

Combating online consumer fraud and counterfeits: A data science perspective

Felix David Soldner

A dissertation submitted in partial fulfilment of the requirements for the
degree of

Doctor of Philosophy

Security and Crime Science

University College London

2023

Student declaration

I, Felix David Soldner, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Date: 18.10.2023

Abstract

Online fraud is a growing problem that impacts many individuals, resulting in billions of dollars' worth of damages. Although various online fraud types exist, they are all easily scalable through online (shopping) platforms, thus reaching many individuals with relatively little effort. The sheer number of fraud cases authorities face is impossible to resolve using traditional investigatory practices, which often require intensive manual work. Data science offers some solutions for the problems presented by online fraud through automation and making manual labour more efficient. This thesis explores various methods of how data science can help combat online consumer fraud and counterfeits.

Chapters 3 and 4 examine the challenges of automated approaches to combating online fraud. Chapter 3 examines the feasibility of creating supervised machine learning training data by asking experts to annotate product listings based on suspiciousness. Annotators show low agreement on what constitutes a suspicious listing, revealing the importance of precise definitions of labels, clear instructions, and thorough records of the annotators' decision-making processes during labelling. In chapter 4, the impact that confounds in training data have on prediction performances (e.g., detecting fake reviews) is evaluated by examining the design choices used to create datasets. The results show that by mixing experimentally created and found (e.g., collected online) data, prediction performances can be artificially boosted, leading to incorrect conclusions about the predictive features.

Chapters 5 and 6 examine the promise of automated methods for combating online fraud by examining data from anonymity networks and cryptomarkets, highly anonymized sections of the internet, often used to trade illicit goods. Counterfeits are openly offered on anonymity networks, and their information could help us to improve our understanding of the counterfeit economy. We analyse a large-scale dataset (2014-2015) to determine their prevalence, types, origins, and (sales) values across multiple markets. Comparing the estimations to other measures by authorities (e.g., border seizures), we found similarities in the type and origins of counterfeited products and that the number of types varied across measures. Finally, we utilise information about counterfeits on anonymity networks to search for the same products on the surface web by matching and ranking them based on image and text similarities. We examined highly similar matches and found that the number of identical products across platforms, such as shoes, smartphones, and watches, would warrant further investigation into whether they are counterfeits.

The thesis closes with a discussion of the results before reviewing the limitations and possible future avenues for research on addressing online fraud. The availability of high-quality data, including ground truth data, is a recurring issue in fraud research, which could be addressed through better data documentation practices and increased data sharing. Future studies should aim to increase the temporal coverage of anonymity network data to allow for a better examination of trends in the counterfeit economy. Practical implications of utilising data science approaches are discussed, highlighting the importance of conveying the limitations and implications of applying data science methods to practitioners.

Impact statement

The thesis includes proof of concept approaches and practical perspectives for practitioners interested in combating online (consumer) fraud and counterfeits. The work reported here examined how data science might facilitate a better understanding of frauds through large-scale analyses, how to apply an automated approach to support practitioners (e.g., speeding up manual tasks), and which pitfalls researchers and practitioners should consider when using supervised machine learning methods. Specifically, parts of the work were conducted in exchange with authorities, such as Trading Standards UK and the Intellectual Property Office UK. Through the collaboration, insights into the challenges of applying data science methods in practice were gathered and are discussed in Chapter 3, which researchers and practitioners can draw on for future work. Results and insights on possible data science applications or integrations within police work were also discussed with Scotland Police. Chapter 6 of this thesis provides a proof-of-concept study exploring the utility of automatically searching for potential counterfeits on the surface web. Thus, providing a basis for future approaches of partial automation of manual work by authorities.

The datasets created for this thesis have been made publicly available or available upon request so that future researchers studying fraud and counterfeits on cryptomarkets can build on the current work without having to construct the datasets themselves, which is a time-consuming and labour-intensive process. By laying out a plan for possible future work, the thesis provides insights into how data science could be further utilised to study fraud-related phenomena and possibly apply it in practical settings. Finally, the thesis provides perspectives on how organisations tackling fraud-related phenomena can be assisted scientifically to increase their impact against fraud.

Parts of the work of this PhD thesis have been presented at conferences, such as the [International Conference on Computational Social Science](#) (2021), the [Annual Cybercrime Conference](#) (2022) at the Cambridge Cybercrime Centre, and the [ODISSEI Conference for Social Science in the Netherlands](#) (2022) in Utrecht. Guest lectures about counterfeits on cryptomarkets derived directly from the empirical insights gained through this PhD project were also delivered at the University of Amsterdam (2021) and about data confounds in fake review detection at University College London (2021).

Chapters 2, 4, and 5 are based on the following publications:

- Soldner, F., Kleinberg, B., & Johnson, S. (2022). Trends in online consumer fraud: A data science perspective. In *A Fresh Look at Fraud* (pp. 167-191). Routledge.
- Soldner, F., Kleinberg, B., & Johnson, S. D. (2022). Confounds and overestimations in fake review detection: Experimentally controlling for product-ownership and data-origin. *Plos one*, 17(12), e0277869. <https://doi.org/10.1371/journal.pone.0277869>
- Soldner, F., Kleinberg, B. & Johnson, S.D. Counterfeits on dark markets: a measurement between Jan-2014 and Sep-2015. *Crime Sci* 12, 18 (2023). <https://doi.org/10.1186/s40163-023-00195-2>

Funding declaration

This research was fully funded by the Dawes Centre for Future Crime as a Student Stipend.

Research Paper Declarations

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- have been uploaded to a preprint server
- are in submission to a peer-reviewed publication
- have been published in a peer-reviewed publication, e.g. journal, textbook.

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. For a research manuscript that has already been published (if not yet published, please skip to section 2)

a) **What is the title of the manuscript?**

Confounds and overestimations in fake review detection: Experimentally controlling for product-ownership and data-origin

b) **Please include a link to or doi for the work**

<https://doi.org/10.1371/journal.pone.0277869>

c) **Where was the work published?**

NA

d) **Who published the work?** (e.g. OUP)

PLOS ONE

e) **When was the work published?**

December 7, 2022

f) **List the manuscript's authors in the order they appear on the publication**

Felix Soldner, Bennett Kleinberg, Shane D. Johnson

g) **Was the work peer reviewed?**

Yes

h) **Have you retained the copyright?**

Yes

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

Yes, ArXiv: <https://doi.org/10.48550/arXiv.2110.15130>

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

a) **What is the current title of the manuscript?**

NA

b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

NA

c) **Where is the work intended to be published?** (e.g. journal names)

NA

d) **List the manuscript's authors in the intended authorship order**

NA

e) **Stage of publication** (e.g. in submission)

NA

3. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4)

Felix Soldner was responsible for data collection, data analyses, drafting, writing, and revising. Bennett Kleinberg and Shane Johnson gave feedback and made language corrections in the individual peer-review rounds.

4. In which chapter(s) of your thesis can this material be found?

Confounds and Overestimations in Fake Review Detection: Experimentally Controlling for Product-Ownership and Data-Origin

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Date:

07.03.2023

Supervisor/ Senior Author (where appropriate)

Date

15/03/2023

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- have been uploaded to a preprint server
- are in submission to a peer-reviewed publication
- have been published in a peer-reviewed publication, e.g. journal, textbook.

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. For a research manuscript that has already been published (if not yet published, please skip to section 2)

j) What is the title of the manuscript?

"Trends in online consumer fraud: A data science perspective" within the Book "A Fresh Look at Fraud"

k) Please include a link to or doi for the work

<https://doi.org/10.4324/9781003017189-9>

l) Where was the work published?

London

m) Who published the work? (e.g. OUP)

Routledge

n) When was the work published?

2022

o) List the manuscript's authors in the order they appear on the publication

Felix Soldner, Bennett Kleinberg, Shane Johnson

p) Was the work peer reviewed?

Yes

q) Have you retained the copyright?

Yes

r) Was an earlier form of the manuscript uploaded to a preprint server? (e.g. medRxiv). If

'Yes', please give a link or doi)

No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

f) **What is the current title of the manuscript?**

NA

g) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

NA

h) **Where is the work intended to be published?** (e.g. journal names)

NA

i) **List the manuscript's authors in the intended authorship order**

NA

j) **Stage of publication** (e.g. in submission)

NA

3. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4)

Felix Soldner was responsible for drafting, writing, and revising the book chapter. Bennett Kleinberg and Shane Johnson gave feedback and made language corrections in the individual peer-review rounds.

4. In which chapter(s) of your thesis can this material be found?

In the chapter "Background and Literature review"

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Date:

07.03.2023

Supervisor/ Senior Author (where appropriate)

Date

15/03/2023

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- have been uploaded to a preprint server
- are in submission to a peer-reviewed publication
- have been published in a peer-reviewed publication, e.g. journal, textbook.

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. For a research manuscript that has already been published (if not yet published, please skip to section 2)

s) **What is the title of the manuscript?**

NA

t) **Please include a link to or doi for the work**

NA

u) **Where was the work published?**

NA

v) **Who published the work?** (e.g. OUP)

NA

w) **When was the work published?**

NA

x) **List the manuscript's authors in the order they appear on the publication**

NA

y) **Was the work peer reviewed?**

NA

z) **Have you retained the copyright?**

NA

aa) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

NA

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

k) **What is the current title of the manuscript?**

Counterfeits on Cryptomarkets: A measurement between Jan-2014 and Sep-2015

l) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

Yes, ArXiv: <https://doi.org/10.48550/arXiv.2212.02945>

m) **Where is the work intended to be published?** (e.g. journal names)

Journal: Crime Science

n) **List the manuscript's authors in the intended authorship order**

Felix Soldner, Bennett Kleinberg, Shane D. Johnson

o) **Stage of publication** (e.g. in submission)

In submission

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

Felix Soldner was responsible for data collection, data analyses, drafting, writing, and revising. Bennett Kleinberg and Shane Johnson gave feedback and made language corrections in the individual peer-review rounds.

4. **In which chapter(s) of your thesis can this material be found?**

Counterfeits on Cryptomarkets: A measurement between Jan-2014 and Sep-2015

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Date:

07.03.2023

Supervisor/ Senior Author (where appropriate)

Date

15/03/2023

Acknowledgments

Working on my thesis has been an incredible journey, and I would not have made it through without support from those around me. The entire PhD was an insightful learning experience that I would never have wanted to miss.

I was continuously accompanied by my supervisors, Bennett, and Shane, who helped me along the way with their advice. Bennett always made time for me and consistently supported me with excellent insights. Bennett, you provided me with great opportunities and showed me how to find new perspectives, for which I am very grateful. Shane was a constant source of new ideas and always showed me how other research might be related to mine. Shane, you helped me to connect with other researchers and practitioners, which taught me how to navigate relationships within and outside of academia, which I greatly value. Both of you have given me the right encouragement at exactly the right time. I remember one meeting in London when I was stuck in my thinking and talking with you gave me the necessary mental shift to continue my work.

Having others around me in similar situations, like Max, Josh, Arianna, or Daniel, was fantastic since I could share all my (typical) PhD frustration with them, which made the journey more enjoyable. But the work was not always stressful, and sharing ideas and excitement about possible future studies was revitalising. Thank you all for a wonderful time. I also want to thank Isabelle, a great companion during the London lockdowns, who was very supportive in keeping an adequate work-life balance. Thanks to you! I have also adopted your no-work policy on weekends, which I still (mostly) uphold—it has likely saved me from many headaches.

Fabian, my great friend, and colleague, I was very happy to have you with me in London. Bouldering with you and sharing our ideas always lifted me up. Thank you for contributing to this work and adding valuable image analyses to my text-analyses-heavy work.

Lauren, I was very fortunate to have you with me during this time (at first, mainly through the internet, but now on the same continent). I am extremely happy we can spend so much time together and that I have you as my life companion. You knew how to support me when I needed it, especially in the last stretches. The days are just better with you.

Florian and Lina, my parents, I am incredibly grateful for having you both as my longest life-guides so far. You made that all possible by showing me how to see the world through curious eyes and understand struggles as opportunities. You taught me how to enjoy learning and how to challenge myself to do my best. These are the skills I needed in this journey. I am very grateful to have you.

Table of Contents

Student declaration	2
Abstract	3
Impact statement	4
Research Paper Declarations.....	6
Acknowledgments	12
Table of Contents.....	13
Table of Tables.....	16
Table of Figures	17
Chapter 1: Introduction	18
1.1 Theoretical perspectives on online consumer fraud.....	19
1.2 Thesis aims and structure	20
Chapter 2: Background and literature review	23
2.1 Common online consumer fraud types on the surface web	23
2.2 Current strategies to prevent and resolve online consumer fraud	26
2.3 The role of data science.....	30
2.4 Possible future data science applications	34
2.5 Anonymity networks	36
2.6 Utilizing cryptomarkets to identify consumer fraud on the surface web	42
2.7 Conclusion	43
Chapter 3: Challenges in annotating training data for supervised machine learning models	44
3.1 Introduction	44
3.2 Method	47
3.3 Results	51
3.4 Discussion	54
3.5 Conclusion	57
Chapter 4: Confounds and Overestimations in Fake Review Detection: Experimentally Controlling for Product-Ownership and Data-Origin	58
4.1 Introduction	58
4.2 Data collection	61
4.3 Supervised learning analysis.....	65
4.4 Results	65
4.5 Discussion	70
4.6 Conclusion	74
Chapter 5: Counterfeits on cryptomarkets:	75

5.1 Introduction	75
5.2 Data	78
5.3 Results	84
5.4 Discussion	94
5.5 Conclusion	99
Chapter 6: From anonymity networks to the surface web: Scouting eBay for counterfeits	101
6.1 Introduction	101
6.2 Data	103
6.3 Similarity metrics.....	107
6.4 Image similarities	108
6.5 Determining similarity between product pairs	111
6.6 Examining similarity scores and product matches	114
6.7 Discussion	122
6.8 Conclusion	128
Chapter 7: General Discussion.....	129
7.1 Summary of main findings.....	129
7.2 How studies relate to each other.....	132
7.3 Generalization of findings.....	133
7.4 Limitations and Outlook	134
7.5 Theoretical perspectives.....	137
7.6 Practical Implications	144
Conclusion	147
References.....	148
Appendix A: Background and literature review	169
A1 Table. 25 techniques of situational crime prevention from Cornish & Clarke (2003)	169
Appendix B: Challenges in annotating training data for supervised machine learning models (Chapter 3).....	170
B1 Full list of scraped eBay categories	170
Appendix C: Confounds and Overestimations in Fake Review Detection: Controlling for Product-Ownership and Data-Origin (Chapter 4)	172
C1 Table. Amazon review replacements	172
C2 Table. All features used in the classification experiments	172
C3 Table. Extra Trees classifier settings	174
C4 Table. Other tested classifiers	174
C5 Table. All classification performance metrics across all experiments.....	174

Appendix D: Counterfeits on Cryptomarkets: A measurement between Jan-2014 and Sep-2015 (Chapter 5)	175
D1 Table. List of considered markets	175
D2 List. Categories included for keyword searches:	176
D3 List. Synonyms used for keyword search:	176
D4 List. Synonyms of authentic:	176
D5 List. Keywords used to exclude listings:	176
D6 Table. Full lists of percentage counterfeits by OECD/EUIPO.....	177
D7 Table. Full list of percentage counterfeits by IPO.....	177
D8 Figure. Product price differences for 10 products in each category between Cryptomarkets and the surface web.	178
Appendix E: From the anonymity network to the surface web: Scouting eBay for Counterfeits (Chapter 6)	179
E1 Table. Descriptive statistics of word occurrences	179
E2 Figure. Example of rescaled cryptomarket images	179
E3 Figure. Siamese network architecture	180
E4 Detailed description of sampling annotation data.....	180
E5 Table. Regression models settings	181
E6 Figure. Detailed model comparison	181
E7 Table. Full product descriptions (eBay scrape period 1)	182
E8 Table. Full product descriptions (eBay scrape period 2)	185

Table of Tables

Table 2.1. Summary of How Different Sectors Address Online Consumer Fraud	29
Table 3.1 Distribution of eBay listings.....	48
Table 3.2. Distribution of Annotations for groups and individuals.	51
Table 3.3. Krippendorff’s alpha within and between annotator groups.	52
Table 3.4. Recommendations for future annotations.....	57
Table 4.1. Average sentiment scores across reviews and their ratings.....	64
Table 4.2. Overview of all filtered reviews	65
Table 4.3. Explanation of all feature names	68
Table 4.4. Top 5 feature differences by BF_{10} for all pure classifications.....	68
Table 4.5. Top 5 feature differences by BF_{10} for all confounded classifications	69
Table 5.1. Markets and their data timeframe in this study.	79
Table 5.2. Annotated categories within counterfeits.....	81
Table 5.3. The performance scores (weighted average) across 10-folds.	82
Table 5.4. Number of found and replaced holding and placeholder prices	84
Table 5.5. Percentage of counterfeit categories.....	87
Table 5.6. Percentage of counterfeit shipping origins by country and product category.	89
Table 5.7. Summary counterfeit prices and volumes for each product category in USD.....	90
Table 5.8. Estimated sales volume (USD) for each category based on the number of feedbacks	91
Table 5.9. Mean [SD] of 10 sample products for each category on surface web markets	93
Table 6.1. Counterfeits for each market across categories	104
Table 6.2. Product price distributions across markets and categories.....	106
Table 6.3. Similarity rating agreements between listings for each product group and overall	112
Table 6.4. Average model performances (10-fold)	113
Table 6.5. SVR coefficients	113
Table 6.6. Cohen’s d effect sizes for product category comparisons.....	115
Table 6.7. Cohen’s d effect sizes for product origins comparisons	115
Table 6.8. Product categories within the top 50 ranked and 50 randomly selected product pairs	116
Table 6.9. Similarity score distribution within both product samples of the first scrape period.....	117
Table 6.10. Examples of titles and prices of matching product pairs.....	118
Table 6.11. Product categories within the top 50 ranked and 50 randomly selected product pairs	119
Table 6.12. Similarity score distribution within both product samples of the second scrape period.	119
Table 6.13. Examples of titles and prices of matching product pairs.....	120
Table 6.14. Examples of eBay product titles.....	122
Table 7.1. Theoretical distribution of predicted counterfeits.....	145

Table of Figures

Figure 2.1. The World Wide Web subdivided into surface, deep, and anonymous web.	37
Figure 2.2. Example screenshot of a counterfeit listing	39
Figure 3.1. Scraper framework for eBay listings	48
Figure 3.2. Example of one eBay listing participants were presented	50
Figure 3.3. Percentage distribution of suspiciousness ratings of experts and non-experts	51
Figure 3.4. Percentage distribution of confidence ratings of experts and non-experts.....	52
Figure 3.5. Distribution of reasoning topics	53
Figure 3.6. Experts (left) and non-experts (right) reasoning topics, split by suspiciousness	54
Figure 4.1. Collecting procedure of reviews from Prolific participants	63
Figure 4.2. Classification accuracies of analyses 1, 2, and 3	66
Figure 4.3. Classification accuracies of analyses 1, 4, 5, and 6	67
Figure 4.4. Topic prevalence and for which review type the topic appears more often.....	70
Figure 5.1. Normalized confusion matrix for true and predicted categories of counterfeits.....	82
Figure 5.2. Monthly volume of products offered across markets.....	85
Figure 5.3. Monthly volume of counterfeits offered across markets.....	86
Figure 5.4. Percentage of shipping origins for all products and counterfeits.....	88
Figure 5.5. Monthly stacked sales volume (based on feedback) across categories.....	92
Figure 5.6. Mean (SE) price differences for each product category.....	94
Figure 6.1 Screenshot of a counterfeit listing on the CM Darkode.....	104
Figure 6.2 Percentage of product origins in percentage.	106
Figure 6.3. Distribution of annotated similarity ratings	112
Figure 6.4. Examples of product images from the first eBay scrape from cryptomarkets (CM).....	118
Figure 6.5. Examples of product images from the second eBay scrape from cryptomarkets (CM)	120
Figure 6.6. Highly similar eBay products resembling counterfeits on cryptomarkets.....	121
Figure 7.1. The crime triangle represents how when the various RAA components converge	138

Chapter 1: Introduction

Over the last few decades, traditional crimes, such as residential burglary, have been declining (Bossler & Berenblum, 2019; Caneppele & Aebi, 2019; Dijk et al., 2012; Tcherni et al., 2016; Walters & Langton, 2013) while others – including online consumer fraud, which is the subject of this thesis – have been increasing. Online consumer fraud can take many forms. It is, in essence, the defrauding of individuals in online environments, leading to financial loss (R. Anderson et al., 2013). Such fraud can include unknowingly purchasing counterfeit items, agreeing to a subscription service, or paying for products and services that are not delivered. Consumers can also be tricked into giving away their login credentials and spending money on non-existent services or products through phishing or fake websites (K. Anderson, 2019, 2022). While such fraud schemes differ in how they are executed, they share a common trait: consumers must interact with an online platform. This interaction between fraudsters and consumers in online environments (e.g., shopping platforms) makes the fraud easily scalable, allowing fraudsters to reach many individuals very quickly with little effort. The scale of online fraud exacerbates the negative impact of the fraud schemes on society.

Online consumer fraud and cybercrime as a whole generate financial and psychological costs to society, including direct monetary losses, criminal revenues, and opportunity costs (Ablon et al., 2014; R. Anderson et al., 2013; van Wegberg et al., 2018). For example, a report by the FBI (2018), which examined complaints and messages from Internet crime victims in the US, found a steady increase for all types of cybercrime, with an estimated total monetary loss of \$2.71 billion in 2018. Notably, it is estimated that only 15% of online crimes are reported (Office for National Statistics, 2020), indicating that losses are probably far more extensive. Additionally, estimates suggest that consumer products and services account for 42.6% of all financial fraud incidents, with the Internet being the dominant solicitation method (30%) in the US (DeLiema et al., 2017). Although accurate estimates of online consumer fraud are difficult to compute due to the lack and granularity of data, estimates such as those above give a sense of the scale of the problem.

With the ever-increasing scale of fraud schemes comes growing pressure on authorities, who are faced with tremendous amounts of fraud cases, and find them impossible to address with traditional methods. The time-intensive manual labour of traditional fraud investigation approaches in particular exacerbates the workload of fraud cases. Here, data science approaches can be helpful. As fraudsters scale up their schemes, we can employ data science methods to scale up analyses to better understand the fraud landscape (e.g., the prevalence and scope of fraud types and affected individuals) and possibly employ large-scale prevention or detection methods through automation. Data science methods, such as machine learning, can uncover previously unknown patterns in online markets (e.g., fraud-related seller behaviour) and reveal information about fraud schemes that humans would not recognize. Thus, various data science approaches could help us understand *how* and *where* fraud occurs and reduce the current workload of authorities.

1.1 Theoretical perspectives on online consumer fraud

Crime theories such as the Routine Activity Approach (RAA) and the rational choice perspective provide a lens for thinking about online fraud and cybercrime and why they might take place (L. E. Cohen & Felson, 1979; Cornish & Clarke, 1987). Both perspectives were initially developed to explain crime in physical spaces but have been increasingly used to understand crime committed in cyberspace (Bossler & Berenblum, 2019; Hutchings & Hayes, 2009; Kigerl, 2021; Ngo et al., 2020; Reyns & Randa, 2020; Simpson et al., 2014). The RAA assumes that a crime will be more likely to take place when a motivated offender and a suitable target (e.g., valuables) converge in space and time absent a capable guardian (e.g., watchful individual, physical hurdle) (L. E. Cohen & Felson, 1979). Thus, influencing the offenders' motivation, the suitability of targets or the guardianship available can all reduce the likelihood that a crime will occur. Put differently, even if offender motivation goes unchanged, crime can be reduced by increasing guardianship or by making targets less suitable. A related theory is the rational choice perspective, which takes the view that perpetrators make rational decisions by weighing the possible risks, effort and benefits of their actions (Cornish & Clarke, 1987). The rational choice perspective is central to situational crime prevention approaches (Clarke, 1995; Freilich & Newman, 2018), the idea of which is that changing situational circumstances to alter offender perceptions of risks and rewards, can reduce crime. Crime reduction is often achieved by increasing the risks and effort associated with offending, reducing the rewards and excuses for committing a crime (Clarke, 1995; Freilich & Newman, 2018), or both. Interventions intended to remove or reduce excuses are often aimed at crimes that are perceived by many as mundane (e.g., speeding, littering) and are often morally excused by the individuals (e.g., "everyone does it") (Cornish & Clarke, 2003). Such common excuses can be challenged by setting rules or increasing the awareness of the prohibitions (e.g., using roadside speed display).

