accounts giving feedback to non-suspended accounts), N2S (non-suspended to suspended accounts), S2N (suspended to non-suspended) and S2S (suspended to suspended). The three payment systems at the bottom are dominated by suspended-to-suspended transactions. Looking at these three payment systems, 8 563 unique users give feedback, and only 391 users receive feedback. Interestingly, all the users giving feedback are from Vietnam – and not Japan, Australia, or Iceland where those three payment methods are reportedly used.

We further confirm that a subset of more than 100 of these users giving feedback send feedback together repeatedly. Those users appear to have been solely created for the purpose of self-promoting attacks, that is, they appear to be Sybils tasked with boosting the reputation of the feedback receivers. For example, one user received feedback from those 103 users through the Zengin payment system. All feedback was sent within 1 400 seconds and all reviews were positive. Several variations of the same comments appear to have been re-used (e.g., *"Excellent trader very fast.," "Good and quick," "Welcome to trade with me again," " He is a reliable trader."*).

In addition, those accounts exhibit unnatural trade distributions. The trade count of the users giving feedback is oddly distributed. For example, among all users that rely on the Zengin payment system, five accounts have engaged in three trades or less, 300 users have exactly four trades, but only two users engaged in five trades. This strongly suggests the presence of Sybils and automation. Similar findings apply to the other two payment methods. Most users receiving feedback have between 200–250 trades, which is markedly different from the overall distribution of trade counts. Based on all of the above, we believe these accounts are most likely engaged in coordinated self-promoting attacks.

Bibliography

[1] S. Tadelis. "Reputation and Feedback Systems in Online Platform Markets". In: *Annual Review of Economics* 8.1 (2016), pp. 321–340. DOI: 10.1146/annurev-economics-080315-015325.

[2] Paul Resnick et al. "The value of reputation on eBay: A controlled experiment". In: *Experimental Economics* 9.2 (June 2006), pp. 79–101. DOI: 10.1007/s10683-006-4309-2.

[3] Jeffrey A. Livingston. "How Valuable Is a Good Reputation? A Sample Selection Model of Internet Auctions". In: *The Review of Economics and Statistics* 87.3 (2005), pp. 453–465. ISSN: 0034-6535. URL: https://www.jstor.org/stable/40042941 (visited on 10/10/2024).

[4] Luís Cabral and Ali Hortaçsu. "The Dynamics of Seller Reputation: Evidence from Ebay". In: *The Journal of Industrial Economics* 58.1 (2010), pp. 54–78. DOI: 10.1111/j.1467-6451.2010.00405.x.

[5] RE Kraut. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press, 2012.

[6] Cliff A.C. Lampe, Nicole Ellison, and Charles Steinfield. "A familiar face(book): profile elements as signals in an online social network". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. Association for Computing Machinery, 2007, pp. 435–444. DOI: 10.1145/1240624.1240695.

[7] Judith Donath. "Signals in Social Supernets". In: *Journal of Computer-Mediated Communication* 13.1 (2007), pp. 231–251. DOI: 10.1111/j.1083-6101.2007.00394.x.

[8] Nicole B. Ellison, Jeffrey T. Hancock, and Catalina L. Toma. "Profile as promise: A framework for conceptualizing veracity in online dating self-presentations". In: *New Media & Society* 14.1 (2012), pp. 45–62. DOI: 10.1177/1461444811410395.

[9] Alexander Bilz, Lynsay A Shepherd, and Graham I Johnson. "Tainted Love: a Systematic Literature Review of Online Romance Scam Research". In: *Interacting with Computers* 35.6 (Oct. 2023), pp. 773–788. ISSN: 1873-7951. DOI:

10.1093/iwc/iwad048. eprint: https://academic.oup.com/iwc/article-pdf/35/6/773/57796351/iwad048.pdf. URL: https://doi.org/10.1093/iwc/iwad048.

[10] Riot Games. *Instant Feedback System FAQ*. Accessed: 2024-11-05. 2024. URL: https://support-leagueoflegends.riotgames.com/hc/en-us/articles/207489286-Instant-Feedback-System-FAQ.

[11] Tina Kuo, Alicia Hernani, and Jens Grossklags. "The Unsung Heroes of Facebook Groups Moderation: A Case Study of Moderation Practices and Tools". In: *Proceedings of the ACM on Human-Computer Interaction*. CSCW '23 7.Cscw1 (Apr. 2023), 97:1–97:38. DOI: 10.1145/3579530.

[12] Oliver L. Haimson et al. "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas". In: *Proceedings of the ACM on Human-Computer Interaction*. CSCW '21 5.Cscw2 (Oct. 2021).

[13] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. "A survey of attack and defense techniques for reputation systems". In: *ACM Computing Surveys*. CSUR '09 42.1 (2009), pp. 1–31.

[14] Taro Tsuchiya et al. "Misbehavior and Account Suspension in an Online Financial Communication Platform". In: *Proceedings of the ACM Web Conference*. WWW '23. Association for Computing Machinery, 2023, pp. 2686–2697. DOI: 10.1145/3543507.3583385.

[15] Madelyne Xiao et al. "Account Verification on Social Media: User Perceptions and Paid Enrollment". In: *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 3099–3116. ISBN: 978-1-939133-37-3. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/xiao-madelyne.

[16] WIRED Staff. "Elon Musk's Twitter Is a Scammer's Paradise". In: *WIRED* (Nov. 2022). Accessed: 2024-07-18. URL: https://www.wired.com/story/twitter-blue-check-verification-buy-scams/.

[17] Christopher Avery, Paul Resnick, and Richard Zeckhauser. "The Market for Evaluations". In: *American Economic Review* 89.3 (June 1999), pp. 564–584. DOI: 10.1257/aer.89.3.564.

[18] Alan Benson, Aaron Sojourner, and Akhmed Umyarov. "Can Reputation Discipline the Gig Economy? Experimental Evidence from an Online Labor Market". In: *Management Science* 66.5 (May 2020), pp. 1802–1825. DOI: 10.1287/mnsc.2019.3303.

[19] Andrey Fradkin et al. "Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb". In: *Proceedings of the Sixteenth ACM Conference*

on Economics and Computation. EC '15. Association for Computing Machinery, 2015, p. 641. DOI: 10.1145/2764468.2764528.

[20] Jiwei Li et al. "Towards a General Rule for Identifying Deceptive Opinion Spam". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Ed. by Kristina Toutanova and Hua Wu. Association for Computational Linguistics, June 2014, pp. 1566–1576. DOI: 10.3115/v1/P14-1147.

[21] Chris Nosko and Steven Tadelis. *The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment*. Working Paper. Jan. 2015. DOI: 10.3386/w20830. URL: https://www.nber.org/papers/w20830 (visited on 02/23/2024).

[22] Nan Hu, Paul A. Pavlou, and Jennifer Zhang. "Can online reviews reveal a product's true quality? empirical findings and analytical modeling of Online word-of-mouth communication". In: *Proceedings of the 7th ACM conference on Electronic commerce*. EC '06. Association for Computing Machinery, 2006, pp. 324–330.

[23] Arpit Merchant et al. "Signals Matter: Understanding Popularity and Impact of Users on Stack Overflow". In: *The World Wide Web Conference*. WWW '19. Association for Computing Machinery, May 2019, pp. 3086–3092. DOI: 10.1145/3308558.3313583.

[24] Ross Anderson et al. *Measuring the changing cost of cybercrime*. 2019. URL: https://orca.cardiff.ac.uk/id/eprint/122684/ (visited on 02/20/2024).

[25] Huilian Sophie Qiu et al. "Going Farther Together: The Impact of Social Capital on Sustained Participation in Open Source". In: *2019 IEEE/ACM 41st International Conference on Software Engineering*. ICSE '19. IEEE, May 2019, pp. 688–699. DOI: 10.1109/icse.2019.00078.

[26] Asher Trockman et al. "Adding sparkle to social coding: an empirical study of repository badges in the *npm* ecosystem". In: *Proceedings of the 40th International Conference on Software Engineering*. ICSE '18. Association for Computing Machinery, May 2018, pp. 511–522. DOI: 10.1145/3180155.3180209.

[27] Louis F. DeKoven et al. "Following Their Footsteps: Characterizing Account Automation Abuse and Defenses". In: *Proceedings of the Internet Measurement Conference 2018*. Imc '18. Association for Computing Machinery, 2018, pp. 43–55. ISBN: 9781450356190. DOI: 10.1145/3278532.3278537. URL: https://doi.org/10.1145/3278532.3278537.

[28] Janith Weerasinghe et al. "The Pod People: Understanding Manipulation of Social Media Popularity via Reciprocity Abuse". In: *Proceedings of The Web Conference*. WWW '20. Association for Computing Machinery, 2020, pp. 1874–1884. DOI: 10.1145/3366423.3380256.

[29] Yuanshun Yao et al. "Automated Crowdturfing Attacks and Defenses in Online Review Systems". In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. CCS '17. Oct. 2017, pp. 1143–1158. DOI: 10.1145/3133956.3133990. (Visited on 02/19/2024).