With the advent of the Internet, the opportunities for motivated offenders to interact with (suitable) targets have changed, often due to the absence of capable or active guardianship. Thus, new forms of crime committed online have emerged, including consumer fraud (Rusch, 1991). With those changes, preventing online consumer fraud becomes difficult due to the nature of the online environment, which enables fraudsters to operate from anywhere with relative ease and anonymity, reducing the perceived effort and risk involved. The online environment also allows fraudsters to reach individuals at scale, meaning that they can interact with many (suitable) targets easily, increasing the fraudsters' potential rewards. This scalability also makes traditional ways of manually detecting and preventing such fraud inefficient and complicates the actions involved (e.g., crimes may be committed in a different jurisdiction to that in which the offender is located) for law enforcement agencies to bring fraudsters to justice (Herley, 2010). While data science could be used to alleviate some of the manual workload authorities face, detection approaches employed by online (shopping) platforms could also increase guardianship capabilities and the risk of detection experienced by fraudsters. Thus, this thesis not only explores how data science approaches could facilitate

a better understanding of online consumer fraud but also how the efforts and risks to offenders could be increased to facilitate the prevention of online consumer fraud.

1.2 Thesis aims and structure

This thesis aims to highlight the different perspectives, challenges, and promises of automated approaches that might be used to combat online consumer fraud. Various data science methods are explored for distinct fraud types, which cover only a part of the entire online fraud landscape. Thus, the chapters of this thesis differ considerably in what kind of fraud they are concerned with, but all share the common goal of investigating the usability of data science methods in combating online consumer fraud.

Chapter 2 describes online consumer fraud, including estimations of financial losses and describes common types. The chapter reviews the current literature about governmental and commercial perspectives on how they deal with online fraud and how data science is and could be utilized to combat it. The chapter continues by introducing anonymity networks on the internet that are used for the illicit trading of (for example) drugs but also for counterfeits (e.g., defrauding guides, falsified documents, fake apparel). Current literature about anonymity networks is reviewed, including data collection approaches, the economy (product and service offers), as well as vendor and user behaviours. Lastly, the chapter describes how information we can obtain from such anonymity networks could be utilized to combat consumer fraud on surface web (shopping) platforms (e.g., eBay, Amazon).

Chapters 3 and 4 examine current challenges and possible pitfalls when using automated methods, such as supervised machine learning methods, in combating online fraud. In particular, Chapter 3 explores the feasibility of using experts to create training data for a supervised machine learning model to detect suspicious, potentially fraudulent eBay listings. By asking experts to label eBay listings based on their suspiciousness of being fraudulent, the chapter aims to answer which annotation practices are important during the dataset creation process. The chapter examines the difficulties associated with creating a suitable training dataset for supervised models, and then describes the different annotation strategies employed by experts and non-experts. The chapter establishes the importance of clear and agreed-upon labelling practices between annotators when ground truth labels are unavailable. Since labelled training data are essential for supervised learning models to learn to associate data features with data classes, a misguided labelling procedure jeopardizes model performance. The chapter exemplifies those problems and concludes with recommendations for addressing these challenges.

Chapter 4 examines automated approaches to detecting fake reviews, which are often encountered on consumer platforms, and which falsely promote products or services or demote competitors' offers. Current supervised learning approaches vary considerably in detection performance, and Chapter 4 investigates how training data created experimentally or sampled from online platforms can affect model performance. More precisely, when

datasets that originate from various sources (e.g., created experimentally or sampled from online platforms) are combined, possible confounds can be introduced that impact model performance, even when ground truth labels are partially available. The chapter illustrates how seemingly similar text data (here: customer reviews of smartphone purchases) can contain platform- and context-related features that confound the outcome labels (here: fake vs genuine) to the degree that they significantly alter the model performance. The chapter closes with suggestions for implementing more robust controls when creating (training) datasets. Chapters 3 and 4 both highlight the importance and difficulties of creating reliable training data for supervised models.

Chapters 5 and 6 examine the potential and possible future applications of automated approaches to combat online fraud. The two chapters cover what we can learn about counterfeits from anonymity networks and how we can use information about them to search for the same products on the surface web. Chapter 5 analyses data from cryptomarkets from January 2014 to September 2015 to examine how prevalent counterfeits are on cryptomarkets and whether we can gather new insights into the counterfeit economy that could be useful for practitioners. This chapter shows how we can learn about counterfeit types on cryptomarkets, their origins, and sales through a computational analysis of digital behavioural data. By using various data science techniques, the chapter also demonstrates how we can increase the granularity of data analyses to improve our understanding of the counterfeit economy. By comparing the generated prevalence estimates to other traditional measures (e.g., border seizures, complaint statistics) created by European and UK authorities, the chapter highlights the similarities and discrepancies between different measurements and how these affect insights concerning the counterfeit economy.

Following the topic of counterfeits on anonymity networks, Chapter 6 investigates how data science methods could be used to search for potential counterfeits on eBay using automated approaches. To do this, information from counterfeit listings on current cryptomarkets (2021) was collected and used to search for the same products on eBay. By automatically generating text and image similarity scores for cryptomarket and eBay products, extensive product comparisons could be made, and highly similar products across platforms were identified. We made product comparisons at two points in time to assess similarity changes for various product types over time. The chapter shows how potential connections between product offers on cryptomarkets and the surface web could be investigated in the future, possibly informing us about potential counterfeit-affected product types.

Viewing the individual chapters together, chapters 3 and 4 show which hurdles are present when utilizing supervised machine learning methods to detect online consumer fraud, while chapters 5 and 6 show how data science methods can also be used to increase our understanding of fraud and how the current manual work typically undertaken by law enforcement could be supported through partial automation.

Lastly, Chapter 7 discusses the results from all chapters of this PhD thesis and what they mean for the bigger picture of online consumer fraud. The chapter proceeds with a discussion of the limitations of the work, possible future research and practical implications for researchers and practitioners.

Chapter 2: Background and literature review

This chapter is based on the following publication:

- Soldner, F., Kleinberg, B., & Johnson, S. (2022). Trends in online consumer fraud: A data science perspective. In *A Fresh Look at Fraud* (pp. 167-191). Routledge.

This chapter starts by examining what online consumer fraud is and describes commonly perpetrated fraud schemes. Subsequently, the chapter reviews some of the current approaches (non-)governmental and commercial institutions take to detect and prevent online consumer fraud, followed by an overview of how methods from data science could be applied to support the detection and prevention of online consumer fraud as well as possible future research directions. The chapter then introduces anonymity networks and cryptomarkets, on which the communication between users is highly anonymized, allowing individuals to offer and purchase (illicit) goods. While drugs are the predominant products on such markets, a small fraction (3%) of the items listed are openly sold counterfeits and fraud-related products. Thus, the chapter describes how information from such networks and markets might be relevant for understanding and combating fraud on the surface web.

2.1 Common online consumer fraud types on the surface web

For this section, we focus on online consumer fraud committed on the surface web (e.g., on eBay, Amazon, etc.) and define fraud in this context as transactions for which the consumer is deceived in some way and is unaware that fraudulent activity is taking place. We limit our attention to fraud, such as unauthorized billing, the non-delivery of goods or services ordered from legitimate or fraudulent websites, or fraud that requires the theft of consumers' credentials (e.g., via phishing websites) (K. Anderson, 2019). These types of fraud can often be deployed at scale (K. Anderson, 2022) and require little sustained attention from the fraudster. These types of offences differ from other types of fraud, such as romance or advanced fee scams, which require sustained effort from the fraudster and typically include activity that cannot be or would be more difficult to automate. We focus our attention on the former (unauthorized billing, non-deliveries, etc.) because we believe that these types of offences – which generally involve little to no interaction with a victim – will be easier to address using approaches from data science by increasing active guardianship, or increasing the effort of offending, through automated detection mechanisms. Data science approaches often rely on automated detection systems, which are difficult to deploy if a personal conversation between individuals is taking place (with the exceptions of e-mail spam filters). Furthermore, to successfully implement automated detection methods, the required data needs to be accessible and usable, as in open or public domains, such as online marketplaces. Thus, we will discuss common online consumer fraud types, which are more suitable to be addressed by data science methods.

2.1.1 Fraudulent billings

One of the most common fraud schemes involves billings for products or services that the customer did not agree to (K. Anderson, 2019). They can occur on websites where a customer believes they are making a one-time order and payment but are in fact signing up for a subscription of some kind. Similarly, free trials for which a subscription is about to terminate can be extended without the approval of – but at a cost to – the customer. It has been estimated that unapproved billings through fraudulent websites accounted for a monetary loss of \$48 million in the USA alone in 2018 (FBI, 2018). However, such estimates are based solely on complaints made to the FBI. Given that most (online) crime goes unreported (up to 85%) (Office for National Statistics, 2020), these estimates provide only a partial picture. Different forms of such web pages exist but include those that masquerade as pages created by legitimate companies or governments, with consumers being directed to them in various ways, including false advertisements on social media platforms, e-mails, text messages¹, and so on. These pages generally mimic the appearance of the official websites in a convincing manner, helping to lure in customers².

2.1.2 Non-deliveries

Another common form of online consumer fraud is non-delivery fraud. In the simplest form of this type of offending online, the buyer and seller transact in an online marketplace, but while the seller receives payment, the product is never sent to the buyer (K. Anderson, 2019). For some online platforms (e.g., eBay), the buyer has the option of getting the money back from the payment system (e.g., PayPal) as well as writing a negative seller review and flagging the fraud quickly, which can make it difficult for offenders to sustain their activity using the same seller account over a prolonged period (Bauerly, 2009). However, this does not prevent fraud from taking place in the first place. According to the FBI, this type of offending accounts for monetary losses of more than \$180 million per annum in the USA (FBI, 2017, 2018).

In other variations of this offence, the seller convinces the customer to pay for the goods or service outside of the marketplace on which it was advertised (e.g., using a cheque, cash, or fake escrow). This type of fraud involves more effort on the part of the seller because it requires them to convince the customer to transact outside of the original platform. However, where successful, the online platform cannot confirm purchases, which makes the flagging of the fraud more difficult, allowing fraudsters to maintain their account for longer.

In other instances of this type of offence, the product may be misrepresented, with the customer receiving an item of lower value (e.g., counterfeit); they may also receive the packaging but not the actual item, or the product may be wrongly delivered on purpose. In the latter case, the seller has proof that an item was delivered and can shift the blame towards

¹ For example, <https://www.actionfraud.police.uk/alert/fake-dpd-messages-lead-to-over-200000-in-losses-since-june>.

² For an example, <https://www.gov.uk/government/publications/phishing-and-bogus-emails-hm-revenue-and-customs-examples/phishing-emails-and-bogus-contact-hm-revenue-and-customs-examples>.

the delivery service, which makes it more difficult to resolve the problem (Abdallah et al., 2016). The same techniques of product misrepresentations can affect not only small individual purchases, but also companies making large-scale industrial purchases, as in the case of painted stones sold as copper for \$36m (Harper, 2021).

2.1.3 Commonalities in online fraud schemes

The above-mentioned fraud schemes differ in how they are executed, but they all typically involve or require the presence of open or publicly available websites. Such websites provide an environment that can facilitate fraud. For example, the number of possible suitable targets (individuals that could be defrauded) fraudsters can reach is scalable, increasing the prospect of financial gains. Fraudsters also seem to have reacted to the changes in online market environments by adjusting their frauds from low-volume goods/services with a high price to high-volume goods/services with low prices (Nikitkov et al., 2014). Financial fraud has previously been explained using the fraud triangle, a perpetrator-centric view, which consists of the perceived (financial) pressure, perceived opportunity, and rationalisation of committing fraud (Cressey, 1953). The fraud triangle also details that the financial pressure is perceived as non-shareable and that opportunities to defraud are only exploitable if the associated risks are perceived as low. While previous studies applied the fraud triangle to explain fraud (Homer, 2020), some point out that the perpetrator-centric perspective is limiting in generating practical preventative measures against fraud (Mui & Mailley, 2015). Nonetheless, the fraud triangle can account for some fraud-facilitating properties of the online environment from the perpetrator's perspective. For example, perceived high anonymity by sellers (as present on online platforms) seems to facilitate the rationalisation of committing fraud (Harrison et al., 2020). Thus, changing the perception of anonymity, such as increased interactions between buyers and sellers, might contribute to fewer fraud instances (Harrison et al., 2020).

Unlike the fraud triangle, the Routine Activity Approach (RAA) and the rational choice perspective provide theoretical frameworks that include the environmental circumstances of offenders when committing a crime (L. E. Cohen & Felson, 1979; Cornish & Clarke, 1987). Some have argued that the fraud triangle can be situated within the RAA to provide a more complete picture of why fraud occurs (Mui & Mailley, 2015). Thus, by inspecting the elements of the fraud triangle (motivation, opportunity, rationalisation) and those of the RAA (e.g., guardians) the reasons for fraud occurrences can be analysed better. The RAA has also been extended to include where a crime occurs (i.e., place) and controllers, who act as supervisors upon the other components of the model (i.e., onto the target, offender, place) (Eck, 1994; Felson, 1995). The controllers include the handler (friends or relatives of the offender), guardian (protecting the target), and manager (responsible towards the place). If any of these controllers is ineffective, a crime is more likely to occur (Felson, 2008).

These theoretical frameworks can guide us on where we could implement possible (data science) measures to detect or prevent fraud. For example, considering online consumer

fraud, we can situate online (shopping) platforms as the place where fraud occurs and where motivated offenders encounter suitable targets. Site administrators or providers can be seen as managers (of the place) and as possible guardians. Other consumers who write reviews could also be considered possible guardians since they provide feedback on products and services that can inform possible buyers about the sellers and whether they might be fraudulent. Payment platforms or solutions (PayPal, banks, etc.) can also act as guardians by supervising monetary transactions. Since information from the buyers and the sellers intersect on the online platform, where fraud occurs, and we can observe and collect such information, we have the opportunity to exploit such data with techniques from data science. Moreover, information collection procedures and data analyses can be automated, and with appropriate planning, such methods might be implemented in a scalable way. If done well, this could potentially facilitate better guardianship of consumers and increase the efforts that offenders must engage in. What such approaches could look like are discussed in section 2.4. Such techniques can also be applied to other types of fraud not mentioned above, such as fraudulent computer repairs (Miramirkhani et al., 2017) or fraud-related phishing websites, which can be prompted through pop-ups or false advertisements on any webpage (K. Anderson, 2019, 2022; Christin et al., 2010).

2.2 Current strategies to prevent and resolve online consumer fraud

Online consumer fraud takes place in an environment in which many different industries and sectors converge. Commercial enterprises, non-governmental and governmental organizations adopt different strategies to deal with such fraud, and these are outlined in Table 2.1. Non-governmental organizations (e.g., Cifas: Credit Industry Fraud Avoidance System) mostly focus on trying to support (vulnerable) customers by informing them about different types of fraud to heighten their awareness of them. The goal of such strategies is to increase the likelihood that customers will spot fraudulent activities or listings while they are online and act accordingly (Beals et al., 2015; Deevy & Beals, 2013; M. DeLiema et al., 2019; Peaston, 2019; Stanford Center on Longevity, 2019). Recommendations to online shoppers often include (but are not limited to) the careful inspection of reviews, ratings, or details about the shop (e.g., physical location). The strategies to increase customer awareness require guidance to be up to date and that consumers are exposed to and act upon it. While this approach is probably useful, a lot must go right for it to work, and it should be noted that awareness does not equate to effectiveness (in prevention).

The commercial sector aims to maximize profit and various actors (e.g., brands, online shopping providers) deploy different strategies. Individual brands might seek support from brand protection agencies or utilize internal divisions to find brand violations and stop them. Since infringements directly impact the brands by denying possible sales and damaging the brands' reputation (e.g., through bad quality products), brands are incentivised to act and prevent infringements. In many cases, brands or brand protection agencies will search for possible violations (or react to complaints) through manual (or semi-automatic) searches and contact the sellers or website providers (Ganguly, 2015; Pointer Brand Protection, 2019;

Yellow Brand Protection, 2019). Preventing further infringements is often achieved through cease-and-desist notices or legal actions, such as civil litigations.

Online shopping platforms want to maximize profits and, similar to law enforcement agencies, often take a reactive approach by responding to complaints and intelligence about fraud. While online shopping platforms remove fraudsters from their websites, law enforcement is also concerned with prosecuting and subsequently deterring fraudsters (FBI, 2018; Intellectual Property Office, personal communication, 2019; Raine et al., 2015; Trading Standards UK, personal communication, February 26, 2019, personal communication, May 17, 2019) (see Table 2.1). Unlike brands, infringement sales do not burden online shopping platforms (e.g., Amazon, eBay) as much, and the number of individuals (sellers and consumers) using their platform is more important. Thus, unless the user numbers are declining and the platform's reputation is damaged, the incentives to invest in fraud prevention strategies or methods are slightly misaligned to brand owners.

Since law enforcement aims to reduce societal costs (e.g., monetary, psychological), their incentives align well with fraud prevention. However, in most cases, they react to consumer complaints and require the possibility of gathering evidence for investigations. Data access (e.g., where and how were frauds committed) and the capabilities to prosecute (due to jurisdictional issues) can limit the authorities' possible actions, often hindering investigations. Action Fraud³, the UK national reporting platform for fraud and scams, was founded in 2009 and aims to alleviate some of the data access issues (Committee of Public Accounts, 2023). Business owners or consumers can report any fraud to the platform, which sends the data to the National Fraud Intelligence Bureau (NFIB). The NFIB then assesses whether the report would warrant further investigations and sends it to the appropriate police agency. Over 300,000 reports from individuals and over 600,000 from businesses or the industry are filed at Action Fraud annually. However, only around 27,000 reports are sent to local police annually and of those, only around 5% result in an offender prosecution (Committee of Public Accounts, 2023). Thus, less than 1% of reports lead to any prosecution. The small number of successful investigations and low response rate to filed reports has led many users of the reporting system to frustrations and feelings of dismissive treatment (Button, 2021; Committee of Public Accounts, 2023). Furthermore, submitting a report takes around 20 minutes and will take an average of 54 days to be transmitted to the appropriate police agency, greatly diminishing the possibility of finding investigatory leads (Button, 2021). Next to issues in fraud reporting mechanisms, authorities are faced with a shortage of personnel capable of dealing with fraud. In 2022, only around 1% of the total UK policing personnel was dedicated to fraud, whilst it was measured to make up around 41% of all crime in the UK in the same year, showing a great mismatch between the needs and capacities (Committee of Public Accounts, 2023).

³ <https://www.actionfraud.police.uk/>

Current strategies used to combat online consumer fraud by the commercial sector and law enforcement are also heavily reliant on identifiable information about online users and vendors. In a non-anonymous space, such as the surface web, identifying individuals is generally feasible as long as users register for accounts using names, e-mails, or other information that can be used to identify them. In addition, the activity of their computers can be tracked through their IP address, making geolocation possible unless they use techniques to obscure this. Thus, authorities have some tools at hand to move against fraudsters. However, the long period of time that elapses between frauds being reported and received by authorities provides fraudsters with the time to erase most traces (e.g., accounts, funds, listings), greatly limiting the possibilities for successful investigations.

Ideally, attempted fraud would be detected as early as possible to prevent it from reaching many individuals. With early detections, workloads for authorities and online platforms, which act mostly reactively, would also be reduced. However, such detection approaches to identify attempted fraud would have to be implemented directly on online platforms outside of brands or the authorities' scope. Since online platforms have only limited incentives to invest in and develop detection methods that are currently not easily enforceable, such implementations are less likely. Thus, two approaches could be followed to combat fraud: reducing the hurdles for online platforms to implement already developed methods and reducing the workload for authorities through partial automation of their manual work. The following two sections (2.3, 2.4) will discuss some data science methods in the academic literature for actively detecting fraud on online platforms, their limitations and what future approaches could look like.

Sector	Motivation	Strategy	Actions/Controls	How to Resolve Fraud
<p>Non-governmental (e.g., Cifas)</p> <p>(Beals et al., 2015; Deevy & Beals, 2013; M. DeLiema et al., 2019; Peaston, 2019; Stanford Center on Longevity, 2019)</p>	<p>Protect (vulnerable) customers</p>	<ul style="list-style-type: none"> Using surveys to estimate fraud prevalence and to identify vulnerable customers Creating fraud avoidance guidelines 	<p>Recommending individuals to scrutinize:</p> <ul style="list-style-type: none"> seller feedback and comments price and seller history 	<ul style="list-style-type: none"> Increase the likelihood of consumers spotting suspicious listings
<p>Commercial industry</p> <p>(Ganguly, 2015; Pointer Brand Protection, 2019; Vistalworks, 2019; Yellow Brand Protection, 2019)</p>	<ul style="list-style-type: none"> Maximizing profit 	<ul style="list-style-type: none"> Hiring brand protection companies 	<ul style="list-style-type: none"> Searching for brand violations (manually, semi-automatically – "AI") in product/service domain Reacting to complaints 	<ul style="list-style-type: none"> Contacting fraudsters Notifying fraud to company/platform pursuing legal actions
		<ul style="list-style-type: none"> Creating internal divisions to detect fraud (e.g., eBay, Amazon) 	<ul style="list-style-type: none"> In-person visits Test purchases Customer flagging systems Automated monitoring of item sales, transactions, and product view ratios 	<ul style="list-style-type: none"> Removing sellers from the webpage
<p>Law Enforcement (e.g., FBI, Trading Standards)</p> <p>(FBI, 2018; Great Britain & National Audit Office, 2016; Intellectual Property Office, personal communication, 2019; Raine et al., 2015; Trading Standards UK, personal communication, February 26, 2019, personal communication, May 17, 2019)</p>	<ul style="list-style-type: none"> Deter/detect criminals, reduce the negative financial impact on society 	<ul style="list-style-type: none"> Respond to fraud complaints, intelligence about fraud Prioritizing fraud types and cases depending on financial damage 	<ul style="list-style-type: none"> Manual investigation of received intelligence Cross-referencing of product and seller information 	<ul style="list-style-type: none"> Find, identify, and deter fraudsters Inform affected platform

Table 2.1. Summary of How Different Sectors Address Online Consumer Fraud

2.3 The role of data science

Data science has been previously defined as "*a set of fundamental principles that support and guide the principled extraction of information and knowledge from data*" (Provost & Fawcett, 2013, p. 2). This definition captures a high-level perspective, but data science also involves "*statistics, or the systematic study of the organization, properties, and analysis of data and its role in inference, including our confidence in the inference*" (Dhar, 2013, p. 1). Thus, data science draws from many disciplines, including computer science, mathematics, statistics, and, importantly, the knowledge domain of the data in question (Dhar, 2013; Provost & Fawcett, 2013). In the realm of online consumer fraud, data science can be utilized to speed up or scale up manual processes, such as collecting, cross-referencing, and analysing information from advertisements, product listings, sellers, or webpages to identify if they are likely to be fraudulent. Such methods will be reviewed in this section by looking at current and possible future applications and their associated advantages and disadvantages.

2.3.1 Current data science approaches in the academic literature

Current data science methods applied to online consumer fraud on the surface web mainly focus on the automated detection of specific types of fraud (Abdallah et al., 2016), which would (in theory) help to increase active guardianship and increase the effort required for offenders to go undetected. For example, Pandit et al. (2007) collected eBay data and modelled the topological connections (i.e., the network) between sellers and buyers using transaction data. They hypothesized that fraudsters, who commit the crime, should be heavily connected to accomplices, who boost the fraudsters' feedback ratings. Accomplices are needed as they can continue to operate once the fraudster is banned from the site, retaining any positive ratings they accumulate. As such, fraudsters would be expected to have fewer connections to honest users than they would have to accomplices. Thus, the authors describe that fraudsters have many connections to accomplices, but the accomplices and the fraudsters are not interconnected, forming near bipartite cores instead of cliques, which exhibit strong interconnectedness. The automated detection of such bipartite cores was evaluated on a synthetic network containing over 66,000 nodes and 795,000 edges. The network was filled with artificial fraudster-accomplice structures of random sizes, which were detected with around 90% accuracy. An additional network was created from scraped eBay data, which contained ten known fraudsters, who were identified through manual inspections and investigative media reports. All ten fraudsters were automatically detected by the authors' devised model. As the researchers discuss, the evaluation based on the analysis of the eBay dataset has limitations since no fraud detections (other than those already detected through a manual investigation) could be verified. As such, additional verification of the models' performance through a well-labelled data set would be needed to determine its applicability (Pandit et al., 2007).

Hernandez-Castro and Roberts (2015) used an automated detection method that can process data without human intervention to try to identify illegal sales of elephant ivory on eBay. Specifically, they used the CN2 induction algorithm (Clark & Niblett, 1989), which is a

supervised machine learning method. Supervised methods require a training phase, in which they learn to infer from features (e.g., metadata or writing style of the advertisements) what the corresponding label (e.g., fraudulent versus legitimate, or same versus different account holder of a vendor profile) of a data point is likely to be. The goal is typically to find a combination of features that enable the best discrimination of the outcome labels. The labels are often obtained through manual annotations, which can be time-consuming, particularly where a large volume of labels is required. However, the advantage of such approaches is that they can uncover previously unknown patterns within the data and utilize these to classify unlabelled data. Such methods find a wide range of applications that could be used in various fraud detection domains. In the study by Hernandez-Castro & Roberts (2015), two former law enforcement officers annotated 1,159 product listings as “selling ivory” or “not”. Utilizing the CN2 induction algorithm (Clark & Niblett, 1989), which made decisions based on metadata parameters of the listings (e.g., item price, number of reviews), the framework was able to categorize almost all listings correctly. The advantage of the CN2 algorithm is that it induces decision rules, which can be inspected to gain an understanding of the sale strategies of potential fraudsters. Other supervised machine learning methods, such as random forest or logistic regressions, are also interpretable, but many others, such as neural networks, can be black boxes since the decision rules are generated automatically, making it difficult to understand the inner workings of the classification system and hence the possible fraud strategies. While the system developed by Hernandez-Castro and Roberts (2015) could find suspicious online listings quickly without manual searchers, it is difficult to assess the reliability of their current classifier, as the true labels of the listings (selling ivory or not) were not known, and no inter-rater reliability score between the annotators was reported (Hernandez-Castro & Roberts, 2015). However, a similar approach of using domain experts to label instances of online consumer fraud (e.g., counterfeits on eBay) could be helpful. For example, classifiers trained using these data could then act as a pre-selection (or triaging) tool by labelling suspicious listings. Although such an approach does not eliminate the problem of falsely labelled fraudsters, it could reduce the number of listings experts subsequently have to investigate, reducing the overall workload. The model could then be updated after each investigation to increase its performance and reduce false positives.

Other researchers also utilized supervised classification methods, which are trained and tested on different subsets of labelled data (Almendrea & Enachescu, 2012; Chang & Chang, 2012; Sahingoz et al., 2019; Xu et al., 2016). For example, Chang and Chang (2012) used decision trees to detect fraud on the Yahoo Taiwan online auction site. Decision trees continuously split the data into subgroups (e.g. [non-] fraudulent seller) based on a binary decision about a predictive feature (e.g., number of negative reviews). As an example, a seller might be labelled as fraudulent if more than half of the reviews are negative. The remaining unlabelled sellers undergo more splits based on other feature values until all data are categorized (for an introduction to machine learning methods, see (Rosenbusch et al., 2021)). In the study by Chang & Chang (2012), the decision trees were trained on the history of sellers' activities, and the sellers were labelled as legitimate or fraudulent, depending upon whether they had been blacklisted from the auction site or not. While the performance of the model

was promising, the labelling process made the evaluation problematic. That is, it remains unclear why users were blacklisted and whether and how many blacklisted sellers were genuine. Supervised machine learning methods have also been employed to detect phishing websites by examining their URL (Sahingoz et al., 2019) or HTML structure (Xu et al., 2016). The approach is the same as that described above – using an annotated dataset, the algorithm learns to associate the URLs or the webpage's HTML structure with the correct label to infer a decision rule. This is then used to classify new incoming data (webpages) that do not have labels.

An important aspect of machine learning is that many (but not all) supervised methods rely on the selection and crafting of informative features, which are the parameters the algorithms learn from. Thus, the goal is often to create features that are readable by the algorithm and convey a high amount of information. Machine learning methods need numerical features to work, which means that in the case of non-numerical data, such as text, the creation of features becomes more complicated as it requires the data to be converted to numerical representations. This can be achieved using Natural Language Processing (NLP) methods, which essentially work at the intersection of computer and human language to bridge the understanding of the two domains (Jurafsky & Martin, 2019; Nadkarni et al., 2011). By utilizing NLP methods, text (e.g., characters, words, sentences) can be converted to numerical data, which can be processed and analysed by a computer. This can be as simple as creating frequencies of words that occur in the text, but also includes numerical representations of grammatical structures or semantic meanings (Goldberg & Levy, 2014). For a more detailed overview of NLP methods and how they function, readers are referred to Goldberg (2016), Jurafsky and Martin (2019), and Nadkarni et al. (2011).

Supervised machine learning methods are powerful tools which can uncover previously unknown patterns in large amounts of data. However, supervised models need well-created data sets with suitable features and labels, which are not always easy to obtain. In particular, the labelling process is often problematic as the "ground-truth" of the labels is often unknown (see above).

2.3.2 Ground truth

In the context of this thesis, ground-truth refers to the true labels of a data set. The knowledge about ground-truth or certainty of the label could be about an item on a selling platform, a confession, or anything else. In an experimental setting, in which researchers have full control, true labels can be obtained. For example, participants might be asked to test a pair of purchased headphones and then follow instructions to write an honest and fake review (with opposing sentiments) about them. However, outside of an experimental setting (and in most other cases), the ground-truth is not as unambiguous, and only a level of certainty can be attributed to a particular data point. For example, how would we know if a shoe advertised on an online platform was counterfeit or legitimate? A common strategy employed by consumers is to choose a reputable seller by inspecting their ratings and reviews (Dellarocas,

2006; Houser & Wooders, 2006; Melnik & Alm, 2003; Resnick et al., 2000). To provide a better label, an expert could assess the listing, or better still, order a pair of shoes and compare them to a sample purchased directly from the manufacturer. Each method will provide an assessment of the listing with different levels of certainty about the labels (counterfeit vs genuine) associated with an item or event.

In many cases, obtaining labels with a high degree of certainty will either be impossible or costly in terms of time, money, or both (Raine et al., 2015). In some instances, officials from online marketplaces or law enforcement (e.g. Trading Standards) order products from sellers or visit the sellers' physical locations to conduct manual inspections (Ganguly, 2015; Raine et al., 2015), but this is costly. Furthermore, transparency during the labelling process is important, as is the clear conveyance of the label's limitations, to ensure that a judgment about the associated certainty of the labels is possible without needing further expert knowledge. This is especially important to law enforcement, for whom resources are limited and misallocations can be costly. The strategy to minimize mislabelling is to reduce false positives labels (e.g., mislabelling a genuine product to be fraudulent) and false-negative labels (e.g., mislabelling a fraudulent product to be genuine). Training a model to work well on a poorly labelled dataset is of little value, and consequently, the determination of the data labels on which a supervised model is trained is paramount. Therefore, the value of a model should not be judged on performance metrics alone but also on how the data were acquired and labelled for each context.

2.3.3 Theoretical perspective for utilizing data science approaches

Situational crime prevention (SCP) is a model of crime prevention that focuses on reducing the opportunities for crime to occur (Clarke, 1980). SCP draws from other theoretical perspectives, such as the rational choice perspective and the routine activity approach and is concerned with practical implementations. Thus, in contrast to understanding why individuals commit a crime, SCP focuses on how a crime is committed while integrating the situational circumstances to create interventions that remove opportunities, aiming at preventing offending (Clarke, 1995; Freilich & Newman, 2018). The SCP framework comprises 25 techniques designed to reduce crime, which can be grouped into: increasing the effort, increasing the risks, reducing the rewards, reducing provocations, and removing excuses for committing a crime (Cornish & Clarke, 2003; Freilich & Newman, 2018). Each category contains five more granular techniques, such as target hardening (e.g., tamper-proof packaging), extending guardianship (e.g., signs of occupancy when leaving home), or reducing the anonymity of possible offenders, to name some examples. See Table A1 for all 25 situation crime prevention techniques provided by Cornish & Clarke (2003).

Drawing from those 25 techniques, we can examine how applicable they might be to cyber-enabled crime, such as online consumer fraud and consider how current and possible future data science approaches could support the implementation of such prevention techniques. Measures that would increase the effort or increase the risks for offenders seem to be most

suitable to be implemented by data science approaches in the context of online consumer fraud. For example, measures that would control access to facilities (e.g., electronic card access, baggage screening), when applied to cyberspace, could translate to further granular automated verifications or checks of individuals who want to trade on consumer platforms. Other measures that would increase the risks for offenders, such as extended guardianship or formal surveillance (e.g., alarms, cameras), could be realised through automated fraud detection approaches in cyberspace, as described above. Moreover, natural surveillance (e.g., improved street lighting) could translate to robust transaction, flagging, or reviewing systems and could be further strengthened by implementing verified purchased reviews. Utilising place managers might entail similar measures to increase the engagement between the platform and its users, such as escrow systems. Reducing anonymity is another measure listed to increase the risks to offenders. Some research shows that offenders believe there is a low risk of being detected partly due to anonymity (Hutchings, 2013). Thus, advanced verification or registration procedures to make individuals who want to trade on consumer platforms more identifiable could be useful. One such measure might entail cross-referencing individuals who want to register on a platform with a list of previously identified fraudsters. Such individuals might have been identified as fraudsters on the same or different platforms, and their details are shared through a centralised database. The national fraud database in the UK, facilitated by Cifas (Non-governmental organisation for economic crime)⁴, supports such an approach for businesses.

Similarly, we can revisit the existing strategies from (non-)commercial sectors, such as the non-governmental institution Cifas, which spends some of its efforts on strategies to increase the awareness of potential fraud victims (suitable targets). Thus, Cifas aims to increase the resilience of individuals to be defrauded (i.e., target hardening), which would increase the effort for fraudsters. Strategies employed by authorities, or the commercial sector are often reactive and are more concerned with punishment than limiting the opportunities for offending. However, potential punishments could ultimately also increase the perceived risk of offending. While law enforcement is essential, it cannot fully deter fraudsters, and current enforcement approaches cannot keep up with the number of fraud cases, as discussed previously. Thus, the next section (2.4) discusses how data science methods could be applied to support fraud prevention and law enforcement.

2.4 Possible future data science applications

As described above, currently discussed data science approaches in the academic literature often focus on supervised machine learning methods, which would support a proactive approach to the detection of possible online frauds before people are affected. Thus, facilitating a preventative approach and limiting the opportunities for offenders. However, given how the class labels (e.g., genuine vs fraudulent listing) used for these approaches are obtained, it is unclear if such detections would always warrant a lawful intervention (Abdallah

⁴ <https://www.cifas.org.uk/>

et al., 2016; Almendra & Enachescu, 2012; Chang & Chang, 2012; Hernandez-Castro & Roberts, 2015; Pandit et al., 2007; Xu et al., 2016). That said, detection approaches could be used to pre-sort and prioritize cases (e.g., suspicious advertisements, sellers, webpages) which authorities, brand owners, or online (shopping) platform provider could then focus on manually, making better use of limited resources.

Other areas in which data science methods could alleviate the current workload of law enforcement or online platform providers are in gathering, pooling, and analysing information about reported items, services, sellers, or webpages (Great Britain & National Audit Office, 2016; Raine et al., 2015; Trading Standards UK, personal communication, February 26, 2019).

Data science can also be used in other disciplines, such as psychology, to investigate fraud on a more individual level. For example, machine learning models could be utilized to uncover underlying patterns of how and why certain people fall victim to fraud. In turn, the same approaches could be used to uncover patterns associated with individuals acting as a fraudster. However, discussing such use cases in more detail is outside of the scope of this thesis, which aims to highlight more direct applications usable by practitioners or law enforcement.

2.4.1 Automation of manual tasks

An initial step in alleviating the workload of practitioners to support law enforcement would be to automate the standard processes of gathering and pooling information relevant to a reported online fraud (e.g., reported from a consumer or a company). Although manual reporting tools⁵ such as provided by Action Fraud already exist, they can be highly time-consuming and might be ineffective. Information gathering and pooling could be addressed by using rule-based systems that could automatically collect information (e.g., telephone numbers, e-mail addresses, prices, product- or seller names) from webpages that are reported. Additional automated web searches could supplement and cross-reference the existing information, including whether they exist and in what capacity they are registered (e.g., does the physical address of the seller exist and is the business legally registered with the relevant authority). Partially automating fraud related investigatory processes would also allow more personnel to work on fraud cases, who lack technical skills. Thereby, lowering the hurdle of training police personnel to work on fraud cases.

2.4.2 Uncovering unknown patterns in existing data

Existing data could be analysed using unsupervised machine learning methods that do not rely on labelled training data. Here, the idea is to cluster the data based on inherent traits to find common features or properties that may not be directly apparent (Arthur & Vassilvitskii, 2006; Ester et al., 1996; Liu et al., 2012). For example, a consumer-reported incident of fraud or

⁵ <https://www.actionfraud.police.uk/reporting-fraud-and-cyber-crime>

counterfeit items on shopping platforms might be clustered into several groups. These groups could be based on similarities in their descriptions, prices, item locations, and so on. Such similarities could point towards meaningful associations, which may open up new leads in an investigation. For example, items that are identified as sharing features, such as similar descriptions, seller locations, and images but are advertised by sellers with different usernames, might point to a common seller who wishes to conceal their identity and fraudulent activity.

2.4.3 Understanding online markets at scale

It is important to understand the online markets in which consumer fraud occurs as it places the reported fraud cases into context and may inform new fraud detection strategies. Data science can help by accumulating, structuring, and analysing data for online markets, such as eBay, Amazon, or Alibaba. By devising scraping methods for these markets, large amounts of data for products and sellers can be collected. Through NLP methods, it is also possible to operationalize and measure text characteristics and styles from titles, descriptions, reviews, and so on. Understanding markets with such data will (we hope) enable us to update our beliefs about common product and seller characteristics, determine what constitutes "normal", and understand how fraudsters might behave. Thus, understanding how frauds are implemented or how fraudulent products might look like will help in devising preventative measures. For example, learning about counterfeits (e.g., shoes, apparel, electronics) sold on cryptomarkets could inform us about, which product types are more prone for counterfeiting, which is valuable for manufacturer and brands, who can act on such information. Manufacturing, product transportation and validation procedures from such products could then be examined by affected brands, aiming to reduce possible exploits and increase the efforts for counterfeiters to manufacture products or re-introduce them into the supply chain (Hollis & Wilson, 2014). What cryptomarkets are, and how we could use data from such markets will be discussed next in more detail.

2.5 Anonymity networks

The Internet can be segmented into three sections, which are characterized by how it can be accessed and how users communicate with each other: the surface web, the deep web, and anonymity networks (Figure 2.1) (Biddle et al., 2003; Mansfield-Devine, 2009).

The surface web – which has been the focus of discussion in this chapter so far – is the Internet we usually encounter and includes platforms such as eBay, YouTube, and news sites. Put differently, it contains the online content that is indexed by openly accessible search engines (Bergman, 2001; He et al., 2007) such as Google's.

The deep web represents the part of the internet that is not indexed by search engines and cannot easily be accessed. Content is often password-protected or restricted in other ways (e.g., requiring authentication). Online banking, webmail, or paywall content would fall within

this category. Historic estimates (Bergman, 2001; He et al., 2007) suggest that the deep web is 400-550 times larger than the surface web in terms of the amount of information (data) and web pages stored. The deep web is believed to be the fastest-growing category of the internet, but it is difficult to estimate its current size precisely, as it is not openly indexed. For both the surface and deep web individual users and servers are not automatically anonymous.

Anonymity networks represent a small portion of the deep web, on which users and hosts are anonymized. Anonymization can be facilitated through different technologies, such as The Onion Router (Tor) (The Tor Project, Inc., 2020) or the Invisible Internet Project (I2P) (The Invisible Internet Project, 2020). Such methods hide the identities of users by sending their data through a network of computers and servers that use protocols to conceal their IP addresses and serve as relays (Gehl, 2018). Hosts using the Tor system can provide web pages called “onion sites”. Navigating to an onion site requires an exact address because they are not indexed by search engines and the administrative markers, such as “.com”, “.net”, or country codes, such as “.de” (Germany), “.us” (United States), are replaced by “.onion” (Ghosh, Porras, et al., 2017). While the anonymity networks were not developed with malicious intent, the safety of anonymous communications makes illegal activities less dangerous for the perpetrators as some of the usual information that is used to track fraudsters is removed.

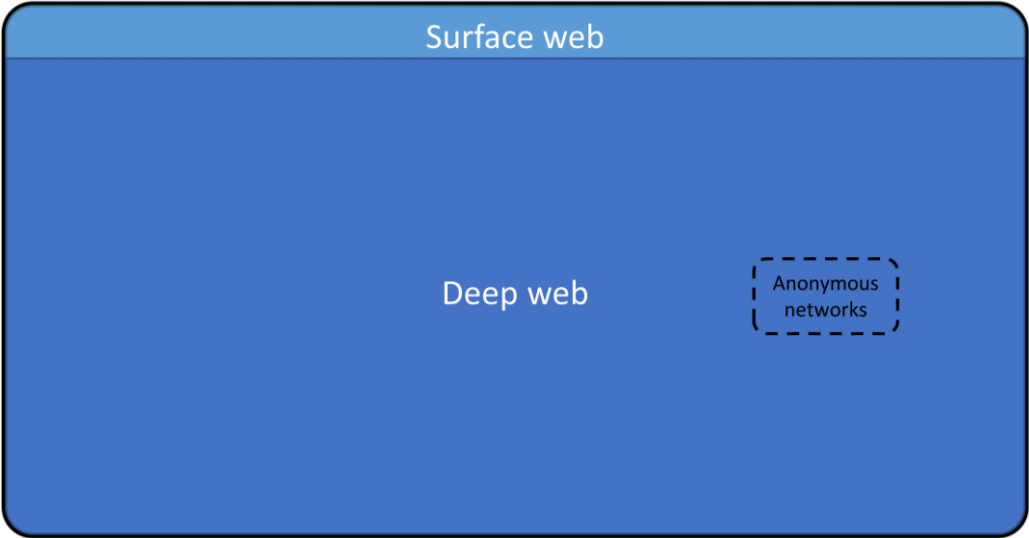


Figure 2.1. The World Wide Web subdivided into surface, deep, and anonymous web. The domain proportions are approximations, as the deep web would dwarf the other domains to be unrecognizable.

The Tor network, which is perhaps the most widely known, can be accessed through the Tor client, which functions like a regular browser. The ease of entering the anonymity network makes it accessible to many individuals, some of whom will want to engage in fraudulent activities. Estimates of the number of “.onion” web pages range from 3,700 to 32,000, with a maximum of around 13,000 exhibiting prolonged activity (Ghosh, Porras, et al., 2017; Gray, 2019; Lewis, 2016).

2.5.1 Markets on anonymity networks

Anonymity networks harbour selling platforms, which are called “cryptomarkets”, but also “black markets”, “dark web markets”, “dark markets”, or “darknet markets”, which offer various products and services (Christin, 2013; Gehl, 2018). These platforms are not online shops but — like markets such as eBay — provide spaces for users to transact (Soska & Christin, 2015). Similarly, platform providers like eBay receive a small margin of each monetary transaction. Cryptomarket transactions are anonymized through the use of cryptocurrencies, which can be obtained from online exchanges (e.g., Coinbase). Most of these currencies are based on a peer-to-peer system that does not rely on a bank or other centralized third party (Soska & Christin, 2015); (for more information on cryptocurrencies, see Kamps et al. (2022)). The use of cryptocurrencies as well as anonymized communications further facilitates illegal activity and makes tracking fraudsters very difficult. The Silk Road, which was the first commonly known cryptomarket, started operating in February 2011 and used Bitcoin (BTC) as a medium of transaction (EMCDDA-Europol, 2017). Buyers did not pay the seller directly but used an escrow system — a form of holding area — embedded within the platform. Escrow systems allow platform operators to compute commissions as well as to supervise transactions between buyers and sellers to ensure that products are shipped (or services are provided) only after payments have been received (Christin, 2013). Thus, transactions can be completed in a highly anonymized space, in which accountability is otherwise almost absent.

Estimates suggest that for larger markets (e.g., Silk Road 1 & 2, Agora, AlphaBay)⁶, about 50-80% of all category listings are for various forms of drugs (e.g. Cannabis, Amphetamines) (Demant et al., 2018; Europol, 2017; Soska & Christin, 2015). However, a large range of other products are also typically offered including weapons, counterfeit goods, guides for malicious activities, stolen credit or debit card information, other personal data (e.g., log-in credentials), or services such as hacking attacks or secure hosting (Adamsson, 2017; Du et al., 2018; van Wegberg et al., 2018). By way of an example, Figure 2.2 shows a screenshot from one cryptomarket, White House Market, which still operates as of today (25.06.2021).

⁶ The mentioned markets do not operate anymore.

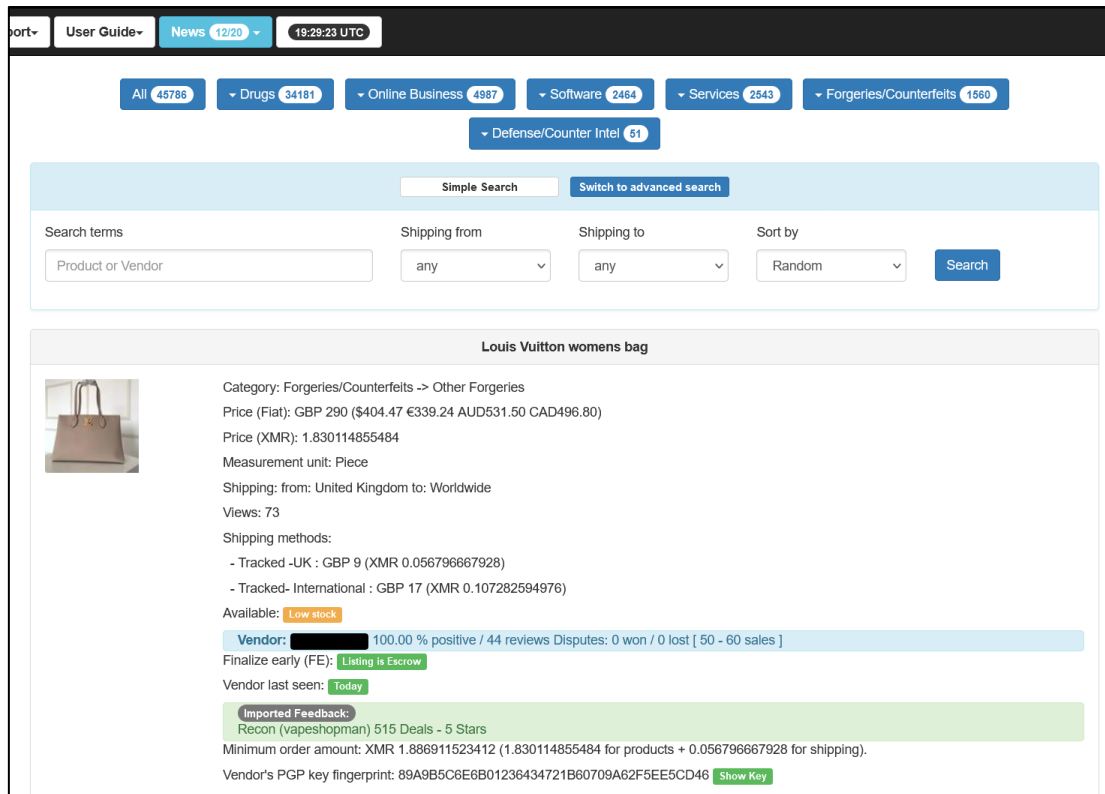


Figure 2.2. Example screenshot of a counterfeit listing (Louis Vuitton handbag) on the cryptomarket White House Market.