[30] Chaz Lever et al. "Domain-Z: 28 Registrations Later Measuring the Exploitation of Residual Trust in Domains". In: *IEEE Symposium on Security and Privacy*. S&P '16. 2016, pp. 691–706. DOI: 10.1109/sp.2016.47.

[31] Johnny So et al. "Domains Do Change Their Spots: Quantifying Potential Abuse of Residual Trust". In: *IEEE Symposium on Security and Privacy*. SP '22. 2022, pp. 2130–2144. DOI: 10.1109/sp46214.2022.9833609. URL: https://ieeexplore.ieee.org/document/9833609 (visited on 09/28/2024).

[32] U.S. Attorney's Office, Southern District of New York. *Dark Web Narcotics Dealer "Fentmaster," Responsible For Overdose Death, Sentenced To 15 Years In Prison*. Press release. July 2021. URL: https://www.justice.gov/usao-sdny/pr/dark-web-narcotics-dealer-fentmaster-responsible-overdose-death-sentenced-15-years.

[33] N. Christin. "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace". In: *The ACM Web Conference (WWW'13)*. May 2013, pp. 213–224.

[34] K. Soska and N. Christin. "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem". In: *24th USENIX Security Symposium*. USENIX '15. 2015, pp. 33–48.

[35] D. Décary-Hétu and L. Giommoni. "Do police crackdowns disrupt drug cryptomarkets? A longitudinal analysis of the effects of Operation Onymous". In: *Crime, Law and Social Change* 67.1 (Feb. 2017), pp. 55–75.

[36] David Décary-Hétu and Anna Leppänen. "Criminals and signals: An Assessment of Criminal Performance in the Carding Underworld". In: *Security Journal* 29 (2016), pp. 442–460.

[37] Alejandro Cuevas et al. "Measurement by Proxy: On the Accuracy of Online Marketplace Measurements". In: *31st USENIX Security Symposium*. USENIX '22. 2022, pp. 2153–2170.

[38] Alejandro Cuevas and Nicolas Christin. "Does Online Anonymous Market Vendor Reputation Matter". In: *Proceedings of the 33rd USENIX Security Symposium*. USENIX '24. 2024.

[39] Taro Tsuchiya, Alejandro Cuevas, and Nicolas Christin. "Identifying Risky Vendors in Cryptocurrency P2P Marketplaces". In: *Proceedings of the ACM Web Conference*. WWW '24. Association for Computing Machinery, 2024, pp. 99–110. DOI: 10.1145/3589334.3645475.

[40] Michael Spence. "Job Market Signaling". In: *The Quarterly Journal of Economics* 87.3 (1973), pp. 355–374.

[41] Amotz Zahavi. "Mate selection–A selection for a handicap". In: *Journal of Theoretical Biology* 53.1 (1975), pp. 205–214. ISSN: 0022-5193. DOI: https://doi.org/10.1016/0022-5193(75)90111-3. URL: https://www.sciencedirect.com/science/article/pii/0022519375901113.

[42] Tim Guilford and Marian Stamp Dawkins. "What are conventional signals?" In: *Animal Behaviour* 49.6 (1995), pp. 1689–1695. ISSN: 0003-3472. DOI: https://doi.org/10.1016/0003-3472(95)90090-X. URL: https://www.sciencedirect.com/science/article/pii/000334729590090X.

[43] John Maynard Smith and David Harper. *Animal signals*. Oxford University Press, 2003.

[44] Joseph Seering. "Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation". In: *Proceedings of the ACM on Human-Computer Interaction*. CSCW '20 4.Cscw2 (Oct. 2020), 107:1–107:28. DOI: 10.1145/3415178.

[45] Wikipedia contributors. *Stolen Valor Act of 2013 — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-July-2025]. 2025. URL: https://en.wikipedia.org/w/index.php?title=Stolen_Valor_Act_of_2013&oldid=1294317613.

[46] YouTube. *YouTube's Community Guidelines*. Accessed: 2025-06-04. 2024. URL: https://support.google.com/youtube/answer/9288567?hl=en.

[47] *YouTube Community Guidelines - Content Removals Transparency Report*. Google Transparency Report - YouTube policy removals. Accessed August 2025. 2025. URL: https://transparencyreport.google.com/youtube-policy/removals?hl=en.

[48] Bryan Ford and Jacob Strauss. "An offline foundation for online accountable pseudonyms". In: *Proceedings of the 1st Workshop on Social Network Systems*. SocialNets '08. Glasgow, Scotland: Association for Computing Machinery, 2008, pp. 31–36. ISBN: 9781605581248. DOI: 10.1145/1435497.1435503. URL: https://doi.org/10.1145/1435497.1435503.

[49] Jess Weatherbed. "Discord Is Verifying Some Users Age with ID and Facial Scans". In: *The Verge* (Apr. 2025). Accessed August 4, 2025. URL: https://www.theverge.com/news/650493/discord-age-verification-face-id-scan-experiment.

[50] Wikipedia contributors. *World (blockchain) — Wikipedia, The Free Encyclopedia*. [Online; accessed 18-July-2025]. 2025. URL: https://en.wikipedia.org/w/index.php?title=World_(blockchain)&oldid=1299388213.

[51] Michael Luca and Georgios Zervas. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud". In: *Management Science* 62.12 (Dec. 2016), pp. 3412–3427. DOI: 10.1287/mnsc.2015.2304.

[52] Yazan Boshmaf et al. "The socialbot network: when bots socialize for fame and money". In: *Proceedings of the 27th Annual Computer Security Applications Conference*. ACSAC '11. Association for Computing Machinery, Dec. 2011, pp. 93–102. DOI: 10.1145/2076732.2076746.

[53] Emilio Ferrara et al. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[54] Gloria Origgi, Stephen Holmes, and Noga Arikha. *Reputation: What It Is and Why It Matters*. Princeton University Press, 2018. DOI: 10.2307/j.ctvc77bzk.

[55] Chrysanthos Dellarocas. "The Digital Economy: Reputation Mechanisms". In: *Economics and Management Strategy* 12.2 (2003), pp. 157–185. DOI: 10.1111/j.1430-9134.2003.00105.x.

[56] George A. Akerlof. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism". In: *The Quarterly Journal of Economics* 84.3 (1970), pp. 488–500. DOI: 10.2307/1879431.

[57] Dina Mayzlin, Yaniv Dover, and Judith Chevalier. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation". In: *American Economic Review* 104.8 (Aug. 2014), pp. 2421–2455. DOI: 10.1257/aer.104.8.2421.

[58] Asle Fagerstrom et al. "That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on Airbnb". In: *Comput. Hum. Behav.* 72.C (July 2017), pp. 123–131. DOI: 10.1016/j.chb.2017.02.029.

[59] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market". In: *Science* 311.5762 (2006), pp. 854–856. DOI: 10.1126/science.1121066. eprint: https://www.science.org/doi/pdf/10.1126/science.1121066. URL: https://www.science.org/doi/abs/10.1126/science.1121066.

[60] Priyanga Gunarathne, Huaxia Rui, and Avi Seidmann. "Customer Service on Social Media: The Effect of Customer Popularity and Sentiment on Airline Response". In: *2015 48th Hawaii International Conference on System Sciences*. 2015, pp. 3288–3297. DOI: 10.1109/hicss.2015.397.

[61] Zsolt Katona. *Competing for Influencers in a Social Network*. Tech. rep. 13-06. Available at SSRN: https://ssrn.com/abstract=2335679. NET Institute, 2018.

[62] Library of Congress. *Influencer Marketing: A Research Guide - Metrics and Costs*. https://guides.loc.gov/influencer-marketing/metrics-and-costs. Accessed on July 7, 2025. 2025.

[63] Lin William Cong and Siguang Li. *A Model of Influencer Economy*. Working Paper 31243. National Bureau of Economic Research, May 2023. DOI: 10.3386/w31243. URL: http://www.nber.org/papers/w31243.

[64] The Economist. "Too many people want to be social-media influencers". In: *The Economist* (Oct. 2024). Accessed: 2025-06-04. URL: https://www.economist.com/business/2024/10/29/too-many-people-want-to-be-social-media-influencers.

[65] Eric Drott. "Fake Streams, Listening Bots, and Click Farms: Counterfeiting Attention in the Streaming Music Economy". In: *American Music* 38.2 (2020), pp. 153–175. ISSN: 07344392, 19452349. URL: https://www.jstor.org/stable/10.5406/americanmusic.38.2.0153 (visited on 06/04/2025).

[66] Galen Stocking et al. *America's News Influencers*. Accessed: 2025-06-04. Nov. 2024. URL: https://www.pewresearch.org/journalism/2024/11/18/americas-news-influencers/.

[67] Wikipedia contributors. *Citizen journalism — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-July-2025]. 2025. URL: https://en.wikipedia.org/w/index.php?title=Citizen_journalism&oldid=1283232915.

[68] X (formerly Twitter). *About X Verified Accounts*. Accessed: 2024-11-22. 2024. URL: https://help.x.com/en/managing-your-account/about-x-verified-accounts.

[69] Carson Powers et al. ""I can say I'm John Travolta...but I'm not John Travolta": Investigating the Impact of Changes to Social Media Verification Policies on User Perceptions of Verified Accounts". In: *Symposium on Usable Privacy and Security*. SOUPS '24. USENIX Association, Aug. 2024, pp. 353–372. ISBN: 978-1-939133-42-7.

[70] Ashton Anderson et al. "Steering user behavior with badges". In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13. Association for Computing Machinery, May 2013, pp. 95–106. DOI: 10.1145/2488388.2488398.

[71] Huilian Sophie Qiu et al. "The Signals that Potential Contributors Look for When Choosing Open-source Projects". In: *Proceedings of the ACM on Human-Computer Interaction (PAMHCI)*. CSCW '19 3.Cscw (Nov. 2019), pp. 1–29. DOI: 10.1145/3359224. (Visited on 02/21/2024).

[72] CmdrTaco (Slashdot). *Slashdot Moderation*. https://www.slashdot.org/moderation.shtml. Last updated September 9, 1999. 1999.

[73] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "How Community Feedback Shapes User Behavior". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '14 8.1 (May 2014), pp. 41–50. DOI: 10.1609/icwsm.v8i1.14518. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14518.

[74] Warut Khern-am-nuai et al. "Haters Gonna Hate? How Removing Downvote Option Impacts Discussion Culture in Online Forum". In: *Proceedings of the 2020 International Conference on Information Systems*. 3. 2020. URL: https://aisel.aisnet.org/icis2020/sharing_economy/sharing_economy/3.

[75] YouTube Official Blog. *An update to dislikes on YouTube*. https://blog.youtube/news-and-events/update-to-youtube/. Accessed on July 7, 2025. 2025.

[76] Youngsam Park, Damon McCoy, and Elaine Shi. "Understanding craigslist rental scams". In: *Proceedings of the 20th International Conference on Financial Cryptography and Data Security*. Springer. 2017, pp. 3–21.

[77] Andrew Chu et al. "Behind the Tube: Exploitative Monetization of Content on YouTube". In: *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Aug. 2022, pp. 2171–2188. ISBN: 978-1-939133-31-1. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/chu.

[78] Cameron Ballard et al. "Conspiracy Brokers: Understanding the Monetization of YouTube Conspiracy Theories". In: *Proceedings of the ACM Web Conference 2022*. Www '22. Association for Computing Machinery, 2022, pp. 2707–2718. ISBN: 9781450390965. DOI: 10.1145/3485447.3512142. URL: https://doi.org/10.1145/3485447.3512142.

[79] Yiqing Hua et al. "Characterizing Alternative Monetization Strategies on YouTube". In: *Proc. ACM Hum.-Comput. Interact.* CSCW '22 6.Cscw2 (Nov. 2022). DOI: 10.1145/3555174. URL: https://doi.org/10.1145/3555174.

[80] Sujha Sundararajan. *Data Reveals Persistent Fake Followers Problem on Crypto Twitter*. https://cryptonews.com/news/data-reveals-persistent-fake-followers-problem-crypto-twitter/. Accessed on July 7, 2025; published approx. early 2024. 2024.

[81] Hao He et al. *4.5 Million (Suspected) Fake Stars in GitHub: A Growing Spiral of Popularity Contests, Scams, and Malware*. 2024. arXiv: 2412.13459 [cs.CR]. URL: https://arxiv.org/abs/2412.13459.

[82] Samuel Woolley. *Manufacturing consensus: Understanding propaganda in the era of automation and anonymity*. Yale University Press, 2023.

[83] Ruben Recabarren et al. "Strategies and Vulnerabilities of Participants in Venezuelan Influence Operations". In: *32nd USENIX Security Symposium*. USENIX '23. 2023, pp. 6683–6700.

[84] Mohammad Hammas Saeed et al. "TrollMagnifier: Detecting State-Sponsored Troll Accounts on Reddit". In: *IEEE Symposium on Security and Privacy*. SP '22. 2022, pp. 2161–2175. DOI: 10.1109/sp46214.2022.9833706.

[85] Savvas Zannettou et al. "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web". In: *Companion Proceedings of the World Wide Web Conference*. WWW '19. Association for Computing Machinery, 2019, pp. 218–226. DOI: 10.1145/3308560.3316495.

[86] Savvas Zannettou et al. "Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls". In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '19. Association for Computing Machinery, 2019, pp. 353–362. DOI: 10.1145/3292522.3326016.

[87] Jacob Wallis et al. "Influence for hire. The Asia-Pacific online shadow economy". In: (Oct. 2021). Accessed: 2025-06-04. URL: https://www.aspi.org.au/report/influence-hire.

[88] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. "Hateful People or Hateful Bots? Detection and Characterization of Bots Spreading Religious Hatred in Arabic Social Media". In: *Proc. ACM Hum.-Comput. Interact.* 3.Cscw (Nov. 2019). DOI: 10.1145/3359163. URL: https://doi.org/10.1145/3359163.

[89] Microsoft Security Blog. *Jasper Sleet: North Korean Remote IT Workers' Evolving Tactics to Infiltrate Organizations*. https://www.microsoft.com/en-us/security/blog/2025/06/30/jasper-sleet-north-korean-remote-it-workers-evolvi. Accessed on July 7, 2025. June 2025.

[90] Kurt Thomas et al. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse". In: *Proceedings of the 22nd USENIX Security Symposium*. USENIX '13. 2013, pp. 195–210.

[91] Peng Gao et al. "SYBILFUSE: Combining Local Attributes with Global Structure to Perform Robust Sybil Detection". In: *IEEE Conference on Communications and Network Security*. IEEE, May 2018, pp. 1–9. DOI: 10.1109/cns.2018.8433147.

[92] Geli Fei et al. "Exploiting Burstiness in Reviews for Review Spammer Detection". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '13 7.1 (2013), pp. 175–184. DOI: 10.1609/icwsm.v7i1.14400.

[93] Nitin Jindal and Bing Liu. "Opinion spam and analysis". In: *Proceedings of the International Conference on Web Search and Data Mining*. WSDM '08. Association for Computing Machinery, Feb. 2008, pp. 219–230. DOI: 10.1145/1341531.1341560.

[94] Qiang Cao et al. "Aiding the detection of fake accounts in large scale social online services". In: *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. NSDI'12. Apr. 2012.

[95] Abedelaziz Mohaisen, Nicholas Hopper, and Yongdae Kim. "Keep your friends close: Incorporating trust into social network-based Sybil defenses". In: *2011 Proceedings IEEE INFOCOM*. 2011, pp. 1943–1951. DOI: 10.1109/infcom.2011.5934998.

[96] Gang Wang et al. *Social Turing Tests: Crowdsourcing Sybil Detection*. arXiv:1205.3856. Dec. 2012. (Visited on 03/28/2024).

[97] Teng Xu et al. "Deep Entity Classification: Abusive Account Detection for Online Social Networks". In: *Proceedings of the 30th USENIX Security Symposium*. USENIX '21. 2021, pp. 4097–4114.

[98] Kai-Cheng Yang et al. "Scalable and generalizable social bot detection through data selection". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. AAAI '20 01. 2020, pp. 1096–1103.

[99] Nitin Jindal and Bing Liu. "Review spam detection". In: *Proceedings of the 16th international conference on World Wide Web*. WWW '07. ACM, May 2007, pp. 1189–1190. DOI: 10.1145/1242572.1242759.

[100] Shagun Jhaver et al. "Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor". In: *Proceedings of the ACM on Human-Computer Interaction*. CSCW '23 7.Cscw2 (Oct. 2023), 289:1–289:33.

[101] Charlotte Schluger et al. "Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support". In: *Proceedings of the ACM on Human-Computer Interaction*. CSCW '22 6.Cscw2 (2022), 370:1–370:27. DOI: 10.1145/3555095.

[102] Jon Roozenbeek et al. "Psychological inoculation improves resilience against misinformation on social media". In: *Science Advances* 8.34 (2022). DOI: 10.1126/sciadv.abo6254. URL: https://www.science.org/doi/abs/10.1126/sciadv.abo6254.

[103] Cameron Martel and David G. Rand. "Fact-checker warning labels are effective even for those who distrust fact-checkers". In: *Nature Human Behaviour* 8 (Sept. 2024), pp. 1957–1967. DOI: 10.1038/s41562-024-01858-1. URL: https://www.nature.com/articles/s41562-024-01858-1.

[104] Chen Ling, Krishna P. Gummadi, and Savvas Zannettou. "Learn the Facts about COVID-19: Analyzing the Use of Warning Labels on TikTok Videos". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '23 17.1 (June 2023), pp. 554–565. DOI: 10.1609/icwsm.v17i1.22168.