2.5.2 Obtaining cryptomarket data

Collecting cryptomarket data can be complicated as such data are rarely shared by the platforms. One way of collecting cryptomarket data is by visiting the websites as a user and retrieve it manually. However, the most commonly used method is to employ a web scraper that revisits markets at set intervals (e.g., once a day) over weeks, months, or longer (Ghosh, Das, et al., 2017). Such web scrapers retrieve information from a web page and download it to a local machine. Employing web scrapers is relatively straightforward on the surface web, it is time-consuming on anonymity networks due to the rerouting of page requests – necessary to create anonymity – through the TOR network, which takes significantly longer than on the surface web. Moreover, since cryptomarket sites often have different layouts, almost all scrapers need to be coded individually for each website (Du et al., 2018; D. Hayes et al., 2018). Also, researchers are often faced with unexpected problems during the scraping process, which requires a constant review of the collected data. This can entail changes in accessibility credentials, changes in the layout of the page, downtime of the host, advanced CAPTCHAs, or blocking when browsing too fast (e.g., clicking on many website links in a short time) (Ball et al., 2019; Ghosh, Porras, et al., 2017). Given that scraping is the predominant method of obtaining cryptomarket data, discussions of these challenges are well-rehearsed elsewhere and the interested reader is referred to studies by (Ball et al., 2019; Du et al., 2018; Ghosh, Porras, et al., 2017; Van Buskirk et al., 2014, 2015, 2016).

2.5.3 Cryptomarkets' economies

The Silk Road 1 operated from 2011-2013 and was the first cryptomarket studied by researchers, who examined what and how much was being sold (Christin, 2013). This market had between 30,000-150,000 active customers, 220 distinct product categories, and vendors who made an overall monthly revenue of \$1.2 million in 2012 (Christin, 2013). Drugs were predominately offered, and half of all products sold were shipped worldwide, mostly originating from the US (43%), followed by the UK (10%) and the Netherlands (6%). In their study, Soska and Christin (2015) estimated transactional and sales volumes through the analysis of product feedback for 32 different marketplaces for the period 2014-2015. They concluded that sales volumes typically varied between \$300,000-\$500,000 per day. With an estimation of 9,386 unique vendors across markets, 70% made less than \$1,000 during their active time on the platform, while 1% accounted for more than half of all sales (Soska & Christin, 2015).

Baravalle and Lee (2018) examined AlphaBay between 2015 and 2017 and estimated sales to reach \$79.8 million over the two years, with \$69.2 million of this attributed to drugs and chemicals alone. They also estimated that \$1.7 million of sales were generated from fraud (e.g., the sale of fake IDs, or accounts), \$1.6 million from counterfeit items, \$1.4 million from hacking attacks or server hosting, \$748,544 from Software and Malware, \$210,000 from Guides and Tutorials about committing fraud, \$198,000 from digital products, \$35,990 from Security and Hosting, and \$125,472 from other listings. This suggested that counterfeit goods and goods and services related to fraud were well represented on the market.

A study by van Wegberg et al. (2018) investigating cybercrime commodities on eight cryptomarkets between 2011-2017 estimated revenue of at least 15 million USD. The study showed that business-to-consumer rather than business-to-business transactions were taking place, and that cybercrime commodity trade exhibited some overall growth.

Multinational governmental institutions, such as Europol are increasingly interested in cryptomarkets, particularly those associated with the drug economy, but also those that facilitate intellectual property (IP) crimes (EMCDDA-Europol, 2017; Europol, 2017). Based on data from five cryptomarket (AlphaBay, Dream Market, Hansa, TradeRoute, and Valhalla), Europol estimate that fraud and counterfeits accounted for around 17% of all listings. Examining AlphaBay specifically, EMCDDA-Europol (2017) estimated that 10,000 products were counterfeit goods. Fake banknotes and IDs were the most frequent, but clothes, electronics, jewellery, software, e-books, subscriptions, and watches were also common. Fraud seemed to be well represented too, accounting for around 22% of all listings on Alphabay (Adamsson, 2017). Europol (2017) reports that vendors often sold small amounts, mostly specialized in one type of product but were present in several different markets at the same time. EMCDDA-Europol (2017) concluded that IP crimes were increasing on cryptomarkets, but that products are not always clearly categorized, which makes it difficult to accurately estimate volumes of offers and sales.

2.5.4 Vendor and user behaviour

The sale of products or services on the Internet requires a level of trust between the customer and vendor. This is true on both the surface and the anonymized web, with trust often established through feedback systems such as reviews or ratings. For many cryptomarkets, reviews can only be submitted after a verified purchase (verified by the escrow system) and some markets make reviews mandatory (Calis, 2018). Since everyone is anonymous and no public or physical stores exist, the review system is a particularly important feature of cryptomarkets, along with escrow. Although there are no findings of fraud on cryptomarkets, the presence of such review, escrow, and registration procedures seem to suggest that fraud occurs frequently. Such security procedures are not new but are still often missing on regular surface web platforms and it would be interesting to investigate their effects on the prevalence of fraud on the surface web.

2.5.5 Methods of vendor identification on cryptomarkets

Identifying individuals on anonymity networks is inherently difficult due to their anonymity (Gehl, 2018; The Invisible Internet Project, 2020; The Tor Project, Inc., 2020). Nonetheless, some studies have analysed text data or product photos to attempt to identify individuals or vendors with different user names that operate across platforms (Ho & Ng, 2016; Wang, 2018; Wang et al., 2018). Photos are important in markets since they serve as a form of proof that the vendor owns the product advertised. At the same time, pictures are only one aspect of building trust, and they should be identifiably different from one vendor to the next, to highlight ownership. Therefore, Wang et al. (2018) suggested that photo styles could serve as an identification of vendors across markets. To examine whether this was plausible, they used transfer learning, for which a machine learning classifier was pre-trained on a large data set (ImageNet) and further fine-tuned by retraining it on a smaller data set (vendor-specific photos). The task for the classifier was to identify whether two or more vendors were the same based solely on the images used. The model performed well for a single market for which a subset of vendor and photo associations were known. However, the performance was more difficult to assess for matches across markets as their true association was unknown. Nevertheless, the approach offers a possible identification methodology for an otherwise anonymized environment, which, although not suggested by the authors, could also prove useful on the surface web. For example, once a fraudulent vendor is identified, other vendor accounts across platforms (e.g., eBay, Amazon) operated by the same fraudster(s) might also be identified in that manner.

Another method of identifying vendors across cryptomarkets is through the use of PGP (Pretty Good Privacy) keys, which allow for encrypted communication between individuals (Ailipoaie & Shortis, 2015; Booij et al., 2021; Soska & Christin, 2015). Two PGP keys are required, a public key provided by users that receive messages (e.g., a key made visible on the vendor page), and a private key held by the same user privately. The public key is used to encrypt messages by anyone who wants to send a message to the public key holder, and the private key is used to decrypt the sent message. Thus, only users with the correct private key can read the

encrypted message. While the intention of PGP keys is to communicate privately, public keys can also be used to identify individuals with different usernames across markets. Such identification is not only used by researchers to estimate the number of vendors but also by users who want to verify their identity across markets to retain previously built reputations (Ailipoaie & Shortis, 2015; Booij et al., 2021; Soska & Christin, 2015). Such identification can be important when vendors want to advertise across markets or when markets unexpectedly close and users migrate to a new one.

In a different study, a series of automated methods, combined with manual investigations were used to identify vendors on cryptomarkets (D. Hayes et al., 2018). The researchers collected data from vendors and their associated listings from an undisclosed collection of cryptomarkets. Using Maltego, which conducts automated surface web cross-referencing (Paterva, 2019), they searched for obtained e-mail addresses, user names, and other personally identifiable information, and were able to identify some cryptomarket vendors on the surface web. As noted by the authors, this approach has some limitations associated with identifiers, which are similar to high-frequency words. For example, a username that resembles a popular brand or product name (e.g., Coca-Cola) would lead to meaningless cross-referencing. Nevertheless, their results suggest that the automated identification of some individuals is possible, but can be easily disrupted when users employ countermeasures, such as adapting their identifiable information (e.g., username, e-mail address) accordingly.

Importantly, the methods outlined show possible ways of identifying vendors across cryptomarket, which might also be applicable, in some circumstances, for the identification of fraudsters across markets on the surface web as well as fraudulent products between anonymized networks and the surface-web.

2.6 Utilizing cryptomarkets to identify consumer fraud on the surface web

Identifying illegal activity on the surface web is a challenge since most products and services are advertised as being genuine or legal. In contrast, on anonymity networks, most of the services are by definition illegal and generally advertised with no effort made to conceal this. This raises an interesting possibility: information found on anonymity networks could serve as a source of intelligence to determine what illegal activity is or might happen on the surface web. Although categories on cryptomarkets surrounding consumer fraud only represent a small portion (17%) of what is sold or traded (EMCDDA-Europol, 2017; Europol, 2017), they are often labelled as such. Thus, such information could serve as a source of ground truth data for illegal products and services on the surface web, where most fraudulent sales occur. Moreover, to increase profits, it appears that vendors sell on multiple cryptomarkets and the surface web (Europol, 2017), suggesting that cross-domain referencing could be of value.

As far as we are aware, to date, research has not examined this possibility of using such data to inform fraudulent product identification on the surface web. The focus of such cross-referencing would be the identification of counterfeit goods, which is the main category of

items on anonymity networks that also appears on the surface web. However, fraud-related categories, such as personal credentials, guides, and tutorials may also be of interest. For example, if hacked seller accounts (e.g., eBay), or specific fraud strategies are frequently advertised, their occurrence could represent an indicator of security breaches, which might be worth investigating. To examine the feasibility of such a cross-referencing approach, historical data from anonymity networks could be compared to law enforcement datasets of online consumer fraud for the same periods. In the case that such connections can be made, further work could focus on monitoring current cryptomarkets and cross-referencing them with current listings on the surface web. Since counterfeits on cryptomarkets are mostly clearly labelled and contain many product details (e.g., title, description, price, and pictures), a similar automated cross-referencing approach could be taken as has previously been described in studies (see above) that have sought to identify individual vendors (D. Hayes et al., 2018). Identifying fraud (enabling) services, such as hosting phishing sites or well established eBay accounts from anonymity networks on the surface web, might not be as easy, as they often do not have similar listed details as fraudulent products before the sale (K. Thomas et al., 2015). However, depending on the level of details provided by vendors offering such services, these listings could also serve as an informative tool for researchers seeking to identify illegal activity on the surface web.

2.7 Conclusion

The costs associated with online consumer fraud are significant, and the detection and prevention of such fraud using data science approaches have seen some progress. However, these approaches are not yet well integrated or used outside of the academic literature. Using and implementing data science methods is complicated by a range of factors, such as the interplay between the different sectors and jurisdictions that are affected, seller anonymity, and data labelling. A pressing issue is the volume of fraud complaints, most of which cannot be dealt with appropriately due to the lack of personnel and the highly time-consuming manual processing of such cases. Data science approaches would offer some help in automating and speeding up this (currently) manual work. Additionally, cryptomarkets are increasingly utilized to sell fraudulent products online, such as counterfeits, or fraud enabling services. However, it is still unclear to what extent and how fraudulent transactions on these markets are related to those on the surface web. Nevertheless, such listings on anonymity networks could serve as an informative tool for detecting and identifying fraudulent behaviour on the surface web.

Chapter 3: Challenges in annotating training data for supervised machine learning models

3.1 Introduction

Fraud on online shopping platforms is an increasing problem calling for adequate detection and prevention methods (M. I. DeLiema et al., 2017; FBI, 2017, 2018). The current approaches adopted by authorities often consist of manual investigations on a case by cases basis (FBI, 2018; Trading Standards UK, personal communication, February 26, 2019), which leaves many fraud instances unaddressed, as manual approaches fail to keep up with the ever-increasing amount of fraud complaints. A possible solution would be to automate the intensive manual labour involved using data science methods. Such approaches often translate into data extraction methods followed by massive data analyses (Kumar & Gunasekaran, 2019; Provost & Fawcett, 2013). A recurring strategy is to utilise supervised machine learning methods (Hernandez-Castro & Roberts, 2015; Sahingoz et al., 2019). Supervised machine learning models require training data from which the models can infer how data properties relate to the associated data labels (e.g., learning to classify phishing e-mails by their text style). Once a model has sufficiently learned how to differentiate the data into the labelled categories, the model can be deployed to categorize unlabelled data automatically (e.g., classify new incoming mail as (non)-phishing). In the case of online consumer fraud, the idea is seemingly simple: annotate a data set of product listings as fraudulent or legitimate, train a classifier based on those annotations and listing properties (e.g., image-, text-styles), and employ the trained classifier to detect suspicious listings on online shopping platforms (e.g., eBay, Amazon). Detected suspicious listings and the associated accounts could then be further investigated or taken down. Such an implemented system would help save financial resources and time by replacing or aiding manual searches and potentially detecting fraudsters quicker and before they can cause any harm. On the face of it, this strategy seems simple and effective. However, obtaining and annotating a data set in a meaningful way brings challenges. This chapter is a pilot study that examines the feasibility of creating a training dataset with annotated product listings as (non-)suspicious. Specifically, the chapter investigates the processes involved in labelling data, including the recruitment of suitable annotators, determining the reliability of the annotations, and assessing the usability of the obtained labels.

First, we determined where we could gather product listings that could be annotated. Many online shopping platforms such as Amazon, Facebook marketplaces, Alibaba, or eBay, could be considered for data collection. However, for this thesis, we decided to collect data from eBay because the platform is one of the most used in the world, has a more prominent presence in Europe than Alibaba or Facebook-marketplace, and provides more detailed information about the products and their sellers (eBay, 2020a, 2020b). In contrast to other platforms, eBay records the interactions between the seller and the buyer on several

dimensions (e.g., item description, communication, dispatch time, postage) through reviews and ratings after purchases have been made. In addition, the sellers can individualize their product listings by supplementing them with descriptions and photos, which are not as common elsewhere. The interactions and individualisations enrich the listings with details visible to all customers without the need for an account to view them. That means that the information is in the public domain, making it more suitable for automated collection. With the help of automated web crawlers implemented in coding languages such as Python or R, large amounts of data can be collected with relative ease.

After determining where and how to collect data, individuals who would label the listings as suspicious needed to be recruited. We defined an eBay listing as suspicious if any annotator perceived signs of fraud, such as non-deliveries, counterfeits, or other frauds leading to monetary loss. While anyone could annotate product listings as suspicious or non-suspicious, not everyone will have the required expertise. Thus, we asked experts who deal with online fraud on a regular basis to annotate the collected product listings. Specifically, we asked employees from Trading Standards UK and the Intellectual Property Office UK to annotate 250 eBay listings. In addition, we are also interested in how labels and annotation strategies differed from experts to non-experts. Thus, we also collected annotations from non-experts and compared agreements between all annotator groups.

3.1.1 Aims of this study

This pilot study investigates the feasibility of creating training data for a supervised machine-learning model to classify suspicious and non-suspicious online shopping listings. Examining the labelling process of a dataset is essential, as the labels are what a supervised model uses to learn how to perform a task better. Thus, the supervised model relies on well-labelled data to make reliable predictions (e.g., if an advertisement is suspicious of fraud). Therefore, we examine how well annotators agree with each other to determine the reliability of the labels created. Also, by comparing the annotation (strategies) within and between experts as well as non-experts, we aim to understand how they differ in identifying suspicious perceived features in online shopping listings. Understanding the annotation strategies might help us to better operationalise suspiciousness so that it could be leveraged for prediction tasks.

3.1.2 Related work

Research about fraud detection on e-commerce platforms mostly focuses on credit card, transactional, or auction fraud (Abdallah et al., 2016; Rodrigues et al., 2022). While most of the researched fraud types include fraudulent customers (i.e., buyers) that employ strategies to defraud sellers, only a few have examined frauds that are directed toward the consumer. Such frauds include fake advertisements, shops, reviews, auctions, or non-deliveries, which are tackled through various supervised and unsupervised machine learning methods (Elshaar & Sadaoui, 2020; Lai et al., 2023; Rodrigues et al., 2022; Weng et al., 2019).

For example, some approaches leveraged the HTML structure of websites to train supervised models to determine whether they are fake (e.g., phishing websites, fraudulent shops), as fraudsters seem to re-use code and content from genuine websites, creating detectable changes (Jain & Gupta, 2018; Xu et al., 2016). Similarly, others leverage the URL structure to train supervised models to detect fraudulent websites (Daeef et al., 2016; Sahingoz et al., 2019). Training data for such approaches are mostly provided by online shopping platforms (e.g., through blocklists) or websites that collect phishing complaints, such as openphish.com or phishtank.com. Others have utilized openly accessible buyer feedback (Weng et al., 2019) or transactional user behaviour provided by the online shopping platform Taobao (part of Alibaba) to train supervised models to detect fraudulent sellers or users (Weng et al., 2018; G. Zhang et al., 2022).

However, most studies that have focused on fraud detection, particularly on eBay, have examined fraudulent auctions, such as shill biddings (Abidi et al., 2021; Alzahrani & Sadaoui, 2018; Bauerly, 2009; Shah et al., 2003). Shill biddings are biddings without the intention to purchase and are intended to increase the auction price artificially. The seller can achieve such an increased auction price by creating multiple auction site accounts and making bids or recruiting accomplices who make repeated bids for the seller. Shell biddings are challenging to detect, and fraudsters are rarely caught (Dong et al., 2009; Ford et al., 2013), which drives the motivation to find possible detection solutions. However, online auctions on shopping platforms, such as eBay, are declining as customers prefer immediate purchases with fixed prices. Thus, online auctions are decreasing in popularity, and shill biddings have become less problematic for most products on eBay. Consequently, they are not examined here.

Other methods, such as unsupervised models (e.g., anomaly detection), have also been used to detect fraudulent sellers on eBay (Pandit et al., 2007). By generating a network of transactions between sellers and buyers, the authors of the study aimed to detect suspicious transactional configuration patterns. While this method is very fast and makes it suitable for large data sets, evaluating it is difficult as it is not clear exactly what constitutes a suspicious pattern, and no ground truth data was available to the study authors.

Some studies have used experts' and non-experts' annotations to train a supervised model to detect fraudulent sellers on the online shopping platform MercadoLivret⁷ (Almendra & Enachescu, 2011, 2012) as well as illegal ivory listings on eBay (Hernandez-Castro & Roberts, 2015). Based on manual inspections of a random sample of 455 neutral and negative feedback comments for MercadiLivret listings, Almendra & Enachescu (2011) annotated the comments with labels such as non-delivery, platform investigation, no response, or out of stock. The labelled comments were then used to train a classifier, using *n*-grams⁸ that would predict the labels of all unlabelled comments (over 4 million). Another classifier, trained on the features

⁷ A Brazilian auction site (mercadolivre.com.br)

⁸ *N*-grams describes the process of splitting a document (e.g., reviews, comments) into a sequence of *n* words. For example, "I walked in the park" into "I walked", "walked in", "in the", "the park", which are bigrams. *N*-grams can also be longer sequences of three (trigrams), or more.

derived from the comments (e.g., number of comments labelled as non-deliver, etc.), was then used to predict known suspended sellers (as a proxy for fraud). Applying the trained seller classifier, the authors predicted 137 from 1,252 unlabelled sellers as fraudulent. Although the authors acknowledge that no ground truth data was available, and predictions were not verifiable, such an annotation process might approximate suspiciousness. In their study, Hernandez-Castro & Roberts (2015) recruited two experts⁹ that annotated 1,159 eBay listings of advertised ivory as potentially legal or illegal (elephant ivory). The data was collected over eight weeks from the eBay UK Antiques section. By training a CN2 classifier (a supervised method) that generates a series of if-then rules on the selected features (metadata of the listings), they achieved a 93% prediction accuracy. To make the rules more generalizable to other platforms, the metadata included most information from the listings, but excluded the images and descriptions. As in previous research, no ground truth data was available that could be used to verify model performance. However, through the if-then rules, the classifier is interpretable and might support faster decision making in the future. Next to attributes such as postage price, merchant feedback, item prices, or number of reviews, the number of bids were also found to be important in determining illegality.

3.2 Method

3.2.1 Automated data collection

To collect product and seller information from eBay, we employed a custom web scraper built in the programming language Python, utilizing the package Beautiful Soup (Richardson, 2019). The entire scraper is divided into four modules embedded in an error-handling script (Figure 3.1). The following steps describe the scraping procedure.

First, all available eBay categories were scraped from "<https://www.ebay.co.uk/n/all-categories>", which were then filtered to include categories found to have a higher prevalence of fraud in international publications (OECD/EUIPO, 2019). These included 191 categories associated with clothing, shoes, bags, electronics, jewellery, and watches (see Appendix B1 for the full list).

Second, a list of proxies from "<https://free-proxy-list.net>" was automatically collected and used during the scraping process to avoid blocking from eBay. Blocking can quickly occur when using the same IP address and without limiting the speed of the scraper. Thus, the proxy was changed after each failed connection for ten attempted connections before skipping the current address call. The user agent (web browser type and version) was also changed for each proxy change. Lastly, a randomly generated pause of 1 to 3 seconds was implemented before each address call.

Third, all product and seller URLs were collected from the first page of each category, including 96 eBay listings (e.g., Mens-Casual-Shirts-Tops, Mens-Coats-Jackets, or Mens-Formal-Shirts).

⁹ Former law enforcement officers with specialist knowledge in illegal wildlife trade

The URLs were used for two separate scrapes that collected detailed product and seller information. In addition, reviews about each product were collected if they were available. The information about the product and the associated seller was stored in a .txt file named after the product category and product number.

Lastly, the individual product information, if provided by the seller, was collected. The individualized information usually contains descriptions and images of the product. Occasionally, this description is replaced by an individual webpage (e.g., the seller’s store) embedded in the eBay page. In such cases, the information was omitted.

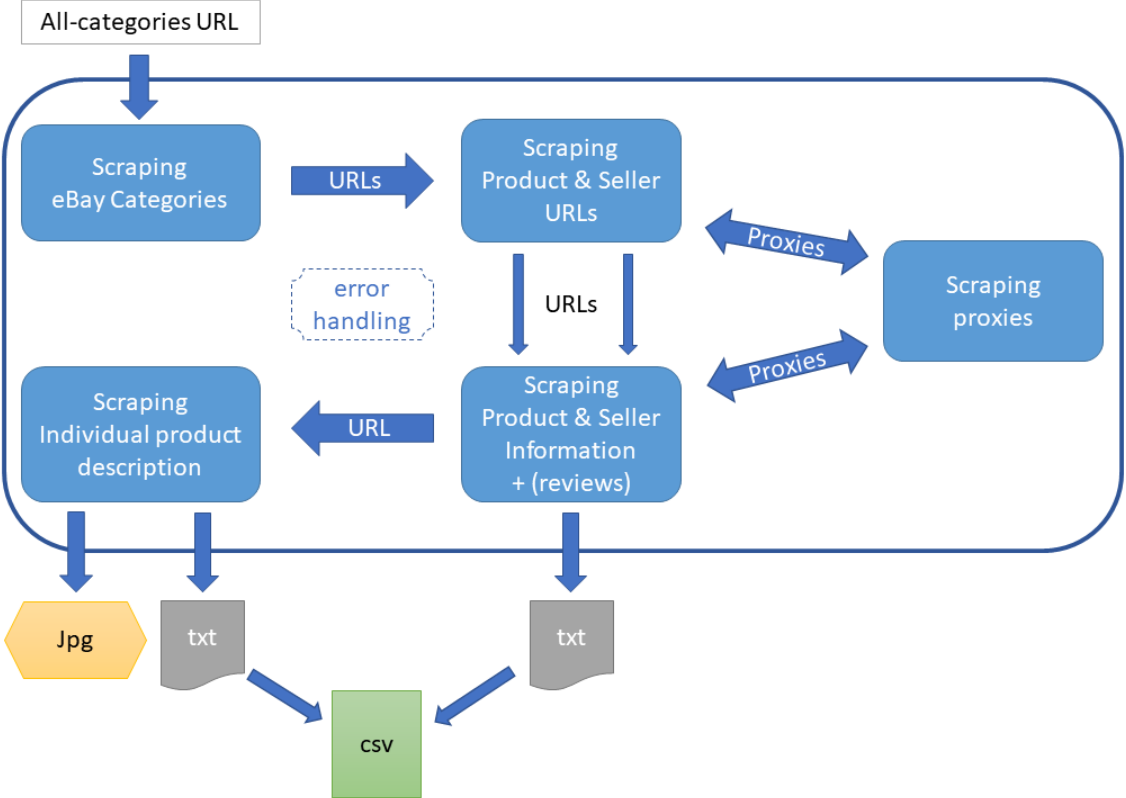


Figure 3.1. Scraper framework for eBay listings.