[105] Tatiana Celadin et al. "Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts". In: *Journal of Online Trust and Safety* 1.5 (Apr. 2023). DOI: 10.54501/jots.v1i5.100. URL: https://tsjournal.org/index.php/jots/article/view/100.

[106] Thao Ngo et al. "Spot the bot: Investigating user's detection cues for social bots and their willingness to verify Twitter profiles". In: *Computers in Human Behavior* 146 (Sept. 2023), p. 107819. DOI: 10.1016/j.chb.2023.107819.

[107] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. "Finding and assessing social media information sources in the context of journalism". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 2451–2460. ISBN: 9781450310154. DOI: 10.1145/2207676.2208409. URL: https://doi.org/10.1145/2207676.2208409.

[108] Jane Im et al. "Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376383. URL: https://doi.org/10.1145/3313831.3376383.

[109] Alejandro Cuevas, Manoel Horta Ribeiro, and Nicolas Christin. *Chameleon Channels: Measuring YouTube Accounts Repurposed for Deception and Profit*. 2025. arXiv: 2507.16045 [cs.CY]. URL: https://arxiv.org/abs/2507.16045.

[110] Mario Beluri et al. *Exploration of the Dynamics of Buy and Sale of Social Media Accounts*. 2024. arXiv: 2412.14985 [cs.CR]. URL: https://arxiv.org/abs/2412.14985.

[111] Steven Tadelis and Oliver E. Williamson. "Transaction Cost Economics". In: *The Handbook of Organizational Economics*. Ed. by Robert Gibbons and John Roberts. Princeton University Press, 2013, pp. 159–190. DOI: 10.1515/9781400845354-006.

[112] Tavish Vaidya et al. "Does Being Verified Make You More Credible?: Account Verification's Effect on Tweet Credibility". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '19. Association for Computing Machinery, 2019, pp. 1–13. DOI: 10.1145/3290605.3300755.

[113] Haitao Xu et al. "E-commerce Reputation Manipulation: The Emergence of Reputation-Escalation-as-a-Service". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. 2015, pp. 1296–1306. DOI: 10.1145/2736277.2741650.

[114] Gianluca Stringhini et al. "Follow the green: growth and dynamics in twitter follower markets". In: *Proceedings of the Conference on Internet Measurement Conference*. IMC '13. Association for Computing Machinery, 2013, pp. 163–176. DOI: 10.1145/2504730.2504731.

[115] David Nevado-Catalan et al. "An analysis of fake social media engagement services". In: *Computers & Security* 124 (2023), p. 103013. ISSN: 0167-4048. DOI:

https://doi.org/10.1016/j.cose.2022.103013. URL: https://www.sciencedirect.com/science/article/pii/S0167404822004059.

[116] Gianluca Stringhini et al. "Poultry markets: on the underground economy of twitter followers". In: *SIGCOMM Computer Communication Review* 42.4 (Sept. 2012), pp. 527–532. DOI: 10.1145/2377677.2377781.

[117] Manuel Egele et al. "Compa: Detecting compromised accounts on social networks." In: *Ndss*. Vol. 13. 2013, pp. 83–91.

[118] Gang Wang et al. "Serf and turf: crowdturfing for fun and profit". In: *Proceedings of the 21st international conference on World Wide Web*. WWW '12. Association for Computing Machinery, Apr. 2012, pp. 679–688. DOI: 10.1145/2187836.2187928.

[119] Marti Motoyama et al. "Dirty jobs: the role of freelance labor in web service abuse". In: *Proceedings of the 20th USENIX Conference on Security*. USENIX'11. 2011, p. 14.

[120] Shehroze Farooqi et al. "Measuring and mitigating oauth access token abuse by collusion networks". In: *Proceedings of the 2017 Internet Measurement Conference*. Imc '17. Association for Computing Machinery, 2017, pp. 355–368. ISBN: 9781450351188. DOI: 10.1145/3131365.3131404. URL: https://doi.org/10.1145/3131365.3131404.

[121] Rafael Grohmann et al. "Click farm platforms: An updating of informal work in Brazil and Colombia". In: *Work Organisation, Labour, and Globalisation* 16.2 (2022), pp. 7–20. ISSN: 1745641x, 17456428. URL: https://www.jstor.org/stable/48691511 (visited on 06/04/2025).

[122] Manuel Egele et al. "Towards Detecting Compromised Accounts on Social Networks". In: *IEEE Transactions on Dependable and Secure Computing* 14.4 (2017), pp. 447–460. DOI: 10.1109/tdsc.2015.2479616.

[123] U.S. Department of Justice. *Alabama Man Pleads Guilty in Connection with Securities and Exchange Commission X Account Hack*. Accessed: 2025-06-04. Mar. 2025. URL: https://www.justice.gov/opa/pr/alabama-man-pleads-guilty-connection-securities-and-exchange-commission-x-account-hack.

[124] Oliver Knight. "Hack of Vitalik Buterin's X Account Leads to 691K Stolen". In: *CoinDesk* (Sept. 2023). Accessed: 2025-06-04. URL: https://www.coindesk.com/business/2023/09/11/691k-stolen-as-hackers-take-over-vitalik-buterins-x-account.

[125] Enze Liu et al. "Give and Take: An End-To-End Investigation of Giveaway Scam Conversion Rates". In: *Proceedings of the ACM on Internet Measurement Conference*. IMC '24. Association for Computing Machinery, 2024, pp. 704–712. DOI: 10.1145/3646547.3689005. URL: https://doi.org/10.1145/3646547.3689005.

[126] Meredith Clark. "Hawk Tuah girl faces 'pump and dump' allegations as crypto coin collapses hours after launch". In: *The Independent* (Dec. 2024). Accessed: 2025-06-04. URL: https://www.the-independent.com/life-style/hawk-tuah-crypto-coin-welch-scam-rugpull-b2659849.html.

[127] Joel Khalili. "The Memecoin Shenanigans Are Just Getting Started". In: *Wired* (Jan. 2025). Accessed: 2025-06-04. URL: https://www.wired.com/story/memecoins-cryptocurrency-regulation/.

[128] Sarah Grevy Gotfredsen. "The Tenet Media Incident". In: *Columbia Journalism Review* (Sept. 2024). Accessed: 2025-06-04. URL: https://www.cjr.org/the_media_today/tenet_media_indictment_russia.php.

[129] Mohammad Hammas Saeed et al. "Unraveling the Web of Disinformation: Exploring the Larger Context of State-Sponsored Influence Campaigns on Twitter". In: *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses*. RAID '24. 2024, pp. 353–367. DOI: 10.1145/3678890.3678911.

[130] Jeff Giesea. "It's Time to Embrace Memetic Warfare". In: *Defence Strategic Communications* 1 (2016). Accessed: 2025-06-04, pp. 67–75. DOI: 10.30966/2018.riga.1.4. URL: https://stratcomcoe.org/publications/its-time-to-embrace-memetic-warfare/164.

[131] Michael B. Prosser. *Memetics–A Growth Industry in U.S. Military Operations*. Master's thesis Ada507172. Accessed: 2025-06-04. Marine Corps University, School of Advanced Warfighting, Dec. 2005. URL: https://apps.dtic.mil/sti/pdfs/ADA507172.pdf.

[132] Rafael Grohmann and Jonathan Corpus Ong. "Disinformation-for-Hire as Everyday Digital Labor: Introduction to the Special Issue". In: *Social Media + Society* 10.1 (2024). DOI: 10.1177/20563051231224723. URL: https://doi.org/10.1177/20563051231224723.

[133] Odanga Madung and Brian Obilo. *Fellow Research: Inside the Shadowy World of Disinformation-for-hire in Kenya*. Accessed: 2025-06-04. Sept. 2021. URL: https://www.mozillafoundation.org/en/blog/fellow-research-inside-the-shadowy-world-of-disinformation-for-hire-in-kenya/.

[134] Michael Safi and Stephanie Kirchgaessner. "The secret world of disinformation for hire". In: *The Guardian* (Feb. 2023). Accessed: 2025-06-04. URL: https://www.theguardian.com/news/audio/2023/feb/22/the-secret-world-of-disinformation-for-hire-podcast.

[135] Catherine Han, Deepak Kumar, and Zakir Durumeric. "On the Infrastructure Providers That Support Misinformation Websites". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '22 16.1 (May 2022), pp. 287–298. DOI: 10.1609/icwsm.v16i1.19292.

[136] Enrico Mariconti et al. "What's in a Name? Understanding Profile Name Reuse on Twitter". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. 2017, pp. 1161–1170. DOI: 10.1145/3038912.3052589.

[137] Jonah Gibbon et al. "Measuring the Unmeasurable: Estimating True Population of Hidden Online Communities". In: *IEEE European Symposium on Security and Privacy Workshops*. EuroS&PW '24. 2024, pp. 56–66.

[138] Swapd. *SWAPD – Digital Rights Marketplace*. Accessed: 2025-06-04. 2025. URL: https://swapd.co/.

[139] Accs-Market. *Accs-Market – Secure Social Media Account Marketplace*. Accessed: 2025-06-04. 2025. URL: https://accs-market.com/.

[140] ViralAccounts. *ViralAccounts – Buy & Sell Social Media Influence*. Accessed: 2025-06-04. 2025. URL: https://viralaccounts.com/.

[141] Fameswap. *Fameswap – Marketplace for Social Media Accounts*. Accessed: 2025-06-04. 2025. URL: https://fameswap.com/.

[142] The YouTube Team. *Introducing Handles: A New Way to Identify Your YouTube Channel*. Accessed: 2025-06-04. Oct. 2022. URL: https://blog.youtube/news-and-events/introducing-handles-a-new-way-to-identify-your-youtube-channel/.

[143] Manoel Horta Ribeiro and Robert West. "YouNiverse: Large-Scale Channel and Video Metadata from English-Speaking YouTube". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '21 15.1 (May 2021), pp. 1016–1024. DOI: 10.1609/icwsm.v15i1.18125.

[144] Manoel Horta Ribeiro et al. "The Evolution of the Manosphere across the Web". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '21 15.1 (May 2021), pp. 196–207. DOI: 10.1609/icwsm.v15i1.18053.

[145] Social Blade LLC. *Social Blade – YouTube, Instagram, Twitch, TikTok, and More Statistics*. Accessed: 2025-06-04. 2025. URL: https://socialblade.com/.

[146] Wikipedia contributors. *Social Blade*. Accessed: 2025-06-04. 2025. URL: https://en.wikipedia.org/wiki/Social_Blade.

[147] Manoel Horta Ribeiro et al. "Auditing radicalization pathways on YouTube". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAccT '20. Association for Computing Machinery, 2020, pp. 131–141. DOI: 10.1145/3351095.3372879.

[148] Gabriel Luis Santos Freire et al. "Understanding Effects of Moderation and Migration on Online Video Sharing Platforms". In: *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*. Ht '22. Association for Computing

Machinery, 2022, pp. 220–224. ISBN: 9781450392334. DOI: 10.1145/3511095.3536377. URL: https://doi.org/10.1145/3511095.3536377.

[149] Wikipedia contributors. *youtube-dl*. Accessed: 2025-06-04. 2025. URL: https://en.wikipedia.org/wiki/Youtube-dl.

[150] Reddit contributors. *r/youtubedl – Community for youtube-dl and yt-dlp*. Accessed: 2025-06-04. 2025. URL: https://www.reddit.com/r/youtubedl/.

[151] Fabian Retkowski and Alexander Waibel. "From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. EACL '24. Association for Computational Linguistics, Mar. 2024, pp. 406–419. URL: https://aclanthology.org/2024.eacl-long.25/.

[152] Internet Archive. *Wayback Machine*. Accessed: 2025-06-04. 2025. URL: https://web.archive.org/.

[153] Wikipedia contributors. *Wayback Machine*. Accessed: 2025-06-04. 2025. URL: https://en.wikipedia.org/wiki/Wayback_Machine.

[154] Caleb Ziems et al. "Can large language models transform computational social science?" In: *Computational Linguistics* 50.1 (2024), pp. 237–291.

[155] Matthew Renze. "The Effect of Sampling Temperature on Problem Solving in Large Language Models". In: *Empirical Methods in Natural Language Processing*. Findings of EMNLP '24. Association for Computational Linguistics, Nov. 2024, pp. 7346–7356. DOI: 10.18653/v1/2024.findings-emnlp.432.

[156] Azwad Anjum Islam and Mark A. Finlayson. "A Semantic Interpreter for Social Media Handles". In: *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM '24 18.1 (May 2024), pp. 676–690. DOI: 10.1609/icwsm.v18i1.31343.

[157] Ralph Peeters and Christian Bizer. "Using ChatGPT for Entity Matching". In: *New Trends in Database and Information Systems*. Springer Nature Switzerland, 2023, pp. 221–230. ISBN: 978-3-031-42941-5.

[158] Hema Yoganarasimhan. "Impact of social network structure on content propagation: A study using YouTube data". In: *Quantitative Marketing and Economics* 10 (2012), pp. 111–150.

[159] Chen Ling et al. "Slapping Cats, Bopping Heads, and Oreo Shakes: Understanding Indicators of Virality in TikTok Short Videos". In: *Proceedings of the 14th ACM Web Science Conference*. WebSci '22. Association for Computing Machinery, 2022, pp. 164–173. ISBN: 9781450391917. DOI: 10.1145/3501247.3531551.

[160] Bertie Vidgen et al. “Challenges and frontiers in abusive content detection”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Aug. 2019, pp. 80–93. DOI: 10.18653/v1/W19-3509.

[161] YouTube. *Child safety policy*. Accessed: 2025-06-04. 2024. URL: https://support.google.com/youtube/answer/2801999?hl=en.

[162] YouTube. *YouTube Copyright Policy*. Accessed: 2025-06-04. 2025. URL: https://support.google.com/youtube/topic/2676339?hl=en&ref_topic=6151248.

[163] YouTube. *Misinformation Policies*. Accessed: 2025-06-04. 2025. URL: https://support.google.com/youtube/answer/10834785.

[164] YouTube. *Medical Misinformation Policy*. Accessed: 2025-06-04. 2025. URL: https://support.google.com/youtube/answer/13813322.

[165] YouTube. *Elections Misinformation Policies*. Accessed: 2025-06-04. 2025. URL: https://support.google.com/youtube/answer/10835034.

[166] Muhammad Haroon, Magdalena Wojcieszak, and Anshuman Chhabra. *””Whose Side Are You On?”” Estimating Ideology of Political and News Content Using Large Language Models and Few-shot Demonstration Selection*. 2025. arXiv: 2503.20797 [cs.CL]. URL: https://arxiv.org/abs/2503.20797.

[167] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. “Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere”. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2018, pp. 69–76. DOI: 10.1109/asonam.2018.8508247.

[168] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. “Deradicalizing YouTube: Characterization, Detection, and Personalization of Religiously Intolerant Arabic Videos”. In: *Proc. ACM Hum.-Comput. Interact.* 6.Cscw2 (Nov. 2022). DOI: 10.1145/3555618. URL: https://doi.org/10.1145/3555618.

[169] YouTube. *Illegal or regulated goods or services policies*. Accessed: 2025-06-04. 2024. URL: https://support.google.com/youtube/answer/9229611?hl=en.

[170] YouTube. *Spam, deceptive practices, & scams policies*. Accessed: 2025-06-04. 2024. URL: https://support.google.com/youtube/answer/2801973?hl=en.

[171] Daisuke Kawai et al. “Is your digital neighbor a reliable investment advisor?” In: *Proceedings of the ACM Web Conference*. WWW ’23. Association for Computing Machinery, 2023, pp. 3581–3591. DOI: 10.1145/3543507.3583502.

[172] Tianle Li et al. *Long-context LLMs Struggle with Long In-context Learning*. 2024. arXiv: 2404.02060 [cs.CL]. URL: https://arxiv.org/abs/2404.02060.

[173] Kostantinos Papadamou et al. “Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children”. In: *Proceed-*

ings of the International AAAI Conference on Web and Social Media. ICWSM '20 14.1 (May 2020), pp. 522–533. DOI: 10.1609/icwsm.v14i1.7320.

[174] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. "ChatGPT Outperforms Crowd Workers For Text-Annotation Tasks". In: *Proceedings of the National Academy of Sciences*. PNAS '23 120.30 (2023). DOI: 10.1073/pnas.2305016120.

[175] Claire Wonjeong Jo, Miki Wesołowska, and Magdalena Wojcieszak. *Harmful YouTube Video Detection: A Taxonomy of Online Harm and MLLMs as Alternative Annotators*. 2024. arXiv: 2411.05854 [cs.MM]. URL: https://arxiv.org/abs/2411.05854.

[176] Naoki Egami et al. "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 68589–68601.

[177] Naoki Egami et al. *dsl: Design-based Supervised Learning*. R package version 0.1.0. 2025. URL: http://naokiegami.com/dsl/.

[178] R. van Wegberg et al. "Plug and prey? Measuring the commoditization of cybercrime via online anonymous markets". In: *Proceedings of the 27th USENIX Security Symposium*. USENIX '18. Aug. 2018, pp. 1009–1026.

[179] Gianluca Stringhini et al. "The harvester, the botmaster, and the spammer: on the relations between the different actors in the spam landscape". In: *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*. Asia CCS '14. Association for Computing Machinery, 2014, pp. 353–364. DOI: 10.1145/2590296.2590302.

[180] KTDY Huang et al. "Framing dependencies introduced by underground commoditization". In: *Workshop on the Economics of Information Security*. 2015.

[181] Felix M. Simon, Sacha Altay, and Hugo Mercier. "Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown". In: *Harvard Kennedy School Misinformation Review* 4.5 (Oct. 2023). DOI: 10.37016/mr-2020-127.