The scraped eBay listings were filtered, excluding duplicate listings, and grouped by domain (Shoes, Clothes, and Electronics), resulting in 7,943 listings (Table 3.1). The listings differ considerably between domains due to the uneven number of eBay categories.

	Electronics	Clothes	Shoes	Jewellery	Watches	Bags
Listings	4281	2223	904	420	71	44

Table 3.1 Distribution of eBay listings.

3.2.2 Annotating eBay listings

To utilize supervised machine learning methods, an annotated dataset is needed that can be used to train and validate a model. Thus, we describe here how we obtained the labels. The study was approved by the UCL Research Ethics Committee.

3.2.3 Participants

We recruited three domain experts and 18 non-experts to annotate a subset of the eBay data as suspicious and non-suspicious. The expert group consisted of two individuals from the Intellectual Property Office UK (IPO) and one from Trading Standards UK (TS). They were considered experts as they are involved in online consumer fraud investigations in their daily work. The non-expert group was recruited from the Security and Crime Science Department at University College London. The non-experts were unfamiliar with online fraud research.

3.2.4 Materials

We randomly selected a subset of 250 eBay listings (product and seller page) from the shoe category intended to be annotated by experts and non-experts. We chose the shoe category as it was found to have a higher prevalence of fraud (OECD/EUIPO, 2019). The listings were provided as HTML files, a copy of the original eBay product, and seller-page that could be viewed through any web browser to imitate a real scenario on eBay. Due to privacy concerns, the seller's name (ID) and address were masked by replacing them with "AnonymousName". In addition, all links which could have forwarded the participants to an online page and possibly revealed the seller were disabled. Figure 3.2 shows an example of how participants would have seen a product listing.

For each listing, the participants were asked to complete a questionnaire through Google forms containing the following three items: "How suspicious is the advert?" (This included any form of fraud: counterfeit, no product shipment after purchase, etc.), "How confident are you in your judgment?", and "Give reasons for your confidence (What made the listing suspicious/trustworthy?)". The first two items were answered on an 11-point scale from 0 (not suspicious/confident) to 10 (very suspicious/confident). The last item was responded to with free text to collect information on their reasoning and beliefs about the appearance of any suspicious listings, thus, allowing us to examine the participants' annotation strategies.

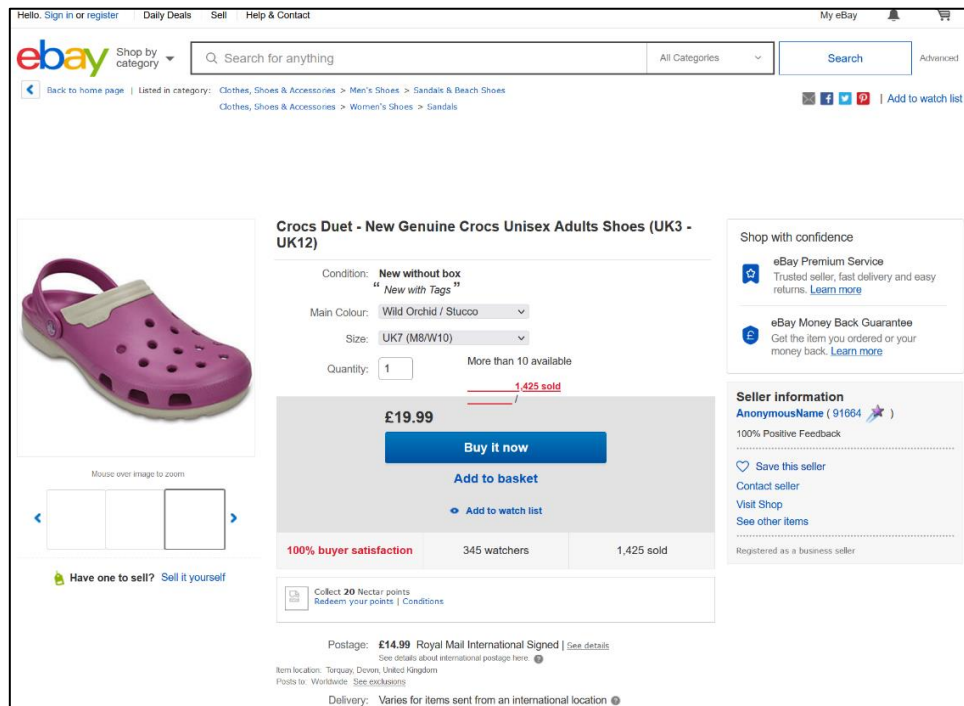


Figure 3.2. Example of one eBay listing participants were presented with.

3.2.5 Procedure

All participants were sent an e-mail with information and informed consent sheets, which were read, signed, and sent back by e-mail before the annotation task. Participants were then sent instructions by e-mail to annotate the listings, which were shared through a Google Drive folder, and answers (i.e., labels) were recorded through Google forms. The annotation task was completed with a computer at the participants' work offices or homes. Experts and non-experts were given two sets of the same listings, ensuring that each listing was annotated twice by experts and non-experts, allowing for a comparison of agreement within and between (non-)experts. Each set of listings was randomly distributed between participants. Non-experts annotated around 28 listings each. Since fewer experts could be recruited to annotate, the TS expert was asked to annotate the full set of 250 listings, while the IPO experts were asked to annotate 125 listings each. Table 3.2 shows the distribution of listings per annotator group. 89 invalid responses (e.g., missing values, unidentifiable ID, a judgment of just product or seller page) were removed, resulting in an overall reduction of annotations in each group. Due to other work duties, the IPO experts were not able to annotate all listings. Thus, 31 listings were annotated twice by experts (TS and IPO), 170 listings were annotated twice by non-experts, and from those listings, 31 were annotated four times by both expert groups and by both non-expert groups.

	Annotators			
	Experts		Non-Experts	
Individuals	1 (TS)	2 (IPO)	9	9
Annotations	248	31	233	180

Table 3.2. Distribution of Annotations for groups and individuals.

3.3 Results

3.3.1 Suspiciousness and Confidence ratings

Figure 3.3 shows the distribution of participants' ratings of how suspicious the items were. The distributions for experts and non-experts seem to follow a similar trend, with most listings being labelled as not being suspicious. Both annotator groups labelled around 2.5-5% of all listings as highly suspicious (ratings 9 and 10).

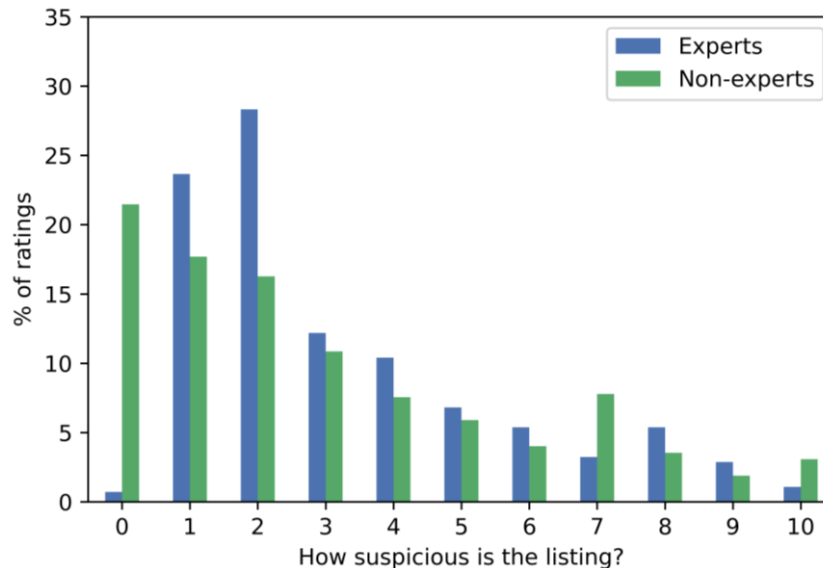


Figure 3.3. Percentage distribution of suspiciousness ratings of experts and non-experts; Ratings range from 0 (Not suspicious) to 10 (Very suspicious).

In contrast, participants differed in the confidence level of their annotations. Experts tended to express lower confidence in their ratings than non-experts did (Figure 3.4).

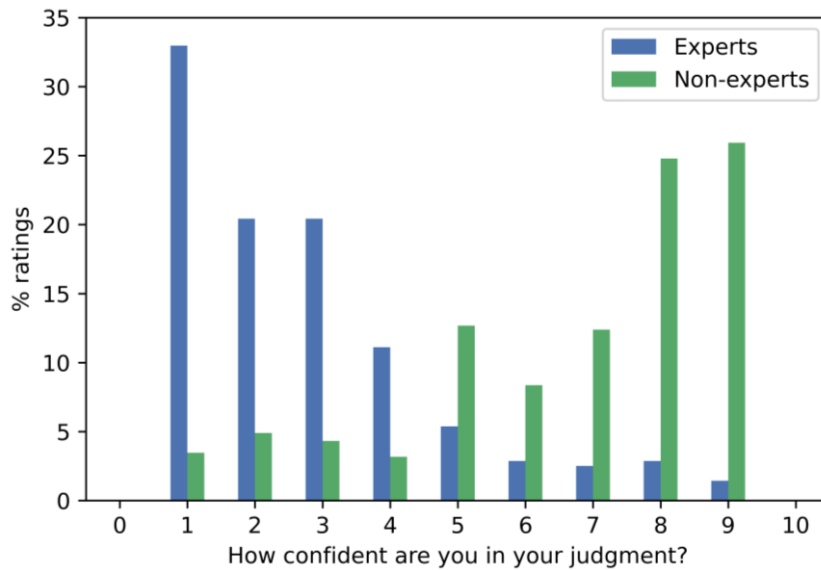


Figure 3.4. Percentage distribution of confidence ratings of experts and non-experts; Ratings range from 0 (not confident) to 10 (very confident).

3.3.2 Annotation agreement

To assess how well annotators agreed with each other in annotating the suspiciousness of listings, we calculated Krippendorff’s alpha values (Table 3.3), which indicates the degree of agreement. We performed the analyses three times, for experts and non-experts individually to assess within group agreement for the two groups, and experts and non-experts together, to assess between group agreement. Krippendorff’s method can account for missing values, which is important since not all annotators annotated all listings within and across groups (experts and non-experts). The alpha value ranges from -1 (complete disagreement) to 1 (complete agreement), while a value of 0 indicates no agreement (A. F. Hayes & Krippendorff, 2007; Krippendorff, 1970). 95% confidence intervals were added to each alpha estimation using a bootstrap procedure with 1,000 iterations. All calculations were performed in the programming language R with the package “krippendorffsalpha” (Hughes, 2022).

	Experts + Non-experts	Experts (IPO + TS)	Non-experts
Krippendorff’s α	-0.04 [-0.02,0.10]	-0.07 [-0.48, 0.23]	0.10 [-0.03, 0.22]
Comparison	Between groups	Within group	

Table 3.3. Krippendorff’s alpha within and between annotator groups (Experts and Non-Experts); 95% confidence interval in brackets.

Based on recommendations of a minimum acceptable agreement value of $\alpha = 0.67$ or higher (A. F. Hayes & Krippendorff, 2007), we cannot assume acceptable agreements between annotators.

3.3.3 Annotation strategies

Since annotators disagreed on how suspicious the eBay listings were, we examined their reasoning for determining the level of suspiciousness. For each answer to the last question of our survey, “Give reasons for your confidence (What made the listing suspicious/trustworthy?)”, we qualitatively assigned topics based on their arguments through a thematic analyses (Ibrahim, 2012; Joffe, 2011). For example, when annotators mentioned that the number of negative reviews led them to their suspiciousness rating, we assigned the topic “buyer feedback”. In most cases, answers included several arguments for which we then assigned multiple topics. Based on the provided arguments, 14 topics emerged: *buyer feedback* (reviews, ratings), *product description* (e.g., formatting, usage of words), *price* (product or postage), *overall look* (e.g., how professional the listings seemed), *product location*, *writing style* (weird letter usage, overuse of special characters), *product images* (e.g., quality, quantity), *product quantity* (availability or sold), *listing details* (e.g., mentioning VAT number, missing terms and conditions, company number, seller contact info), *unbranded*, *product variation* (e.g., size, colour), *seller membership time*, *transaction methods*, and *no reason*. Figure 3.5 shows the relative frequency (i.e., percentage) with which each theme was discussed by experts and non-experts.

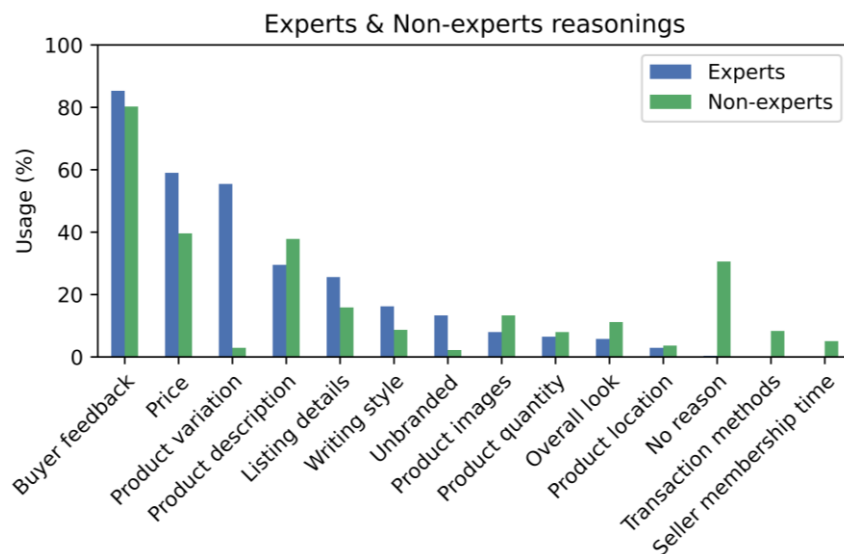


Figure 3.5. Distribution of reasoning topics derived from the annotators' answers during the labelling process; Topic usage is expressed in percentage (based on the number of annotations); Multiple topics can be present in each answer.

Experts and non-experts seemed to use similar reasoning for how they reached their suspiciousness ratings, both often referred to the buyer feedback, such as ratings and reviews, to determine the suspiciousness of the listings. Similarly, the price also appeared to be important, with too low prices often considered suspicious. However, experts seemed to focus more on product variation in the seller offers than the non-experts did. Interestingly, non-experts also appeared to have considered the transaction methods and the time the seller was a platform member as useful indicators. Long membership and various payment methods were considered indicators of non-fraudulent activity in most cases. Examining the

annotators' reasonings also showed that the absence of any negative feedback was often considered suspicious, and “organic” or “well distributed” feedback (including positive and negative) were considered non-suspicious.

To further inspect if experts and non-experts focused on different listing properties for different suspiciousness ratings, we aggregated the reasoning topics by suspiciousness, split into low (ratings 0-2), medium-high (ratings 3-7), and very high (ratings 8-10) suspiciousness (Figure 3.6).

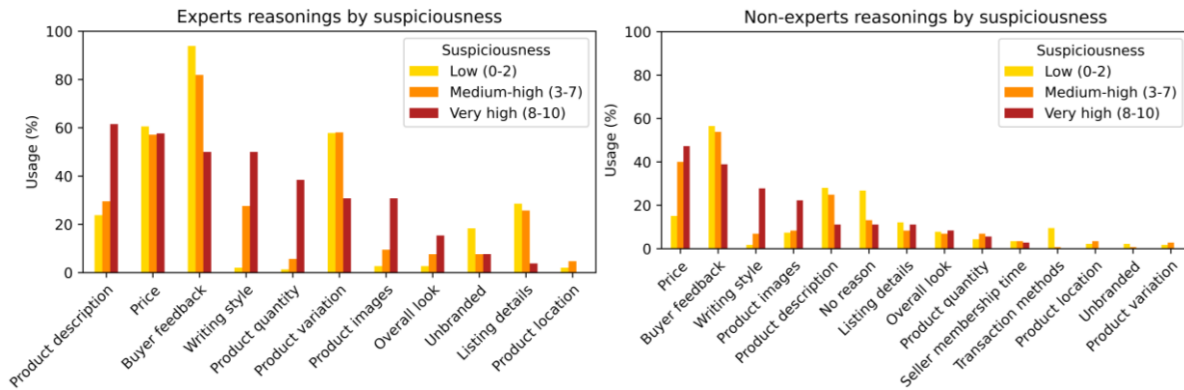


Figure 3.6. Experts (left) and non-experts (right) reasoning topics, split by suspiciousness: low (0-2), medium-high (3-7), and very high (8-10).

Experts seem to focus more on the descriptions, writing style, product quantity, and images when considering a listing as highly suspicious. High quantity and availability were considered suspicious, but so too was a strong discrepancy between availability and sold products. Descriptions overemphasizing item locations or stressing words such as “legit” or “genuine” were also regarded as signs of suspiciousness. Similarly, non-experts seemed to focus more on the writing style and product images to determine high suspiciousness. For example, low-quality images and the overuse of special characters were considered suspicious.

3.4 Discussion

The aim of this work was to test the feasibility of building a reliable dataset that can be used for training a supervised machine learning model to detect suspicious online shopping listings. We automatically collected eBay listings from product categories previously found to be most affected by fraud (OECD/EUIPO, 2019). To generate a training dataset, we required labelled data, which we aimed to generate with the help of fraud experts. By comparing the annotations from experts and non-experts, we investigated how annotators determine the suspiciousness of product listings. Building reliable training data seemed difficult due to the low agreements between annotators, making the use of machine learning models unfeasible. Below we will discuss the annotators' labelling strategies and the difficulties surrounding the labelling process, as well as give recommendations for future annotation tasks in the fraud domain.

3.4.1 Difficulties accompanying the annotation process

We encountered several issues while generating the labelled dataset, which touched on broader annotation issues. We recruited three experts, which limited the number of possible annotations and within-group comparisons. Ideally, more experts would have been recruited. However, the annotation process is very time-consuming, and consequently, tasking many employees with the labelling task is costly. Therefore, our collaborators (Trading Standards, Intellectual Property Office) had difficulties assigning more employees to annotate the eBay listings. Of those who did participate, the annotators had great difficulties agreeing on what a suspicious listing looked like.

It is possible that the definition of a suspicious listing might have been too vague and not conveyed clearly, leading to various ideas on suspiciousness amongst the annotators. We wanted to capture all types of online consumer fraud under the umbrella of “suspiciousness”, which may have been too broad to capture any specific knowledge individuals might have about specific online fraud types. For example, some individuals might know what burner accounts (i.e., hacked accounts offering and selling products well under their value, which are never shipped) look like but have limited knowledge about other fraud types and how they manifest in product listings. Thus, their insight on burner accounts might have been left unleveraged and not properly captured by asking about general suspiciousness. Therefore, experts might have been artificially pushed out of their specific expertise.

The results of this pilot study and discussed annotation issues above highlight that no clear criteria exist for identifying suspicious listings. The lack of such criteria also points toward a more complex problem that humans might not be able to grasp potential indicators of suspiciousness. The information load listings provide, the constant change in the online environment (e.g., how advertisements are presented), and the shift in fraud strategies make it difficult for humans to keep up with their knowledge. Thus, it is possible that individual judgments on suspiciousness might not be reliable enough to generate training datasets that can be used to make predictions about whether online listings are suspicious or fraudulent. Similarly, inconsistent judgments of suspiciousness could also be problematic for manual inspections, as it would possibly make the procedures on who to investigate unreliable.

A possible confound to the annotation task might also be the representativeness of the provided data. Although shoes were selected because previous findings showed that the category is affected by fraudulent listings (e.g., counterfeits) more than other categories (OECD/EUIPO, 2019), we do not know true underlying distribution of fraudulent listings.

3.4.2 Annotation strategies

Based on the qualitative assessment of the annotators' reasons for their labelling choices, we can see clear indications that individuals have different ideas of what a suspicious listing is. However, buyer feedback, listed prices, and product descriptions were often central in the annotator's evaluation, although for various reasons. For example, annotators' understanding

of a weird description based on the use of specific words appeared to differ. Some reasoning examples¹⁰ are stated below:

- *"I'm always sceptical of items listed that state "UK Seller" in larger writing on the advert"*
- *"Over emphasis on the fact they are "Genuine", size chart reminds me of things I have seen from suspected counterfeit sellers. The fact it states to go a size higher than your normal UK size, weird letters and icons in product description"*
- *"It says it's a 'big name brand' but doesn't state which"*
- *"No negative reviews, states " New, 100% Authentic " with various sizes??, unable to find product, new without box"*

Some expert annotators remarked that the seller and store information, which was excluded due to privacy concerns, would have influenced their evaluation. For example, investigators would perform a web search of the indicated store or seller, such as validating the physical location provided by eBay listings. In some cases, Google Street View would be used to assess the provided address quickly. Furthermore, investigators would cross-reference local or federal business registers with the seller of the eBay listings. Thus, with the seller information missing, some experts raised the concern that they could not adequately judge the listings' suspiciousness, which might have contributed to the overall low confidence ratings (Figure 3.4).

Lastly, annotators' time spent on labelling the eBay listings seemed to vary considerably. Non-experts seemed to make relatively quick judgments, using only a minute or two per listing. However, some expert annotators reportedly spent 6-8 minutes for each listing.

3.4.3 Recommendations for future annotations

Based on the issues faced here, we can make several recommendations for future studies aiming to label suspicious product listings. First, we can make suggestions on how to improve the annotation procedure, which are summarized in Table 3.4. Second, future studies could reassess who would be suitable as expert annotators. For example, brand experts could be asked to label suspiciousness (e.g., if the product is a counterfeit) of the products they are familiar with. Thus, their expertise could be leveraged more efficiently, which would likely result in higher confidence associated with labels. The annotation strategies could also be further examined to uncover how annotators determine the labels (i.e., which product features annotators examine). Although such product details are challenging to incorporate as features for supervised models, more reliable labels could help reveal other suspicious patterns within seller behaviours or the presentations of online listings (e.g., specific writing style of product descriptions).

¹⁰ All spelling and grammatical errors have been transcribed from the survey.

Problem	Consequence	Recommendation
Ill-defined labels (e.g., suspiciousness).	Annotators label unreliably.	Clearly define and convey annotation construct (i.e., label meaning).
Ill-defined labelling process.	Low validity between annotator agreement assessment (e.g., due to unequal time spent on labelling).	Provide a well-controlled annotation environment (e.g., through clear instructions or an annotation interface), ensuring participants conduct the task in the same manner.
Missing information needed for labelling (e.g., seller information).	Annotators label unreliably and have low confidence.	If possible, provide all available information to annotators (might not be possible due to ethical considerations)
The labelling task is outside the annotator's expertise.	Expert knowledge remains unleveraged, and derived labels are less reliable.	Survey annotators on their expertise to tailor the labelling task. (e.g., experts are only asked to annotate a specific fraud type they are familiar with).
The annotator's decision process is not recorded.	No conclusions are possible on how annotation decisions are made.	Survey experts about their decision and reasoning on how they derive the labels. Ask experts to comment on their decision-making process. Ask experts to rank the importance of decision-relevant aspects.

Table 3.4. Recommendations for future annotations.