[182] Anh V. Vu, Alice Hutchings, and Ross Anderson. "No Easy Way Out: the Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment". In: *2024 IEEE Symposium on Security and Privacy (SP)*. 2024, pp. 717–734. DOI: 10.1109/sp54263.2024.00007. URL: https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00007.

[183] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. "A Survey on Automated Fact-Checking". In: *Transactions of the Association for Computational Linguistics* 10 (Feb. 2022), pp. 178–206. ISSN: 2307-387x. DOI: 10.1162/tacl_a_00454.

[184] Mohammad Karami, Youngsam Park, and Damon McCoy. "Stress Testing the Booters: Understanding and Undermining the Business of DDoS Services". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. 2016, pp. 1033–1043. DOI: 10.1145/2872427.2883004.

[185] K. Levchenko et al. "Click Trajectories: End-to-End Analysis of the Spam Value Chain". In: *IEEE Symposium on Security and Privacy*. S&P '11. May 2011.

[186] C. Kanich et al. "Show Me the Money: Characterizing Spam-advertised Revenue". In: *Proceedings of the 20th USENIX Security Symposium*. USENIX '11. Aug. 2011.

[187] Ariana Mirian et al. "Hack for Hire: Exploring the Emerging Market for Account Hijacking". In: *The World Wide Web Conference*. WWW '19. Association for Computing Machinery, 2019, pp. 1279–1289. DOI: 10.1145/3308558.3313489.

[188] Michele Campobasso and Luca Allodi. "Know Your Cybercriminal: Evaluating Attacker Preferences by Measuring Profile Sales on an Active, Leading Criminal Market for User Impersonation at Scale". In: *Proceedings of USENIX Security 2023*. Anaheim, CA, Aug. 2023.

[189] Catherine Han et al. "Characterizing the MrDeepFakes Sexual Deepfake Marketplace". In: *Proceedings of USENIX Security 2025*. Seattle, WA, Aug. 2025.

[190] Cassidy Gibson et al. "Analyzing the AI Nudification Application Ecosystem". In: *Proceedings of USENIX Security 2025*. Seattle, WA, Aug. 2025.

[191] *Internet Archive Developer Portal*. URL: https://archive.org/developers/index.html (visited on 08/14/2025).

[192] Kyle Beadle et al. "SoK: A Privacy Framework for Security Research Using Social Media Data". In: *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, May 2025, pp. 1178–1196. DOI: 10.1109/sp61157.2025.00145. URL: https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00145.

[193] Wikipedia contributors. *List of social networking services*. Accessed: 2025-06-05. 2025. URL: https://en.wikipedia.org/wiki/List_of_social_networking_services.

[194] Judith Aldridge and David Décary-Hétu. "Hidden wholesale: The drug diffusing capacity of online drug cryptomarkets". In: *International Journal of Drug Policy* 35 (Sept. 2016), pp. 7–15. DOI: 10.1016/j.drugpo.2016.04.020. (Visited on 09/02/2020).

[195] M. Dittus, J. Wright, and M. Graham. "Platform Criminalism: The 'Last-Mile' Geography of the Darknet Market Supply Chain". In: *Proc. Web Conf.* 2018, pp. 277–286.

[196] C. Zhang, R. Wei, and X. Liu. "Drugs and bitcoins: What role do bitcoins play in the darknet market? A preliminary study". In: *Proceedings of the Association for Information Science and Technology* 55 (2018), pp. 944–945.

[197] K. Turk, S. Pastrana, and B. Collier. "A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments". In: *Proceedings of the IEEE European Symposium on Security and Privacy Workshops*. EuroS&PW '20. 2020, pp. 428–437.

[198] N. Christin. *An EU-focused analysis of drug supply on the AlphaBay marketplace*. EMCDDA commissionned paper. Oct. 2017.

[199] J. van de Laarschot and R. van Wegberg. "Risky Business? Investigating the Security Practices of Vendors on an Online Anonymous Market using Ground-Truth Data". In: *Proceedings of the 30th USENIX Security Symposium*. USENIX '21. Aug. 2021, pp. 4079–4095.

[200] X. Wang et al. "You Are Your Photographs: Detecting Multiple Identities of Vendors in the Darknet Marketplaces". In: *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. Asia CCS '18. Association for Computing Machinery, 2018, pp. 431–442.

[201] Y. Zhang et al. "Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network". In: *The World Wide Web Conference*. WWW '19. Association for Computing Machinery, May 2019, pp. 3448–3454.

[202] X. Tai, K. Soska, and N. Christin. "Adversarial Matching of Dark Net Market Vendor Accounts". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Association for Computing Machinery, July 2019, pp. 1871–1880.

[203] R. van Wegberg and T. Verburgh. "Lost in the Dream? Measuring the effects of Operation Bayonet on vendors migrating to Dream Market". In: *Proceedings of the Evolution of the Darknet Workshop at Web. Sci. Conf.* 2018, pp. 1–5.

[204] A. Celestini, G. Me, and N. Mignone. "Tor Marketplaces Exploratory Data Analysis: The Drugs Case". In: *Proc. ICG3S*. Vol. 630. Jan. 2016, pp. 218–229.

[205] A. Baravalle, M. Lopez, and S. Lee. "Mining the Dark Web: Drugs and Fake Ids". In: *Proc. IEEE ICDM 2016 Workshops*. Dec. 2016, pp. 350–356.

[206] A. Baravalle and S. Lee. "Dark Web Markets: Turning the Lights on AlphaBay". In: *The Workshop on the Economics of Information Security*. WEIS '18. Nov. 2018, pp. 502–514.

[207] D. Hayes, F. Cappa, and J. Cardon. "A Framework for More Effective Dark Web Marketplace Investigations". In: *Information* 9.8 (July 2018), p. 186.

[208] J. Aldridge and D. Décary-Hétu. *Not an 'Ebay for Drugs': The Cryptomarket 'Silk Road' as a Paradigm Shifting Criminal Innovation*. Available at SSRN: https://ssrn.com/abstract=2436643. May 2014.

[209] D. Dolliver. "Evaluating drug trafficking on the Tor Network: Silk Road 2, the sequel". In: *Int. J. Drug Pol.* 26.11 (Nov. 2015), pp. 1113–1123.

[210] S. Ahmad et al. "Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web". In: *Proc. Web Conf.* May 2020, pp. 271–280.

[211] J. Aldridge and D. Décary-Hétu. "Cryptomarkets and the future of illicit drug markets". In: *The Internet and drug markets*. Emcdda, 2015, pp. 23–32.

[212] Anonymous. *Grams: Search the Darknet*. Was at http://grams7enufi7jmdl.onion. Taken offline in December 2017. 2017.

[213] Q. Rossy et al. *Drogues sur Internet: État des lieux sur la situation en Suisse*. Tech. rep. 98. Addiction Suisse & ESC/UNIL, Nov. 2018.

[214] C. Bradley. "On the resilience of the Dark Net Market ecosystem to law enforcement intervention". PhD thesis. University College London, 2019.

[215] D. McCoy et al. "PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs". In: *Proceedings of the 21st USENIX Security Symposium*. USENIX '12. Aug. 2012.

[216] K. Kruithof et al. *Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands*. RAND corporation. https://www.rand.org/pubs/research_reports/RR1607.html. 2016.

[217] S. Lewis. *OnionScan Report: Reconstructing the Finances of Darknet Markets through Reputation Systems*. Jan. 2017. URL: https://mascherari.press/onionscan-report-forensic-finances-dark-markets/ (visited on 05/27/2021).

[218] *Scrapy: An open source web scraping framework for Python (v.1.0.0-1.4.0)*. http://scrapy.org.

[219] R. Dingledine, N. Mathewson, and P. Syverson. "Tor: The Second-Generation Onion Router". In: *Proceedings of the USENIX Security Symposium*. Aug. 2004.

[220] Europol. *Massive blow to criminal Dark Web activities after globally coordinated operation*. https://www.europol.europa.eu/media-press/newsroom/news/massive-blow-to-criminal-dark-web-activities-after-globally-coordinated-operation. 2017. URL: https://www.europol.europa.eu/media-press/newsroom/news/massive-blow-to-criminal-dark-web-activities-after-globally-coordinated-operation.

[221] H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *Ann. Math. Stat.* 18.1 (1947), pp. 50–60.

[222] K. Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *London Edinburgh and Dublin Philosophical Magazine and Journal of Science* 50.302 (July 1900), pp. 157–175.

[223] *CoinCap API 2.0*. https://docs.coincap.io. (Visited on 09/23/2021).

[224] J. Broséus et al. "Studying illicit drug trafficking on Darknet markets: Structure and organisation from a Canadian perspective". In: *Forensic Sci. Int.* 264 (July 2016), pp. 7–14.

[225] Z. Schnabel. "The estimation of the total fish population of a lake". In: *American Mathematical Monthly* 45.6 (1938), pp. 348–352.