3.5 Conclusion

The current work suggests that labelling suspicious online product listings is difficult. While we identified potential reasons for the current issues, they are hard to pinpoint. As the recommendations summarize, future studies have room to improve on several fronts. Labelling data could also be replaced with ground truth data, instances from which we know fraud has occurred (i.e., historical data), to reliably inform a supervised model. However, such data are often unavailable or might become outdated due to changes in the online environment or fraud strategies. Furthermore, since only detected fraud instances are recorded in historical data, undetected fraud might introduce biases that can affect machine learning performances, which will be discussed in the following chapter.

The problem of reliable annotations or ground truth data is not easily addressed. A possible strategy to address this might lie in a closer collaboration between authorities and researchers by better understanding what circumstances need to be satisfied to warrant an investigation and how to translate them into measurable parameters. Such collaborations could also facilitate more reliable annotation tasks that could benefit practitioners in the future.

Chapter 4: Confounds and Overestimations in Fake Review Detection: Experimentally Controlling for Product-Ownership and Data-Origin

This chapter is based on the following publication:

- Soldner, F., Kleinberg, B., & Johnson, S. D. (2022). Confounds and overestimations in fake review detection: Experimentally controlling for product-ownership and data-origin. *Plos one*, 17(12), e0277869. <https://doi.org/10.1371/journal.pone.0277869>

4.1 Introduction

Online shopping is not new, but it is increasing in popularity as seen by the growth of companies such as Amazon and e-Bay (Palmer, 2020; Soper, 2021; Weise, 2020). Previous work shows that consumers rely heavily on product reviews posted by other people to guide their purchasing decisions (M. Anderson & Magruder, 2012; Chevalier & Mayzlin, 2006; Watson, 2018). While sensible, this has created the opportunity and market for deceptive reviews, which are currently among the most critical problems faced by online shopping platforms and those who use them (Dwoskin & Timberg, 2018; Nguyen, 2018). Research suggests that for a range of deception detection tasks (e.g. identifying written or verbal lies about an individual's experience, biographical facts, or any non-personal events), humans typically perform at the chance level (DePaulo et al., 2003; Kleinberg & Verschuere, 2021). Furthermore, in the context of considering online reviews, the sheer volume of reviews (Woolf, 2014) makes the task of deception detection implausible for all but the most diligent consumers. With this in mind, the research effort has shifted towards the use and calibration of automated approaches. For written reviews, which are the focus of this article, such approaches typically rely on text mining and supervised machine learning algorithms (Newman et al., 2003; Ott et al., 2011, 2013; Pérez-Rosas et al., 2018). However, while the general approach is consistent, classification performance varies greatly between studies, as do the approaches to constructing the datasets used. Higher rates of performance are usually found in studies for which the review dataset (Mohawesh et al., 2021; Nagi Alsubari et al., 2022; Ren & Ji, 2019; Santos et al., 2020; Shojaee et al., 2013; Singh & Chatterjee, 2022) is constructed from several different sources, namely a crowdsourcing platform and an online review platform (Ott et al., 2011, 2013). High classification performances are also found in studies using data scraped or donated from a single review platform, such as Yelp or Amazon (Barbado et al., 2019; Fazzolari et al., 2021; Ren & Ji, 2019; D. Zhang et al., 2016). Lower rates of performance are typically found in studies for which data is extracted from a single source and for which greater experimental control is exercised (Kleinberg & Verschuere, 2021; Mihalcea & Strapparava, 2009; Perez-Rosas & Mihalcea, 2014; Pérez-Rosas & Mihalcea, 2015). Why we can observe such strong differences of classification performances between studies is unclear. However, such findings suggest that confounds associated with the construction of

datasets may explain some of the variation in classification performance between studies and highlights the need for the exploration of such issues. In the current study, we will explore two possible confounds and estimate their effects on automated classification performance. In what follows, we first identify and explain the two confounds. Next, we provide an outline of how we control for them through a highly controlled data collection procedure. Lastly, we run six analyses on subsets of the data to demonstrate the pure and combined effects of the confounds in automated veracity classification tasks.

4.1.1 Confounding factors

In an experiment, confounding variables can lead to an omitted variable bias, in which the omitted variables affect the dependent variable, and the effects are falsely attributed to the independent variables(s). In the case of the detection of fake reviews, two potential confounds might explain why some studies report higher and possibly overestimated automated classification performances than others. The first concerns the provenance of some of the data used. For example, deceptive reviews are often collected from participants recruited through crowdsourcing platforms, while “truthful” reviews are scraped from online platforms (Ott et al., 2011, 2013), such as TripAdvisor, Amazon, Trustpilot, or Yelp. Creating datasets in this way is efficient but introduces a potential confound. That is, not only do the reviews differ in veracity but also their *origin*. If origin and veracity were counterbalanced so that half of the fake (and genuine) reviews were generated using each source, this would be unproblematic but unfortunately in some existing studies, the two are confounded. A second potential confound concerns ownership. In existing studies, participants who write fake reviews are asked to write about products (or services) that they do not own. In contrast, in the case of the scraped reviews – assuming that they are genuine (which is also a problematic assumption) – these will be written by those who own the products (or have used the services). As such, ownership and review veracity (fake or genuine) will also be confounded.

Besides these two confounds, it is worth noting that some of the studies that have examined fake review detection have used scraped data that does not have “ground truth” labels (Barbado et al., 2019; Fazzolari et al., 2021; Mukherjee et al., 2013b; Rahman et al., 2015; Ren & Ji, 2019; D. Zhang et al., 2016). That is, they have used data for reviews for which the veracity of the content is not known but is instead inferred through either hand-crafted rules (e.g., labelling a review as fake when multiple “elite” reviewers argue it is fake) or inferred by the platforms own filtering system, which is non-transparent (e.g., Yelp). While utilizing such data to investigate how platforms filter reviews is helpful, studying the effects of deception without ground truth labels is problematic because any found class-specific properties cannot be reliably attributed to the class label. Furthermore, any efforts to improve automated deception detection methods will be limited by algorithmically filtered data, because classification performances cannot exceed the preceding filter. Thus, ground truth labels are imperative for investigating deception detection in supervised approaches, and such labels can be obtained through experimental study designs. However, previous studies collecting

data experimentally (Ott et al., 2011, 2013) suffer from the confounds mentioned above, and an altered study design is required.

4.1.2 Confounds in fake review detection

Studies of possible confounding factors in deception detection tasks that involve reviews are scarce. In their study, Salvetti et al. (2016) investigated whether a machine learning classifier could disentangle the effects of two different types of deception – lies vs. fabrications. In the case of the former, participants recruited using Amazon’s Mechanical Turk (AMT) were asked to write a truthful and deceptive review about an electronic product or a hotel they knew. In the case of the latter, a second group of AMT participants was asked to write deceptive reviews about the same products or hotels. However, this time they were required to do this for products or hotels they had no knowledge of, resulting in entirely fabricated reviews. Salvetti et al. (2016) found that the classifier was able to differentiate between truthful reviews and fabricated ones but not particularly well. However, it could not differentiate between truthful reviews and lies – classification performance was around the chance level. These findings suggest that product ownership (measured here in terms of fabrications vs truthful reviews) is a potentially important factor in deceptive review detection.

A different study examined the ability of a classifier to differentiate truthful and deceptive reviews from Amazon (Fornaciari et al., 2020) using the “DeRev” dataset (Fornaciari & Poesio, 2014). The dataset contains fake Amazon book reviews that were identified through investigative journalism (Flood, 2012; Streitfeld, 2011). Truthful reviews were selected from Amazon about other books from famous authors, such as Arthur Conan Doyle, Rudyard Kipling, Ken Follett, or Stephen King, for which it was assumed that it would not make sense for someone to write fake reviews about them. A second corpus of fake reviews – written about the same books – was then generated by participants recruited through crowdsourcing to provide a comparison with the “DeRev” reviews. The researchers then compared the performance of a machine learning classifier in distinguishing between different classes of reviews (e.g., crowdsourced-fake vs. Amazon-fake, crowdsourced-fake vs. Amazon-truthful). Most pertinent here was the finding that the study authors found that the crowdsourced-fake reviews differed from the Amazon-fake reviews. Both studies (Fornaciari et al., 2020; Salvetti et al., 2016) hint at the problems of confounding factors in deception detection tasks. Although Fornaciari et al. (2020) uses a well-designed setup to test hypotheses, book reviews were not always about the same books between classes, introducing a potential content related confound. Similarly, the machine learning classifiers used were not always cross-validated with the same data type (i.e., the training and testing data were sourced from different data subsets), complicating the interpretation of the results. In contrast, in the current study, we match product types, hold the cross-validating procedure constant across all data subsets, and extend the analyses to positive and negative reviews.

4.1.3 Aims of this study

With this study, we want to examine how confounds in datasets can affect the training of supervised machine learning models, which could lead to wrongful conclusions about their performance and the interpretation of which data features are important for the prediction task. Confounding variables have the potential to distort the findings of studies, leading researchers to conclude that a classifier can distinguish between truthful and deceptive reviews when, in reality, it is actually leveraging additional characteristics of the data, such as the way in which it was generated. Such confounds would mean that the real-world value of the research is limited (at best). In the current study, we employ an experimental approach to systematically manipulate these possible confounders and to measure their effects for reviews of smartphones. Specifically, we estimate the effect of *product-ownership* by collecting truthful and deceptive reviews from participants who do and do not own the products they were asked to review. To examine the effect of data-origin, we also use data (for the same products) scraped from an online shopping platform. We first examine how well reviews can be differentiated by *veracity* alone (i.e., without confounds), and whether classification performance changes when this is confounded with *product-ownership*, *data-origin*, or both. If *ownership* or *data-origin* do influence review content (we hypothesize that they do), reviews should be easier to differentiate when either of the two confounds is present in veracity classification, but reviews should be most easily classifiable if both confounds (*ownership*, *data-origin*) are present at the same time. Thus, our experiments allow us to assess how well a classifier can differentiate reviews based on veracity alone, and how much the confounds discussed above influence detection performances.

4.2 Data collection

4.2.1 Participants

Data were collected with Qualtrics forms (www.qualtrics.com) from participants recruited using the academic research crowd-sourcing platform Prolific (www.prolific.co). Since we wanted to collect reviews about smartphones, we wanted to make sure that all participants owned a smartphone they could write about. We achieved this by using a pre-screener question to limit the participant pool. In this case, only prolific users who use a mobile phone on a near-daily basis could take part. 1169 participants (male = 62.19%, female = 37.13%, prefer not to say = 0.007%) ranging between 18 and 65 years of age ($M = 24.96$, $SD = 7.33$) wrote reviews. Participants completed the task with a median time of 10 min. and were paid for that time with 0.79 GBP. The study was reviewed by the ethics committee of the UCL Department of Security and Crime Science and was exempted from requiring approval by the central UCL Research Ethics Committee. Participants provided written informed consent online by clicking all consent statement boxes affirming their consent before taking part in the study.

4.2.2 Experimental manipulation

We collected 1,600 reviews from participants who owned the products (both truthful and deceptive reviews), and 800 from those who did not (deceptive review only). For the former, these were organized to generate 400 positive and 400 negative reviews for each of the factor (positive vs negative, deceptive vs truthful) combinations (i.e., positive-deceptive, negative-deceptive, positive-truthful, negative-truthful). For the latter, participants could only write deceptive reviews, and we collected 400 positive and negative of each. Reviews from owners and non-owners were collected using two Qualtrics survey forms. For both, participants were introduced to the task and asked to provide informed consent. They were then asked to indicate the current and previous brands of phones that they owned. Participants selected all applicable brands from ten choices (Samsung, Apple, Huawei, LG, Motorola, Xiaomi, OnePlus, Google, Oppo, Other) without selecting a specific phone. The brands were selected based on the top selling smartphones by unit sales and market shares in 2018 and 2019 within Europe (CNET, 2020; Counterpoint Reserach, 2020; Hong, 2020; Mishra, 2020).

4.2.2.1 Smartphone owners

For survey 1, participants were asked which phone they liked and disliked the most from their selected brands (Figure 4.1). The questions were presented in a randomized order, and participants had to rate their phones on a 5-point scale, replicating the Amazon product rating scale (1 star = very bad; 2 stars = bad; 3 stars = neutral; 4 stars = good; 5 stars = very good). Subsequently, each participant was randomly allocated to either write a truthful or deceptive review about their most liked and their most disliked phone. The truthful review corresponded to their given phone rating. For deceptive reviews, participants were asked to write reviews that were the polar opposite of the rating they had provided. For example, for a smartphone they liked the most (or least), they were asked to write a 1- or 2-star (or 4- or 5- star) rating for that smartphone. The exact ratings (1 or 2 for a negative review, and 4 or 5 for a positive review) used for each condition (truthful or fake) were also randomized. Participants were presented with an attention check after writing each review, by asking what type of review they were instructed to write (truthful or deceptive).

4.2.2.2 Smartphone non-owners

For survey 2 (Figure 4.1), participants were instructed to write a negative (1- or 2-star) review as well as a positive (4- or 5-star) review about two separate phones they did not own. The two randomly selected phones were selected from a list of 60 phones from the top sold phones by the top brands previously established. The brands of both phones were randomly selected from those participants who had indicated not owning them. Asking them to write about brands they did not and had not owned meant that participants could not use personal knowledge about that brand while writing the reviews. The allocation of reviews to negative and positive conditions was counterbalanced using a random number generator. Participants were allowed to perform an online search of the smartphone. Since the shortest Amazon reviews for electronic products contain around 50 characters, but most range between 100 to

150 characters (Woolf, 2014), the minimum length of reviews participants had to write was set to 50 characters. To prevent participants from using existing reviews found online, they were prevented from being able to copy-and-paste text into the text field in which they were required to provide their review. Also, participants were presented with an attention check in both ownership conditions after writing each review, by asking what type of review they were instructed to write (truthful or deceptive).

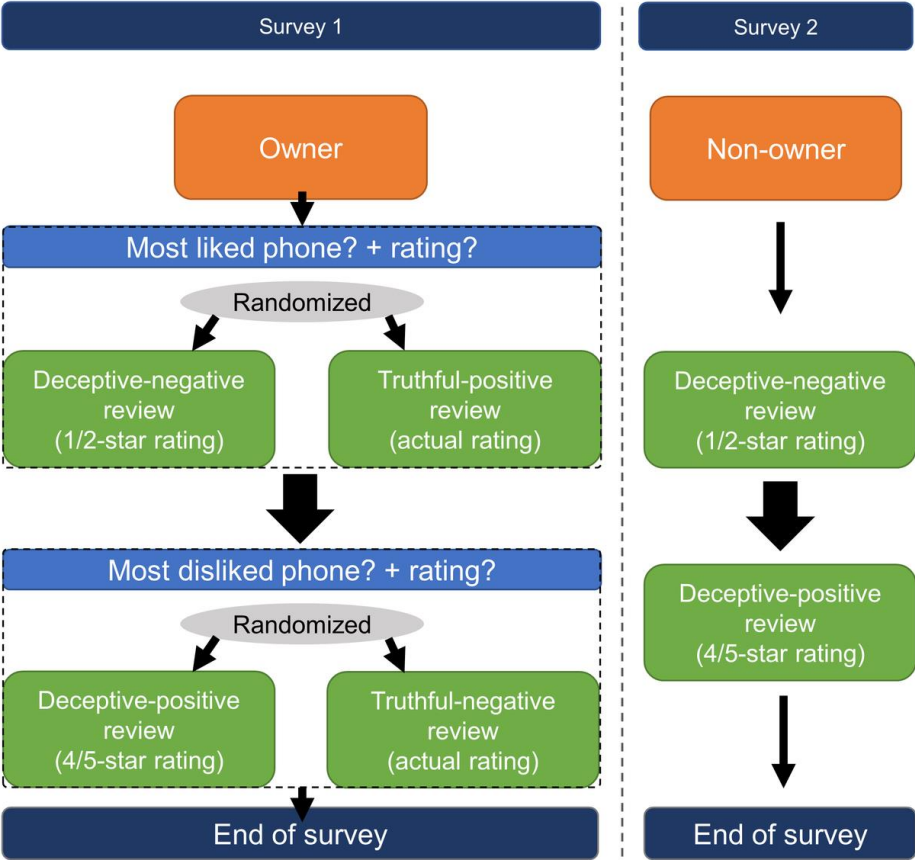


Figure 4.1. Collecting procedure of reviews from Prolific participants.

4.2.2.3 Amazon reviews

To obtain reviews that differed in *origin*, we collected Amazon reviews for the same phones that participants had written about. To do this, a list of all the smartphones reviewed (by owners and non-owners) was created, and product links were manually created for all of those that were available on Amazon and had received reviews. We used the “selectorlib” Python package (Rajeev, 2019) to collect all reviews from each product link. To reduce the likelihood of collecting fake reviews, only those for which there was a verified purchase were used. The collection procedure adhered to the Amazons terms and conditions.

4.2.3 Final dataset

4.2.3.1 Data filtering

Crowdsourced reviews were excluded if participants failed the attention check (see above) or reviews were not written in English, the latter tested using the Python “langdetect” package (Danilák, 2014/2021). Reviews were also removed if they did not follow the instructions. To detect the latter, we obtained the sentiment for each review using the “TextBlob” Python package (Loria, 2013/2021). All reviews that had a 4 or 5-star rating but for which the sentiment score was below the neutral value of 0.00, or those with a 1, 2, or 3-star rating that had a sentiment score higher than +0.50, were manually inspected. Twenty-nine reviews were removed using this procedure. Reviews with ratings of 4 or 5 stars were considered positive for the remainder of the analysis, while reviews with ratings of 1, 2, or 3 star(s) were considered negative. From the 327 most-disliked phones, 158 were assigned 3-stars, but the associated reviews were sufficiently negative to be considered negative reviews. Table 4.1 shows the average sentiment scores (positive values indicate positive sentiment) for all review types and their associated ratings. It can be seen that the mean scores – including those for Amazon reviews – were consistent with the review ratings.

Review type	Ratings				
	1	2	3	4	5
Prolific [owners]	-0.18	-0.03	0.11	0.35	0.42
Prolific [non-owners]	-0.09	-0.03	x	0.36	0.44
Amazon [owners]	-0.05	0.03	0.10	0.32	0.43

Table 4.1. Average sentiment scores across reviews and their ratings.

4.2.3.2 Matching Amazon and Prolific reviews

After all (Prolific and Amazon) reviews were filtered as described, they were matched according to smartphone and rating to generate complementary data sets. To do this, three review sets from Amazon were generated to mirror the three Prolific reviews sets. These were matched in terms of the smartphones reviewed and the ratings provided to reduce content-related confounds when comparing and classifying Prolific and Amazon reviews in later analyses. All Amazon reviews were considered truthful and from owners (as we only included those for which a purchase had been verified). Smartphone models that were not sold on Amazon or had only a limited number of reviews, were replaced with reviews for smartphone models from the same brand with the same ratings, resulting in 1,060 replacements (Appendix C1). This was not possible for 127 reviews. For these, they were replaced with a smartphone review from a randomly selected brand with the same rating. The final dataset consisted of 4,168 reviews (Table 4.2) which is publicly available at: <https://osf.io/29euc/>

Review type	Truthful		Deceptive		Total
	Pos	Neg	Pos	Neg	
Prolific owners	384	327	302	348	1,361
Prolific non-owners	-	-	352	371	723
Amazon owners	1,038	1,046	-	-	2,084
Total	1,422	1,373	654	719	4,168

Table 4.2. Overview of all filtered reviews.

4.3 Supervised learning analysis

N-grams, part of speech frequencies (POS), and LIWC (Linguistic Inquiry and Word Count, (Pennebaker et al., 2015)) features were extracted from the reviews. To do this, URLs and emoticons were removed from all reviews, and all characters were converted to lowercase. LIWC features were then extracted. Subsequently, we removed punctuation, tokenized the text, removed stop-words, and stemmed the text data. Since the LIWC software performs text cleaning internally and to retain the measures on punctuations, the LIWC features were generated first. From the cleaned data, we extracted unigrams, bigrams, and POS proportions for each text. The “WC” (word count) category from LIWC features was excluded. The Python package nltk (Bird et al., 2009) was utilized for text cleaning and feature generation. Lastly, during feature preprocessing, features with a variance of 0 in each class (e.g., truthful-positive-owners, deceptive-positive-owners) were excluded to avoid any non-content related features (e.g., Amazon-specific website signs or words, such as “verified purchased”) affecting the results. Appendix C2 provides the list of features, which were present across all analyses. In total, we removed 69,927 (99.92%) bigrams, 6,520 (94.31%) unigrams, 13 (37.14%) POS features (WRB, WP\$, WP, VBG, UH, TO, RBS, PRP, POS, PDT, EX, ", \$), and 4 (4.3%) LIWC features (we, sexual, filler, female).

Instead of using a pre-trained model from other published work, we decided to train and test our own classifier. Doing so meant that we could ensure that all observed changes in classification performance could be attributed to our experimental manipulations as opposed to other factors (e.g., domain differences or class imbalances in the training data of other models). We tested several different classifiers, but the “Extra Trees” classifier (Geurts et al., 2006) showed the best performance in most scenarios and is therefore reported throughout all analyses (see Appendix C3 for the classifier settings and Appendix C4 for a list of other tested classifiers). All classification models were implemented in Python with the “scikit-learn” package (Pedregosa et al., 2011). No hyperparameter changes were made.

4.4 Results

A total of six analyses were performed to investigate how well reviews could be classified in terms of their *veracity*, *ownership*, and *data-origin* alone, as well as how strongly *ownership*, and *data-origin* affected the classification of truthful and fake reviews individually and

combined. In each analysis, wherever needed, we balanced the classes by downsampling the reviews of the majority class.

4.4.1 Classification performance

Each analysis involves a binary classification task for a subset of the data (e.g., fake vs. truthful reviews, reviews from phone owners vs. non-owners, etc.), each separated into negative and positive reviews. Thus, each classification analysis contains reviews that exhibit one or more of the following: *veracity*, *ownership*, and *data-origin*. Classification performance is measured in accuracy (acc.), precision (pre.), recall, and F1, each of which is averaged across a 10-fold cross-validation procedure. Since the performance metrics behaved the same between classes, we only report accuracies here (see Appendix C5 for a full list of all performance metrics).

4.4.1.1 Pure classifications of *veracity*, *ownership*, and *data-origin*

The first three analyses examined how well reviews can be distinguished in terms of *veracity*, *ownership*, and *data-origin*. The binary classification analyses were carried out for the following pure (i.e., removing any confounds) comparisons: **(1)** Participant [owners, fake] and participant [owners, truthful] (assessing *veracity*), **(2)** participant [non-owners, fake] and participant [owners, fake] (assessing *ownership*), and **(3)** participant [owners, truthful] and Amazon [owners, truthful] (assessing *data-origin*). Classification performance for each analysis is reported in Figure 4.2. The results show that the classifier found it difficult to differentiate reviews that differed solely in terms of *veracity* or *ownership* but performed better for *data-origin*. However, classification performance was different by review sentiment. Specifically, *veracity* seemed to be more easily classified if reviews were negative.

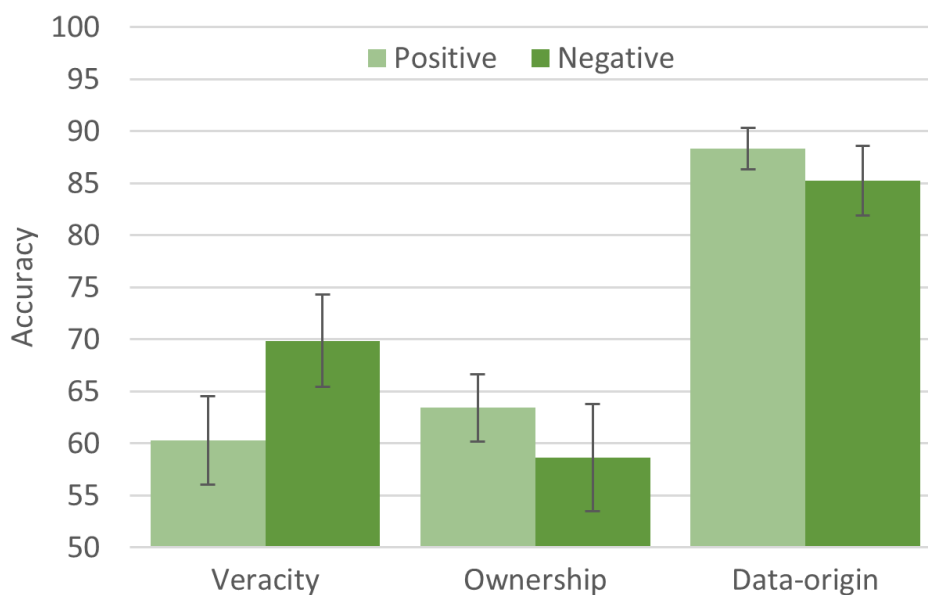


Figure 4.2. Classification accuracies (with 99% CI) of analyses 1 (*veracity*), 2 (*ownership*), and 3 (*data-origin*); Accuracy ranges from 50% (chance level) to 100%.

4.4.1.2 Confounded classifications of *veracity*

The last three analyses focus on the classification of *veracity* but examine how this is affected by – or confounded with – *ownership*, *data-origin*, and the two combined. The goal was to assess the strength of these factors to estimate the extent to which they (as confounders) might have affected the accuracy of classifiers in other studies. Specifically, we compared **(4)** participant [non-owners, fake] and participant [owners, truthful] (assessing *veracity confounded with ownership*), **(5)** participant [owners, fake] and Amazon [owners, truthful] (assessing *veracity confounded with data-origin*), as well as **(6)** participant [non-owners, fake] and Amazon [owners, truthful] (assessing *veracity confounded with ownership and data-origin*). Classification performance for each analysis (and analyses 1 for comparison) is reported in Figure 4.3. The results show that all confounds have a boosting effect on the *veracity* classification, but with different strengths. Compared to analysis 1 (Figure 4.2, assessing *veracity*) the *veracity* classification seems to be overestimated with the confound of *ownership* by 6.15 - 9.84%, with *data-origin* by 21.11 - 44.27%, and with *ownership* and *data-origin* combined by 24.89 - 46.23%, depending on sentiment.

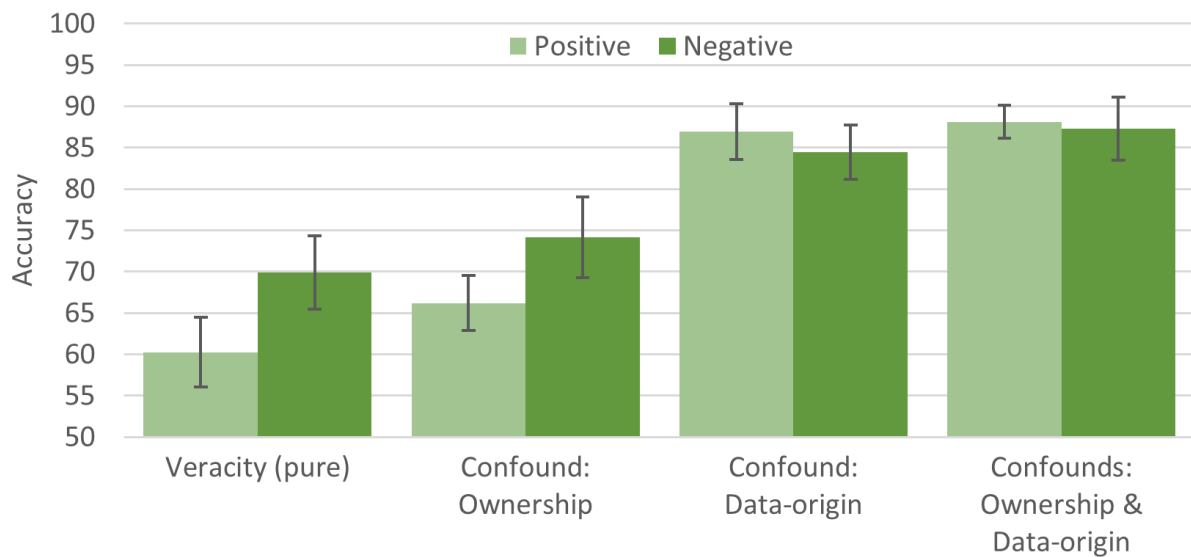


Figure 4.3. Classification accuracies (with 99% CI) of analyses 1 (*veracity*), 4 (*veracity, ownership*), 5 (*veracity, data-origin*), and 6 (*veracity, ownership, data-origin*); Accuracy ranges from 50% (chance level) to 100%.

4.4.2 Linguistic properties in pure and confounded classification experiments

The linguistic properties of all analyses were examined using Bayesian hypothesis testing (Kruschke, 2013; Ortega & Navarrete, 2017; van der Vegt & Kleinberg, 2020). The aim was to investigate which linguistic features drive each classification. To do this, we inspected the top 5 highest Bayes Factors (BF_{10}) and reported features with scores of 10 or greater. A BF_{10} indicates the likelihood of the data if there was a difference of occurrences in the feature between the compared classes (alternative hypothesis) relative to the null hypothesis (no difference). A BF_{10} of 1 represents an equal likelihood of the null- and alternative hypothesis.

Each reported feature name is tagged with one of the following indications: “POS” (part of speech), “LIWC” (Linguistic Inquiry and Word Count), “UNI” (unigram), “BI” (bigrams), to categorize its feature type. For example, “POS_CD” means cardinal digit, which is a number (e.g., 42) or “LIWC_social” means social processes, which includes all words the LIWC software relates to that construct. Table 4.3 describes all features that showed the five highest Bayes Factors in the analyses. More details about the LIWC features (e.g., which words are included in the categories) can be found in the LIWC documentation (Pennebaker et al., 2015).

Name	Meaning	Name	Meaning
Part of Speech (POS) features			
CD	cardinal digit	JJS	adjective, superlative (e.g., “biggest”)
LIWC features			
social	Social processes	Period	Periods
WPS	Words/sentence	ppron	Personal pronouns
focuspresent	Time orientations: Present focus	Time	Relativity: Time
focuspast	Time orientations: Past focus	Exclam	Exclamation mark
money	Money	percept	Perceptual processes
Tone	Emotional tone	Authentic	Authentic
conj	Conjunctions	auxverb	Auxiliary verbs
see	Perceptual processes: See	Article	Articles
i	Personal pronouns: 1st person singular	function	Function Words
Comma	Commas	work	Personal concerns: Work

Table 4.3. Explanation of all feature names.

Table 4.4 shows the top 5 linguistic features for the pure classification experiments, while Table 4.5 shows the top 5 features for the classification experiments in which we introduced confounds.

Testing		Experiment		
		Veracity	Ownership	Data-origin
Sentiment	Positive	LIWC_social (-)	POS_CD (-)	LIWC_Comma (+)
		POS_CD (+)	LIWC_see (-)	UNI_smartphon (+)
		UNI_phone (-)	UNI_im (-)	UNI_camera (+)
		UNI_iphon (+)	LIWC_WPS (-)	LIWC_social (-)
		LIWC_WPS (+)	LIWC_i (-)	UNI_best (+)
	Negative	LIWC_focuspresent (-)	UNI_samsung (+)	UNI_smartphon (+)
		LIWC_focuspast (+)		LIWC_Comma (+)
		LIWC_money (-)		UNI_slow (+)
		LIWC_Tone (+)		LIWC_Period (+)
		LIWC_conj (+)		LIWC_ppron (-)

Table 4.4. Top 5 feature differences by BF_{10} for all pure classifications; $BF_{10} > 11$ for each feature; (+) = feature appears more often in truthful reviews, in reviews by smartphone non-owners, or in Prolific reviews; (-) = feature appears more often in deceptive reviews, in reviews by smartphone owners, or in Amazon reviews; See Table 4.3 for all feature explanations.

Testing		Experiment		
		Veracity, Ownership	Veracity, Data-origin	Veracity, Ownership, Data-origin
Sentiment	Positive	UNI_year (+)	UNI_smartphon (-)	UNI_camera (-)
		LIWC_time (+)	UNI_camera (-)	UNI_smartphon (-)
		LIWC_Exclam (-)	LIWC_Authentic (+)	LIWC_article (-)
		LIWC_percept (-)	LIWC_auxverb (-)	LIWC_function (-)
		UNI_still (+)	POS_JJS (-)	UNI_photo (-)
	Negative	LIWC_focuspast (+)	LIWC_focuspast (+)	UNI_smartphon (-)
		LIWC_money (-)	LIWC_Authentic (+)	LIWC_Comma (-)
		LIWC_focuspresent (-)	UNI_qualiti (-)	LIWC_work (+)
		UNI_buy (-)	UNI_bad (-)	UNI_slow (-)
		LIWC_Exclam (-)	LIWC_social (+)	UNI_bad (-)

Table 4.5. Top 5 feature differences by BF_{10} for all confounded classifications; $BF_{10}>188$ for each feature; (+) = feature appears more often in truthful reviews; (-) = feature appears more often in deceptive reviews; See Table 4.3 for all feature explanations.

To complement the linguistic analyses of testing the feature distributions between review types, we also used structural topic modelling, which can be used to discover underlying topics within text data as a function of covariate variables (Blei et al., 2003; M. E. Roberts et al., 2019; van der Vegt et al., 2021). A topic model, such as Latent Dirichlet Allocation (Blei et al., 2003) or Correlated Topic Model (Blei & Lafferty, 2007), is probabilistic and assumes that a document is a mixture of topics and topics are a mixture of words. Thus, instead of utilizing a top-down approach, such as with the LIWC software, in which pre-selected words are assigned to pre-selected topics, topic models represent a data-driven, bottom-up approach. We used a structural topic model with the R package “stm”, which is a correlated topic model (M. E. Roberts et al., 2019). The structural topic model can account for document covariates, such as veracity or data-origin and whether they impact the prevalence of the derived topics within the corpus (e.g., the degree to which smartphone reviews are associated with a specific topic) and the terms (e.g., unigrams, bigrams) within the topics. Thus, we provided the model with unigrams and bigrams as terms and veracity (truthful, deceptive), data-origin (Prolific, Amazon), ownership (owner, non-owner), and sentiment (positive, negative) as covariates to assess how they might covary with topic prevalence. Based on the semantic coherence and exclusivity of terms, a model with 13 topics (tested with 2-100 topics) was selected. Semantic coherence measures the co-occurrence of highly probable words within a topic (Mimno et al., 2011), whereas exclusivity measures how often highly prevalent words in a topic do not appear as highly in other topics (M. E. Roberts et al., 2014). The idea is that by maximizing both metrics, semantically useful topics can be created.

We derived topic names by manually inspecting each topic’s top prevalent and unique words. Twelve of the 13 topics differed significantly in their prevalence for at least one review covariate (i.e., veracity, data-origin, ownership, sentiment). Figure 4.4 shows all topics that significantly differed in topic prevalence based on review type and the overall topic proportions across all reviews. Topic names are on the left of the horizontal bars, the top ten

frequent and exclusive (FREX) words are within the bars, and the review types for which the topic is most associated are indicated on the right of the bars in square brackets.

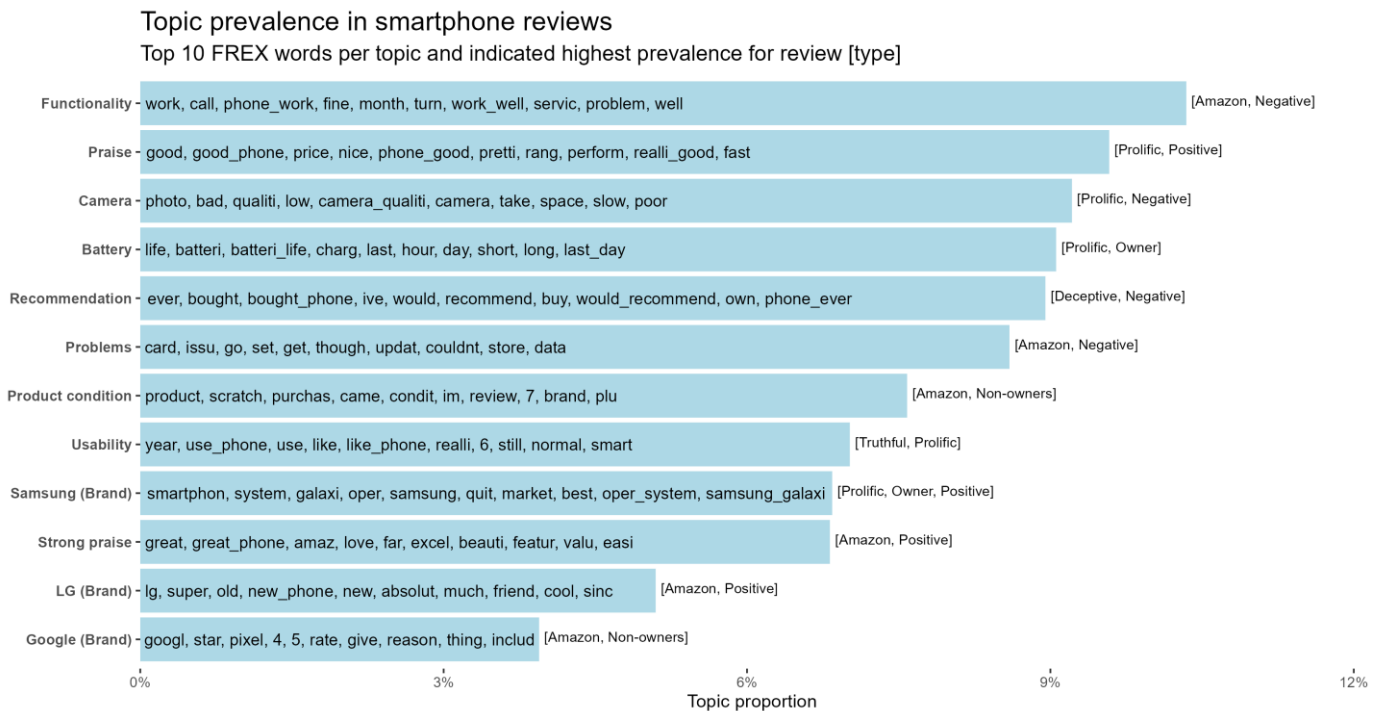


Figure 4.4. Shows the overall topic prevalence in smartphone reviews within the corpus (topic proportion) and for which review covariate the topic appeared more often (indicated in square brackets), including veracity (truthful, deceptive), data-origin (Amazon, Prolific), and ownership (owners, non-owners). Topic names are on the left of the horizontal bars, and the top ten frequent and exclusive (FREX) words are within the bars.

Data-origin seems to significantly affect the prevalence of most topics, with six topics most prevalent for Amazon reviews (Functionality, Problems, Product condition, Strong praise, LG, Google) and five for Prolific reviews (Praise, Camera, Battery, Samsung). Eight topics are significantly affected by sentiment, with four showing increased prevalence for positive (Praise, Samsung, Strong praise, LG) and four for negative reviews (Functionality, Camera, Recommendations, Problems). Two topics are more prevalent for owners (Batteries, Samsung), and two for non-owners (Product condition, Google). Truthful reviews refer to Usability more often, while deceptive reviews Recommendations.

4.5 Discussion

This study investigated how product *ownership* and *data-origin* might confound interpretation of the accuracy of a machine learning classifier used to differentiate between truthful and fake reviews (*veracity*), using smartphones as a case study. To disentangle the unique contributions of each factor, we devised an experimental data collection procedure and created a dataset balanced on all factors. We used supervised learning to examine pure and stepwise confounded classification performance.

4.5.1 Classifying veracity

The supervised machine learning analyses showed that after controlling for two possible confounders (*ownership* and *data-origin*) reviews can be classified as to their pure *veracity*, but with difficulty, suggesting that detecting fake reviews may be harder than other studies have reported (Ott et al., 2011, 2013). Furthermore, at least for our data, negative reviews appear to be easier to classify (by almost 10%) than positive ones, suggesting that the way individuals deceive differs depending on sentiment.

4.5.2 Classifying confounded veracity

As discussed, the two confounds tested led to an overestimation in the classification of *veracity* of between 6-46%, depending on which confound, and sentiment was involved. The combined confounding effect of *ownership* and *data-origin* (24.89-46.23% overestimation, depending on sentiment) seems to have the strongest effects, followed by *data-origin* (20.85-44.27%) and *ownership* (6.15-9.84%) alone. The ordering of these effects is the same for both positive and negative reviews. However, classification performance for positive and negative reviews differs. Specifically, negative reviews are easier to classify when *veracity* is confounded with *ownership*, but the reverse is true when *veracity* is confounded with *data-origin* or *data-origin* and *ownership* combined. Additionally, the difference in performance by review sentiment is most clear when *ownership* is involved (8.22%) than for *data-origin* (2.5%) or both combined (0.86%). The performance differences associated with the change in sentiment (e.g., *veracity* vs. *ownership* classification) further supports the idea that the *veracity* classification is sentiment dependent and might not be as easily generalizable.

Interestingly, when comparing the pure classification of *data-origin* to the confounded classification of *veracity* and *data-origin*, performance was almost identical. The similarity in classification performance seems slightly counter-intuitive, as one would expect that an increase in difference would lead to an increase in performance. A possible explanation is that some of the Amazon reviews were deceptive, which would mean that they add no additional information to the classification task. The interpretability rests on the assumption that Amazon reviews are truthful, which is further discussed below in section 4.5.5.

4.5.3 Linguistic properties

Examining the top 5 features ranked by their Bayes Factor₁₀ for each classification task provides insight into each class's text differences. As expected from the classification performance metrics, we observe that features are not consistent across sentiment nor between or across classes.

Both deceptive and truthful reviews highlight non-psychological or non-perceptual constructs (except LIWC_{social}). Given the increased usage of “phone” in deceptive positive reviews,

individuals writing such reviews might reiterate what review type they are writing. Similarly, truthful review writers seem to highlight that their product originates from the brand Apple. Truthful reviews also seem to focus on the past, which might support the idea of highlighting past-owned phones. Thus, truthful negative reviews might exhibit a stronger emphasis on the ownership of the phone and when it was owned. In turn, deceptive negative reviews seem to focus more on the present and money, suggesting that the phone price might be over-emphasized in such reviews. For differences in *ownership*, we observe fewer differences, but smartphone owners seem to express their experience more in personal terms than do non-owners. However, this was only the case for positive reviews. Prolific reviews seem to follow a more factual, syntactical, and sentiment-specific style, suggesting a stronger emphasis on the product. In contrast, Amazon reviews seem to include more social processes (LIWC_social) and personal pronouns (LIWC_ppron), which could be attributed to an increased focus on services (delivery, refunds, customer support, etc.). However, the increased usage of the words “best” and “slow” in Prolific reviews could serve a similar function.

An examination of the linguistic properties of the confounded classification experiments shows a mixed picture of classification features that appeared in the initial experiment (i.e., when no confounds were introduced) and some new ones. For example, feature differences for the classification of *veracity* and *ownership* show a new set of features for positive reviews. For the negative reviews, however, the features identified were those previously seen in the pure *veracity* classification. Since almost no differences were present for the *ownership* condition features for negative reviews, such an effect seemed expected. Interestingly, when *veracity* is confounded with *data-origin*, or with *ownership*, and *data-origin* combined, we see recurring features from the purely *data-origin* classification, which is strongest when both confounds are present. Thus, *data-origin* seems to have a strong linguistic impact, which is reflected in strong classification performance.

Lastly, examining the results of the structural topic model, we also see that many topics are affected by *data-origin* (11), some by *ownership* (4) and two by *veracity*, reflecting the trend of the confounds impacting classification performances. We can further understand the language used in reviews by inspecting how topics differ within and between the review types. Reviews on Amazon refer more often to the brands LG and Google, the smartphone’s functionality, or usability and in what conditions the phone is (presumably after shipping). In contrast, Prolific reviews refer more often to the brand Samsung, and focus more on the phone’s features and technical details, such as the camera’s quality and the batteries’ capabilities (e.g., battery life, charging). Amazon and Prolific reviews praise the smartphone’s features (topic Praise and Strong Praise). While smartphone owners seem to refer more often to Samsung phones and were more concerned about the smartphone’s features, such as batteries, non-owners referred more often to Google phones and were more concerned with the condition of the phone (e.g., scratches) after the purchase. Some of the top FREX words in the topics Samsung and Camera were also listed within the top five BF scores in the Bayesian hypothesis testing of feature differences between review types (section 4.4.2). In both

analyses, the words (features) were affected by the same covariates, data-origin and ownership, showing some overlap in the results.

4.5.4 Practical implications and generalization

Previous studies that have examined the effects of supervised deception (veracity) detection suggest, that model performance does not easily transfer across domains, datasets or languages (Belavadi et al., 2020; Capuozzo et al., 2020; Levitan et al., 2016; Li et al., 2014). While domain specific language features (e.g., words specifically related to a product or service) probably contribute to the difficulties of model generalizability, other features related to the research setup or data collection procedures have also been candidate explanations for low model transferability (Li et al., 2014). The current study supports the idea that the research design can have a strong effect on the classification performances in deception detection. Specifically, how the data is sourced (*data-origin*) seems to substantially effect classification performance. Consequently, models that are trained on data with confounds, such as in *data-origin*, will most likely not generalize well and will probably perform poorly when employed on data that is sourced differently than the models' training data. Thus, our findings indicate the importance of controlling for confounds in the training of classifiers.

Two other studies have examined possible confounds in fake review detection (Fornaciari et al., 2020; Salvetti et al., 2016), utilizing reviews about books, electronics or hotels. While their study designs differ from that used in this study, their results seem to indicate that confounds are also present for other review types. Both studies sourced their reviews from crowd-working platforms, such as Amazon Mechanical Turk (Salvetti et al., 2016), CrowdFlower (now called Appen) and the online shopping platform Amazon (Fornaciari et al., 2020), suggesting that confounds might also translate when utilizing a different crowdsourcing platform, such as Prolific. Further tests will have to determine if such confounding effects are also present for other online (shopping) platforms that contain reviews (e.g., Yelp, TripAdvisor, Google reviews, Etsy, Airbnb). In addition, future studies would also have to examine whether current results translate to other domains and frauds that include image or (online) behavioural data.

4.5.5 Limitations

4.5.5.1 Are verified Amazon reviews truthful?

This study is not without limitations. Chief among these is the assumption that Amazon reviews are truthful, which rests upon the “verified purchased” seal. However, the current system used by Amazon does seem to be exploitable (D. Lee, 2020; Schiffer, 2020). Investigative journalists have found several potential ways to circumvent the verified purchased seal: (1) Companies send customers free products in exchange for positive reviews (Keegan, 2020); (2) Companies send packages to random addresses, which are registered with Amazon accounts, to obtain fake reviews (Business Insider, 2021; Eisenbrand, 2018; Peteranderl, 2019); (3) Sellers hijack reviews from other products (Swearingen, 2019). Thus, it is plausible that some of the Amazon reviews used in this study were deceptive. Since we

cannot control for reviews posted on Amazon, this is difficult to test. However, it could explain findings that showed almost equal performance when the veracity was confounded with *data-origin* (i.e., deceptive reviews from Prolific participants and truthful reviews from Amazon customers) compared to the pure *data-origin* (participants vs Amazon, both truthful reviews) test. A small part of the Amazon reviews, which are assumed to be truthful, could be deceptive and might lead to an increased similarity with deceptive participant reviews, making it more difficult to differentiate them. Future studies that use truthful and deceptive participant reviews could test this hypothesis. By incrementally contaminating the truthful reviews with deceptive reviews (i.e., deceptive reviews labelled as truthfully) and reporting the classification performance for each step, the changes in classification performance could be estimated. Since the differentiation should become more difficult, a drop in classification performance would be expected, which can then be tested for association with the degree of contamination. If observed, the performance drop could then serve as an indirect indicator of fake review contamination. We could then compare the percentage difference of classification performance of *veracity* (analysis 1), *data-origin* (analysis 3) and *veracity* and *data-origin* (analysis 5) to estimate the contamination of fake reviews within Amazon reviews. However, the idea that deceptive reviews from Amazon and a crowdsourcing platform are similar contradicts other findings (Fornaciari et al., 2020). Nonetheless, fake Amazon and fake crowdsourced reviews would only need to show some similarities or at least only be more similar to each other than fake crowdsourced to truthful Amazon reviews to have a negative effect on classification performances.