[226] Carl James Schwarz and A. Neil Arnason. "A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations". In: *Biometrics* 52.3 (Sept. 1996), p. 860. DOI: 10.2307/2533048. (Visited on 05/27/2021).

[227] New York Power Authority. *Use of Buckhorn Marsh and Grand island tributaries by northern pike for spawning and as a nursery*. Tech. rep. FERC, 2004.

[228] D. Ogle. *fishR Vignette – Closed Mark-Recapture Abundance Estimates*. Dec. 2013. URL: http://derekogle.com/fishR/examples/oldFishRVignettes/MRClosed.pdf.

[229] G. White and K. Burnham. "Program MARK: survival estimation from populations of marked animals". In: *Bird Study* 46 (1999), S120–s139.

[230] N. Christin and J. Thomas. *An analysis of the supply of drugs and new psychoactive substances by EU-based vendors via darknet markets in 2017–18*. EMCDDA commissionned paper. Nov. 2019.

[231] R. Portnoff et al. "Tools for Automated Analysis of Cybercriminal Markets". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. 2017, pp. 657–666.

[232] P. Andreas and K. Greenhill, eds. *Sex, Drugs, and Body Counts: The Politics of Numbers in Global Crime and Conflict*. Cornell University Press, 2010. ISBN: 9780801476181. URL: http://www.jstor.org/stable/10.7591/j.ctt7zg8b (visited on 05/13/2022).

[233] A. Noroozian et al. "Platforms in Everything: Analyzing Ground-Truth Data on the Anatomy and Economics of Bullet-Proof Hosting". In: *Proceedings of the 28th USENIX Security Symposium*. USENIX '19. Aug. 2019.

[234] S. Hao et al. "Drops for Stuff: An Analysis of Reshipping Mule Scams". In: *Proceedings of the SIGSAC Conference on Computer and Communications Security*. CCS '15. Association for Computing Machinery, Oct. 2015, pp. 1081–1092.

[235] J. Martin and N. Christin. "Ethics in Cryptomarket Research". In: *International Journal of Drug Policy* 25 (6 2016), pp. 84–91.

[236] Heski Bar-Isaac and Steven Tadelis. "Seller Reputation". In: *Foundations and Trends in Microeconomics* 4.4 (2008), pp. 273–351. DOI: 10.1561/0700000027.

[237] T. J. Holt. "Examining the Forces Shaping Cybercrime Markets Online". In: *Social Science Computer Review* 31.2 (2013), pp. 165–177.

[238] Rolf van Wegberg et al. "Go See a Specialist? Predicting Cybercrime Sales on Online Anonymous Markets from Vendor and Product Characteristics". In: *Proceedings of The Web Conference*. WWW '20. Association for Computing Machinery, 2020, pp. 816–826. DOI: 10.1145/3366423.3380162.

[239] Carlo Morselli et al. "Conflict Management in Illicit Drug Cryptomarkets". In: *International Criminal Justice Review* 27.4 (2017), pp. 237–254.

[240] P. Reuter and J. P. Caulkins. "Illegal 'Lemons': Price Dispersion in Cocaine and Heroin Markets". In: *Bulletin on Narcotics* 56.1-2 (2004), pp. 141–165.

[241] J. Martin. *Drugs on the Dark Net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs*. Springer, 2014.

[242] M. Tzanetakis et al. "The Transparency Paradox. Building Trust, Resolving Disputes and Optimising Logistics on Conventional and Online Drugs Markets". In: *International Journal of Drug Policy* 35 (2016), pp. 58–68.

[243] M. J. Barratt, J. A. Ferris, and A. R. Winstock. "Use of Silk Road, the Online Drug Marketplace, in the United Kingdom, Australia and the United States". In: *Addiction* 109.5 (2014), pp. 774–783.

[244] M. J. Barratt, J. A. Ferris, and A. R. Winstock. "Safer Scoring? Cryptomarkets, Social Supply and Drug market Violence". In: *International Journal of Drug Policy* 35 (2016), pp. 24–31.

[245] T. M. Booij et al. "Get Rich or Keep Tryin' Trajectories in Dark Net Market Vendor careers". In: *Proceedings of the IEEE European Symposium on Security and Privacy Workshops*. EuroS&PW '21. 2021, pp. 202–212.

[246] Diego Gambetta. *Codes of the Underworld: How Criminals Communicate*. Princeton University Press, 2009.

[247] Y. Zhang et al. "Key Player Identification in Underground Forums over Attributed Heterogeneous Information Network Embedding Framework". In: *28th ACM International Conference on Information and Knowledge Management (CIKM'19)*. Nov. 2019, pp. 549–558.

[248] A. Caines et al. "Automatically identifying the function and intent of posts in underground forums". In: *Crime Science* 7.1 (2018), p. 19.

[249] J. Hughes, B. Collier, and A. Hutchings. "From Playing Games to Committing Crimes: A Multi-Technique Approach to Predicting Key Actors on an Online

Gaming Forum". In: *APWG Symposium on Electronic Crime Research*. eCrime '19. Nov. 2019, pp. 1–12.

[250] S. Pastrana et al. "Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum". In: *International Symposium on Research in Attacks, Intrusions, and Defenses*. RAID '18. Sept. 2018, pp. 207–227.

[251] M. Motoyama et al. "An analysis of underground forums". In: *ACM Internet Measurement Conference*. IMC '11. Nov. 2011, pp. 71–80.

[252] Z. Sun et al. "Understanding and Predicting Private Interactions in Underground Forums". In: *9th ACM Conference on Data and Application Security and Privacy*. CODASPY '19. Mar. 2019, pp. 303–314.

[253] R. Overdorf et al. "Under the Underground: Predicting Private Interactions in Underground Forums". In: 2018. eprint: 1805.04494 (cs.CR).

[254] A. Maddox et al. "Constructive Activism in the Dark Web: Cryptomarkets and Illicit Drugs in the Digital 'Demimonde'". In: *Information, Communication & Society* 19.1 (2016), pp. 111–126.

[255] T. J. Holt. "Exploring the Social Organisation and Structure of Stolen Data Markets". In: *Global Crime* 14.2-3 (2013), pp. 155–174.

[256] T. J. Holt et al. "Examining the Risk Reduction Strategies of Actors in Online Criminal Markets". In: *Global Crime* 16.2 (2015), pp. 81–103.

[257] F. Wehinger. "The Dark Net: Self-Regulation Dynamics of Illegal Online Markets for Identities and Related Services". In: *European Intelligence and Security Informatics Conference*. Sept. 2011, pp. 209–213.

[258] L. Franceschi-Bicchierai. *Reddit Bans Subreddits Dedicated to Dark Web Drug Markets and Selling Guns*. Vice. 2018. URL: https://www.vice.com/en/article/ne9v5k/reddit-bans-subreddits-dark-web-drug-markets-and-guns.

[259] *Pushshift*. https://pushshift.io/. Accessed: September 19, 2023.

[260] 'u/lift_ticket83'. *Reddit Data API Update: Changes to Pushshift Access*. Website. Accessed: September 19, 2023. URL: https://www.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/.

[261] stuck-in-the-matrix and Watchful1. *'Reddit comments/submissions 2005-06 to 2022-12'*.

[262] D. J. Nutt, L. A. King, L. D. Phillips, et al. "Drug Harms in the UK: A Multicriteria Decision Analysis". In: *The Lancet* 376.9752 (2010), pp. 1558–1565.

[263] T. J. Holt, O. Smirnova, and A. Hutchings. "Examining Signals of Trust in Criminal Markets Online". In: *Journal of Cybersecurity* 2.2 (2016), pp. 137–145.

[264] M. Paquet-Clouston, D. Décary-Hétu, and C. Morselli. "Assessing market competition and vendors' size and scope on AlphaBay". In: *International Journal of Drug Policy* 54 (May 2018), pp. 87–98. (Visited on 09/02/2020).

[265] J. Franklin et al. "An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants". In: *14th ACM Conference on Computer and Communications Security*. Vol. 7. CCS '07. Oct. 2007, pp. 375–388.

[266] M. Pal. "Random Forest Classifier for Remote Sensing Classification". In: *International Journal of Remote Sensing* 26.1 (2005), pp. 217–222. DOI: 10.1080/01431160412331269698.

[267] H. Deng et al. "A Time Series Forest for Classification and Feature Extraction". In: *Information Sciences* 239 (2013), pp. 142–153.

[268] A. Greenberg. *How Dutch Police Took Over Hansa, a Top Dark Web Market*. Accessed: September 19, 2023. URL: https://www.wired.com/story/hansa-dutch-police-sting-operation/.

[269] R. Mac. *Silk Road 2.0's Blake Benthall Arrested, Charged With Running The Massive Dark Web Drug Site*. https://www.forbes.com/sites/ryanmac/2014/11/06/silk-road-2-blake-benthall-fbi-shutdown/?sh=7bdde138170f. 2014.

[270] E. Hammersvik, S. Sandberg, and W. Pedersen. "Why Small-scale Cannabis Growers Stay Small: Five Mechanisms that Prevent Small-scale Growers from Going Large Scale". In: *International Journal of Drug Policy* 23.6 (2012), pp. 458–464.

[271] F. E. Harrell, K. L. Lee, and D. B. Mark. "Multivariate Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors". In: *Statistics in Medicine* 15 (1996), pp. 361–387.

[272] Z. Xiao et al. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding". In: *28th ACM International Conference on Intelligent User Interfaces*. IUI '23. 2023, pp. 75–78.

[273] P. Törnberg. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning". In: arXiv (2304.06588), 2023. eprint: 2304.06588 (cs.CL).

[274] Z Banković et al. "Detecting bad-mouthing attacks on reputation systems using self-organizing maps". In: *Proceedings of Computational Intelligence in Security for Information Systems: 4th International Conference (CISIS 2011)*. June 2011, pp. 9–16.

[275] Satoshi Nakamoto. *Bitcoin: A peer-to-peer electronic cash system*. Tech. rep. 2008.

[276] Clemens Graf von Luckner, Carmen M Reinhart, and Kenneth Rogoff. "Decrypting new age international capital flows". In: *Journal of Monetary Economics* (2023).

[277] Lin William Cong et al. "An anatomy of crypto-enabled cybercrimes". In: *Available at SSRN 4188661* (2022).

[278] Emma Fletcher. *Reports show scammers cashing in on crypto craze*. July 2022. URL: https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/06/reports-show-scammers-cashing-crypto-craze%5C#crypto1.

[279] David Yermack. "Is Bitcoin a real currency? An economic appraisal". In: *Handbook of digital currency*. Elsevier, 2015, pp. 31–43.

[280] Kyle Soska et al. "Towards understanding cryptocurrency derivatives: A case study of BitMEX". In: *WWW'21: Proceedings of the ACM Web Conference 2021, Ljubljana, Slovenia (online)*. 2021.

[281] Global Legal Research Directorate The Law Library of Congress. *Regulation of cryptocurrency around the world: November 2021 update*. https://tile.loc.gov/storage-services/service/ll/llglrd/2021687419/2021687419.pdf?roistat_visit=513132. The Law Library of Congress, Global Legal Research Directorate, Nov. 2021.

[282] Matt Ahlborg. *Paxful is the Most Important Bitcoin Company You Aren't Paying Attention to.* https://medium.com/dlabvc/paxful-is-the-most-important-bitcoin-company-you-arent-paying-attention-to-4e699db0c5ca. Accessed Mar. 28th, 2023. 2019.

[283] Paxful. *What Countries Require ID Verification?* https://web.archive.org/web/20230610111357/https://support.paxful.com/hc/en-us/articles/360021175133-What-Countries-Require-ID-Verification-. Accessed Mar. 28th, 2023, but page deleted. 2023.

[284] Paul Resnick et al. "Reputation systems". In: *Communications of the ACM* 43.12 (2000), pp. 45–48.

[285] Yuewen Liu, Juan Feng, and Kwok Kee Wei. "Negative price premium effect in online market–The impact of competition and buyer informativeness on the pricing strategies of sellers with different reputation levels". In: *Decision Support Systems* 54.1 (2012), pp. 681–690.

[286] Audun Jøsang and Jennifer Golbeck. "Challenges for robust trust and reputation systems". In: *Proceedings of the 5th International Workshop on Security and Trust Management*. Vol. 5. 9. 2009.

[287] Paxful. *Why is My Paxful Account Frozen or Restricted?* Accessed Mar. 28th, 2023. 2023. URL: https://paxful.zendesk.com/hc/en-us/articles/360014551640-Why-is-My-Paxful-Account-Frozen-or-Restricted.

[288] Paxful. *List of Banned Countries, OFAC (Through Wayback Machine)*. https://web.archive.org/web/20230324163159/https://paxful.zendesk.com/hc/en-us/articles/360013470374-List-of-Banned-Countries-OFAC. Accessed Mar. 28th, 2023, but page deleted. 2022.

[289] LocalCoinSwap. *Terms of Service and Use - LocalCoinSwap*. https://localcoinswap.com/terms-of-service. Accessed Jul. 13th, 2023. 2023.

[290] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[291] Teddy Fusaro et al. *Meeting with Bitwise Asset Management, Inc., NYSE Arca, Inc., and Vedder Price P.C.* Official Memorandum. 2019. URL: https://www.sec.gov/comments/sr-nysearca-2019-01/srnysearca201901-5164833-183434.pdf.

[292] Clayton Allen Davis et al. "Botornot: A system to evaluate social bots". In: *Proceedings of the 25th international conference companion on world wide web*. 2016, pp. 273–274.

[293] Hung Tran et al. "Spam detection in online classified advertisements". In: *Proceedings of the Joint WICOW/AIRWeb Workshop on Web Quality*. 2011, pp. 35–41.

[294] Youngsam Park et al. "Scambaiter: Understanding targeted nigerian scams on craigslist". In: NDSS '14 1 (2014), p. 2. DOI: 10.14722/ndss.2014.23284.

[295] *Make Trading More Rewarding with the Paxful Rewards Program*. https://paxful.com/university/new-rewards-program/. Accessed Jun. 27th, 2023. 2021.

[296] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[297] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems*. NeurIPS '17 30 (2017).

[298] Hamad Alsaleh and Lina Zhou. "A Heuristic Method for Identifying Scam Ads on Craigslist". In: *2018 European Intelligence and Security Informatics Conference (EISIC)*. Ieee. 2018, pp. 69–72.

[299] Madelyne Xiao et al. "Account Verification on Social Media: User Perceptions and Paid Enrollment". In: *arXiv preprint arXiv:2304.14939* (2023).

[300] Marina Andreianova et al. "Bitcoin Usage: Study on Bitcoin Usage Around the World 2020". In: *The Journal of FinTech* 1.02 (2021), p. 2150005.

[301] Vaibhav Garg and Shirin Nilizadeh. "Craigslist scams and community composition: Investigating online fraud victimization". In: *2013 IEEE Security and Privacy Workshops*. SPW '13. 2013, pp. 123–126.

[302] Chris Hays et al. "Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection". In: *arXiv preprint arXiv:2301.07015* (2023).

[303] Kurt Thomas et al. "Suspended accounts in retrospect: an analysis of twitter spam". In: *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*. IMC '11. Association for Computing Machinery, 2011, pp. 243–258.

[304] Abdullah Almaatouq et al. "Twitter: who gets caught? Observed trends in social micro-blogging spam". In: *Proceedings of the 2014 ACM conference on Web science*. 2014, pp. 33–41.

[305] Chao Yang et al. "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter". In: *Proceedings of the 21st international conference on World Wide Web*. WWW '12. 2012, pp. 71–80.

[306] Qiang Cao et al. "Uncovering large groups of active malicious accounts in online social networks". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 2014, pp. 477–488.

[307] Huyen Le et al. "A postmortem of suspended Twitter accounts in the 2016 US presidential election". In: *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '19. IEEE. 2019, pp. 258–265.

[308] Joshua Sunshine et al. "Crying wolf: an empirical study of SSL warning effectiveness". In: *Proceedings of the 2009 USENIX Security Symposium*.

[309] Robert W. Reeder et al. "An Experience Sampling Study of User Reactions to Browser Warnings in the Field". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

[310] Adrienne Porter Felt et al. "Android permissions: user attention, comprehension, and behavior". In: *Proceedings of the 2012 Symposium on Usable Privacy and Security (SOUPS '12)*.

[311] Tim Miller. *Explanation in Artificial Intelligence: Insights from the Social Sciences*. 2018. arXiv: 1706.07269 [cs.AI]. URL: https://arxiv.org/abs/1706.07269.

[312] Binance. *How to Spot and Avoid P2P Scams and Fraud*. Accessed Jun. 27th, 2023. 2021. URL: https://www.binance.com/en/blog/p2p/how-to-spot-and-avoid-p2p-scams-and-fraud-6416111153825153913.

[313] R. Cattell. "The Scree Test For The Number Of Factors". In: *Multi. Behavior. Res.* 1.2 (1966), pp. 245–276.

[314] J. Horn. "A rationale and test for the number of factors in factor analysis". In: *Psychometrika* 30 (2 1965), pp. 179–185.

[315] J. Stevens. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

[316] W. Sutherland. *Ecological census techniques: a handbook*. Cambridge University Press, 2006.

[317] C. Schwarz and A. Arnason. "Jolly-Seber Models in MARK". In: *MARK: A Gentle Introduction*. 8th. 2009.

[318] Reddit. https://www.reddit.com/r/paxful/comments/jjq1qd/nigerians_using_vpn/. Accessed Jun. 27th, 2023. 2020.

[319] Reddit. https://www.reddit.com/r/paxful/comments/xh22gx/what_does_user_not_active_mean/. Accessed Mar. 28th, 2023. 2021.