4.5.5.2 Quality of Prolific reviews

We cannot be certain that Prolific participants who wrote the smartphones reviews were honest when instructed to be. However, previous research has shown that in most cases crowdsourcing platforms produce high-quality data and are better suited to the collection of large amounts of text data than other traditional collection methods, such as student samples (Chandler et al., 2019; Peer et al., 2017). Research also suggests that compared to other crowdsourcing platforms (e.g., Amazon Mechanical Turk, CloudResearch), Prolific seems to produce data of higher quality and with the most honest responses (Eyal et al., 2021).

4.6 Conclusion

Through careful experimental control, we found that product *ownership* and *data-origin* do confound fake review detection. This may have resulted in the overestimation of model performance in detecting veracity in previous work. In particular, *data-origin* seems to boost classification performance, and this could easily be misattributed to the classification of *veracity* alone. Our findings suggest an overestimation of 24.89-46.23% when data is sourced from different platforms. Consequently, more effort and experimental control are necessary to create datasets when investigating complex concepts such as deception.

Chapter 5: Counterfeits on cryptomarkets: A measurement between Jan-2014 and Sep-2015

This chapter is based on the following publication:

- Soldner, F., Kleinberg, B. & Johnson, S.D. Counterfeits on dark markets: a measurement between Jan-2014 and Sep-2015. *Crime Sci* **12**, 18 (2023). <https://doi.org/10.1186/s40163-023-00195-2>

5.1 Introduction

Counterfeits are illicit goods that violate intellectual property (IP) rights such as copyrights, trademarks, design rights, or patents, and they can exist physically or digitally (OECD/EUIPO, 2019; WTO, 1994). The purpose of a counterfeit is to make a monetary profit by deceiving a customer into believing that the product is of a higher value than it is (OECD/EUIPO, 2019). Counterfeits can cause a variety of problems, such as physical (e.g., through foods or pharmaceuticals) and monetary harms to the consumer, the IP holder (e.g., through damages to the brand value, loss of sales), or the government (e.g., through the loss of tax income) (EMCDDA-Europol, 2017; OECD/EUIPO, 2019). In turn, the sales of counterfeits can support organized crime groups financially and facilitate other illegal activities, such as money laundering (EMCDDA-Europol, 2017; UNICRI & ICC BASCAP, 2013; UNODC, 2014). The OECD/EUIPO (2019) estimated that counterfeits made up 3.3% of worldwide trades in 2016, worth USD 509 billion. Furthermore, the proportion of counterfeits seems to increase and be exacerbated within developed regions, such as the European Union (EU). However, estimating counterfeit goods' trade (value) is difficult and is mostly achieved through auditing goods seized at borders (OECD, 2018; OECD/EUIPO, 2019). Thus, current estimates often do not include domestically traded counterfeits or digital products, and since not all counterfeits will be seized at ports, estimates of what is traded may be incomplete. For example, the number of routinely checked containers at major ports a Genoa (Italy), Melbourne (Australia), Montreal (Canada), New York (USA), and Liverpool (UK) together, only account for 2-5% of all traffic (Sergi, 2022). Since only a limited number of containers can be checked, the selection procedure can strongly impact possible finds.

A theoretical and empirical understanding of how counterfeiting occurs is currently not well developed, perhaps due to the complex involvement of various stakeholders, which results in difficulties for researchers to obtain reliable data (Sullivan et al., 2017). For example, many companies affected by counterfeiting operate across nations, affecting the ease with which authorities can monitor and combat counterfeits. Moreover, the definition of counterfeits varies across nations, further complicating how counterfeiting is measured. However, theories provide perspectives as to why counterfeiting occurs and how it might be addressed. As previously discussed, the Rational Choice perspective considers the offender's choice to

commit a crime (e.g., counterfeiting a product) and the factors that influence this (Clarke & Cornish, 1985). The perspective suggests that changing offender's perceptions of the risks and rewards – such as increasing the perceived risks of detection or by increasing the general efforts needed to commit a crime – can affect the likelihood of offending. Within the context of counterfeits, facilitating the traceability of genuine products within a supply chain (e.g., through watermarks) is one approach to increasing the efforts to counterfeit (Gayialis et al., 2022). Another perspective, such as the Routine Activity Approach (RAA), discussed by Spink et al. (2013, 2014), states that crime is more likely when a suitable target (e.g., a product that can be counterfeited) and a motivated offender converge absent a capable guardian (L. E. Cohen & Felson, 1979). Capable guardians can include those involved in security at country borders or those involved in inspecting goods at other stages of the supply chain (Maruchek et al., 2011; Tang, 2006). For example, when manufactured products are transported, transport personnel and employees could also act as guardians (Hollis & Wilson, 2014). However, effective guardianship requires a clear understanding of the problem and processes to monitor it, such as reporting procedures. Absent this understanding, guardianship will be less effective.

With this in mind, risk assessments are often conducted to help with decisions based on intelligence from federal and local authorities and custom officer experiences (Sergi, 2022). Checks can also be random or may only be informed by the country of origin or how the delivery is labelled, as in the case of parcel shipments (Männistö et al., 2021). Another indicator of possible biases present in the check-selection procedures are large prevalence estimation differences in counterfeit product types from different agencies (IP Crime Group, 2015; OECD/EUIPO, 2019). For example, estimations can strongly differ for footwear by 20 percentage points or in electronics and clothing with differences of 11 and 10 percentage points, respectively. For other products, estimations may be missing entirely, as in the case of tobacco, which was measured to make up 28.15% of all counterfeits by the IP Crime Group (2015) but was not measured to be counterfeited by OECD/EUIPO (2019). Since the different agencies can have different data sources (e.g., border or inland seizures), some measurement differences are expected, but they also illustrate how inconsistent seizures reflect the true prevalence of counterfeits. Thus, additional data sources to estimate counterfeit affected products would be helpful to better understand the counterfeit landscape.

With the emergence of cryptomarkets in recent years, new ways of trading illicit goods, including counterfeits, have appeared (Christin, 2013; van Wegberg et al., 2018), which may serve as an additional data source to measure counterfeit prevalence. Cryptomarkets are online shopping platforms on the deep web, a highly anonymized part of the internet that is not indexed by traditional search engines, and which operate like their surface web counterparts, eBay or Amazon. Vendors on cryptomarkets offer a range of illegal products and services, mainly consisting of drugs, but also including hacking services, weapons, guides on how to defraud people, and counterfeits (Baravalle & Lee, 2018; D. L. Roberts & Hernandez-Castro, 2017; Soska & Christin, 2015; van Wegberg et al., 2018). During the COVID-19 pandemic, markets also started to offer a mix of genuine and fake protective gear (masks,

gloves, etc.), medicines, and COVID-19 vaccines (Bracci et al., 2021a, 2021b; Broadhurst & Ball, 2020). Even with successful disruptions and the closing of markets by law enforcement, cryptomarkets increasingly trade in such products and services (Décary-Hétu & Giommoni, 2017; ElBahrawy et al., 2020; EMCDDA-Europol, 2017). On cryptomarkets, vendors openly sell counterfeits and forgeries, which provides an interesting opportunity to gain insight into the counterfeit market from a new angle. Since some cryptomarkets also register the number of goods sold and buyers leave reviews, we can use such information to generate estimates of sales volume and counterfeits' monetary value over time. By comparing counterfeit listings and their sales on cryptomarkets to border seizures, we can see if they differ and provide a more comprehensive picture, which would be of value to law enforcement and policymakers.

Therefore, to better understand the counterfeit economy on anonymity networks, we examined the prevalence and sales of counterfeits sold on 89 cryptomarkets, covering three years (January 2014 – January 2017). Specifically, we quantified the (price) volume, type, and origins of advertised counterfeits and estimated their sales volume and the value the same counterfeits would attract on the surface web. We then compare the results to measures and estimations from border seizures conducted by law enforcement over the same period. By highlighting discrepancies, we can identify product groups for which counterfeiting appears to be a problem and would be overlooked based on an analysis of seizures alone. Thus, the aim of this study is to investigate whether monitoring the counterfeit economy on cryptomarkets could inform authorities in the future by highlighting individual and groups of products that are known to be counterfeited.

5.1.1 Fraud and counterfeits on cryptomarkets

Studies that have investigated the types of products listed and sold on anonymity networks mostly cover illegal drugs, which often account for 60-80% of all listings on a cryptomarkets (Baravalle & Lee, 2018; EMCDDA-Europol, 2017). However, some studies have examined less frequently listed products, such as art, wildlife, and plane tickets (Hutchings, 2018; Paul, 2018; D. L. Roberts & Hernandez-Castro, 2017). Others focus on fraud-related products or services, such as credit card information, online accounts (e.g., eBay), social engineering guides and tutorials, or financial malware (e.g., ransomware) (Garg et al., 2015; Marin et al., 2016; Schafer et al., 2019; van Wegberg et al., 2018). Although some of these studies have considered the sale of forged documents (e.g., passports, licenses, diplomas), none have investigated or quantified the sales of counterfeits in a systematic way, such as differentiating between clothing, shoes, electronics, or jewellery; product types which can also be found on surface web markets (e.g., eBay, Amazon). Europol (2017) draws attention to IP crime on anonymity networks and estimates that solely counterfeit goods make up around 1.5-2.5% of all listings on such markets. The report lists some of the types of counterfeits sold on cryptomarkets (e.g., clothes, accessories, electronics, jewellery, pirated goods) and assumes that products and services are sold with high frequency and low volume, but estimating actual sales is difficult. According to this report, counterfeits seem to be sold for 1/3 of the price of the equivalent genuine product, and digital goods for around 1/6 of their original price (Europol, 2017). The

report concludes that while the sale of IP goods is limited, there is potential for growth on cryptomarkets, and IP goods on cryptomarkets should be monitored and investigated in more detail. However, the report explains neither how the mentioned statics were obtained nor which cryptomarkets were included in the analyses, making it difficult to assess the extent of counterfeits on anonymity networks. Furthermore, the lack of granularity prevents us from understanding which product types are offered, how frequently, how much they are sold, and where they originate. Lastly, the Europol (2017) report does not differentiate between counterfeits that could be sold on the surface web (e.g., shoes, clothes, electronics) and counterfeits that are limited to anonymity networks (e.g., fake banknotes or IDs), which is important if we want to inform authorities on potentially affected product types that could be sold on the surface web.

Therefore, we aim to address the shortcomings of previous work by examining an extensive collection of cryptomarket datasets to (I) understand the prevalence of counterfeit goods on anonymity networks and (II) determine the product types, occurrences, and origins of the identified counterfeits. Determining those details will help us (III) report counterfeit prices more accurately (by product types) and make sales volume estimations through product feedback, which can help us better understand the counterfeit economy on anonymity networks. Subsequently, we (IV) compare counterfeit prices on anonymity networks with prices of the same products on the surface web to understand possible profit margins for the various found product types. Lastly, we (V) compare our results to observations made through border seizures, complaint statistics, and activities from authorities to contribute to the overall understanding of the counterfeit economy and highlight differences between our and other estimates, discussing possible reasons and future research avenues.

5.2 Data

The data used in this study originated from the “Darknet Market Archive”¹¹, a collection of 89 markets and associated forums (Branwen et al., 2015) for which data were initially collected between 2014-2015 and continuously supplemented thereafter. To facilitate the selection of relevant markets, we cross-referenced the available market data with a list of markets documented by EMCDDA-Europol (2017). Through this comparison, we identified 38 markets (see Appendix D1), each of which operated for at least six months and was captured in the data archive. The reason for including markets that operated for at least six months was to ensure that the markets were able to attract enough vendors and customers, allowing for a broader range of product offers and trades¹². The market archive contained data on 30 of the 38 identified markets, but five of them contained data spanning less than six months, and data on eight markets did not include a sufficient self-organizing structure (e.g., categories), which would have allowed for the identification of counterfeit goods. For example, some market data contained products (e.g., shoes, handbags) without categorization or a detailed

¹¹ Data: <https://www.gwern.net/DNM-archives>

¹² Manual inspections of the data revealed that markets with a shorter lifespan did not sell any counterfeits and often harbored very few vendors.

description, making it impossible to determine if they were counterfeits or originals that had been stolen or otherwise illegally obtained. Furthermore, six markets were either highly specialized (e.g., solely carding or marijuana markets) or did not contain any counterfeits. Thus, we included the remaining 11 markets in our study (see Table 5.1).

Name	Data-start	Data-end
The Marketplace	03.01.2014	09.11.2014
Agora	01.01.2014	07.07.2015
Evolution	21.01.2014	17.03.2015
Cloud 9	11.02.2014	01.11.2014
BlackBank Market	06.02.2014	17.05.2015
Andromeda	12.04.2014	18.11.2014
Middle Earth Market	23.06.2014	05.07.2015
Diabolus/Silk Road 3	17.10.2014	05.07.2015
Abraxas	16.12.2014	05.07.2015
AlphaBay	21.12.2014	28.01.2017
Crypto Market	19.02.2015	06.07.2015

Table 5.1. Markets and their data timeframe in this study.

5.2.1 Data filtering

Each market listed a range of products that were not counterfeits (e.g., drugs, services, weapons). Consequently, it was necessary to exclude such listings prior to analysis. To do this, we created a corpus of counterfeit products in two steps. First, we included products that were clearly categorized as counterfeits based on the categories used on the markets, such as “Counterfeit[s]”, “Replica[s]”, “Counterfeit Items”, and “Replica watches”. Second, listings that were not included on this basis were filtered using an advanced keyword search. These keywords were for 29 other categories of items that, through the manual inspection of the data, were identified as including counterfeits (see Appendix D2 for the complete list of the categories). To facilitate the advanced keyword search, we merged the title and description of each listing in those 29 categories, lowercased, tokenized, and stemmed the text, and removed all punctuation. We then searched for 44 stemmed synonyms of the word “counterfeit” (e.g., “fake”, “clone”; a complete list is provided in Appendix D3) as well as six negated synonyms of “authentic”, using bigrams (e.g., “genuine”, “original”; a complete list is provided in Appendix D4) in each merged listing text. Lastly, a list of keywords was used to exclude listings (Appendix D5) that sold templates or tutorials on how to counterfeit. 124,379 listings were clearly marked as counterfeits, while 42,775 listings were identified through the keyword searches, resulting in a total of 158,228 counterfeit listings overall. Of these, 11,633 were completely unique listings for which at least the title, description, and vendor name differed. Text processing was conducted using the python package “nltk” (Bird et al., 2009).

5.2.2 Categorizing counterfeits

To determine the distribution of product types among those identified as counterfeits, we trained a machine-learning classifier on a subset of human-annotated data. The classifier was then used to predict the categories of the remaining unannotated products. To generate the annotated data, we randomly extracted 2200 unique listings from the counterfeit data set, which participants from the crowdsourcing platform *Prolific* subsequently annotated¹³. To ensure that we obtained accurate annotations, each listing was annotated by at least three participants. We recruited 220 participants, each annotating 30 listings based on the listing title. The final category label for each listing was determined using the majority vote.¹⁴ When annotating, participants were presented with an online interface and were required to select one of the following category labels: “Watches”, “Handbags”, “Wallets”, “Sunglasses”, “Other accessories”, “Clothing”, “Footwear”, “Articles of leather”, “Fabrics (silk, rugs)”, “Phones”, “Electronics”, “Jewelry”, “Cosmetics”, “Pharmaceuticals”, “Metals”, “Tobacco”, “Forgeries (Money, Coupons, IDs, etc.)”, “Services”, “Other”.¹⁵ We calculated Krippendorff's alpha to determine how much annotators agreed on the labels they generated¹⁶ (Feng, 2015). The value of $\alpha = 0.75$ demonstrated good agreement (A. F. Hayes & Krippendorff, 2007; Krippendorff, 1970).

From the distribution of categorized products, it was apparent that the product types were not uniformly distributed, with watches representing the majority of all counterfeits annotated. Because some of the categories had low numbers, which would likely affect the classifier's performance, when training the classifier, we manually added eight listings to the tobacco category and six listings to the cosmetics category. Table 5.2 shows the resulting distribution (after manually adding listings) of the labelled categories for the randomly selected subset of counterfeits.

¹³ www.prolific.co

¹⁴ Due to the allocation procedure of participants from Prolific to our annotation task, some listings were only annotated twice while some were annotated more than three times. 192 annotations ties were manually resolved by the first author.

¹⁵ The categories were determined based on reported counterfeits for seized goods by law enforcement (OECD/EUIPO, 2019).

¹⁶ Krippendorff's alpha ranges between -1 (perfect disagreement) and 1 (perfect agreement) and can account for unequal numbers of annotators and annotations per item as well as missing annotations.

Category	#	%
Watches	902	40.76
Forgeries	168	7.59
Clothing	166	7.50
Other	148	6.69
Footwear	138	6.24
Electronics	121	5.47
Handbags	120	5.42
Sunglasses	108	4.88
Jewelry	81	3.66
Other accessories	69	3.12
Services	64	2.89
Wallets	34	1.54
Metals	34	1.54
Pharmaceuticals	22	0.99
Cosmetics	21	0.95
Tobacco	17	0.77

Table 5.2. Annotated categories within counterfeits.

5.2.3 Automated labelling

Inspired by previous research (van Wegberg et al., 2018), we used the annotated listings to train a multiclass classifier to predict the labels of the remaining unlabelled counterfeits. Obtaining labels for all the listings has the advantage of allowing us to conduct our analyses for the whole dataset, including the price or individual texts of the listings, which would be more difficult through estimations from a sub-sample. We generated text features from the merged product title and description to train the classifier. However, we first converted the text to be all lowercase and removed all punctuation. We then tokenized the text, removed all English stop words, and stemmed the remaining words. Subsequently, we generated part of speech tags, unigrams, and bigrams, which were weighted with a tf-idf (term frequency-inverse document frequency) score. The python package “nltk” (Bird et al., 2009) was used for all text cleaning and feature generation steps. To increase the classifier's performance, we used a mix of under- and over-sampling methods to balance the number of product listings between the categories. First, the category watches was under-sampled, reducing the number of listings in the sample. This was followed by oversampling of the remaining categories to increase the number of these listings in the sample, resulting in an equal representation between all categories, each consisting of 450 listings. To reduce the number of listings within each category, we randomly selected listings (without replacement) from the data until we reached 450 listings. To increase the number of listings within a category, we used “SMOTE” (Synthetic Minority Oversampling Technique), which synthesizes new unseen data points (Chawla et al., 2002). Such new data is generated by first randomly selecting a listing of that category and finding the k (5) nearest neighbours of that listing within the feature space. Then, one of the neighbours is selected at random, and a new data point is created at a random point between the two listings in their feature space. Both under- and over-sampling methods

were implemented in python using the package “imblearn” (Lemaître et al., 2017). Next, we utilized the “LinearSVC” classifier with an “l2” penalty (the default regularization parameter used to reduce complexity in the model and avoid overfitting) using a 10-fold cross-validation procedure. The under- and over-sampling, training, and testing steps were embedded within a pipeline so that the classifier was trained on the balanced listings (450 in each category) but tested on the unbalanced listings (as in Table 5.2), ensuring a fair assessment. The test performances were evaluated using the average accuracy, and the weighted average of precision, recall, and F1 scores across all folds, as shown in Table 5.3. The python package “scikit-learn” (Pedregosa et al., 2011) was utilized for training, testing, and evaluating the classifier.

Accuracy	Precision	Recall	F1
0.85	0.85	0.85	0.85

Table 5.3. The performance scores (weighted average) across 10-folds.

To better understand the classifier's performance for each category, we generated a normalized confusion matrix for all classes (Figure 5.1). The matrix shows the cases of true (rows) and predicted (columns) categories of the listings. Thus, the values in the matrix show the proportion of items for which the true class was predicted. The diagonal cells (left-top to right-bottom) indicate the correct proportion for each category.

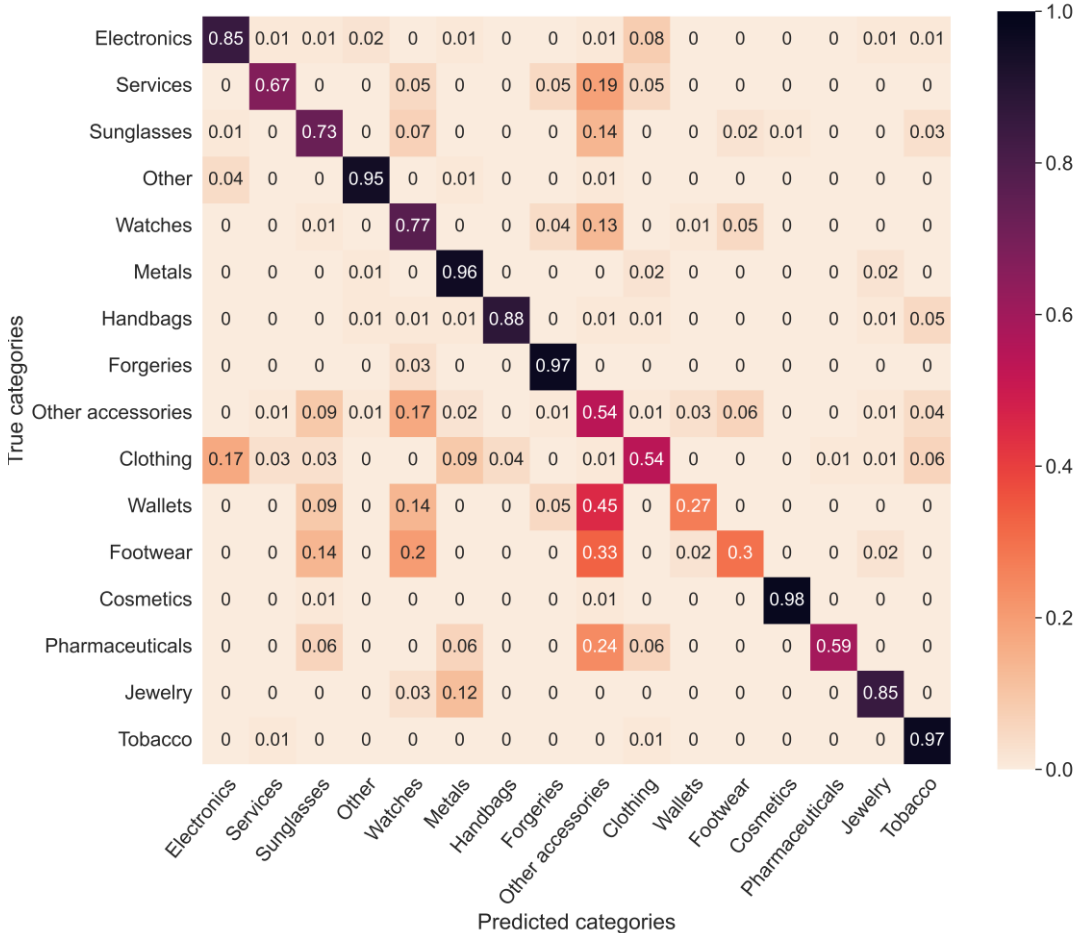


Figure 5.1. Normalized confusion matrix for true and predicted categories of counterfeits.

Classification performance was generally good, but we observed that six categories showed low (cosmetics, tobacco, other accessories, other) or very low (pharmaceuticals, services) categorization performance. Since low performances are only present with classes exhibiting few listings in the test set, most of the listings are well categorized, which is also reflected in the weighted performance scores (Table 5.3). An exception was for the category other, which was also less well categorized despite containing more listings than the other low-performing categories. The category other often contained custom orders, with product titles such as “custom [customer name]”, complicating the annotation process. Since the classifier received additional information from the product description, which was not available to the annotators, it is possible that mismatches between the annotations and product descriptions led to more misclassifications in the category other. For example, some custom orders might have similar descriptions as other counterfeits. Besides custom orders, the category other also included guides, instructions, counterfeit art (e.g., paintings), or cars.

Having established the accuracy of the classifier to predict the unlabelled listings (i.e., label all the unannotated listings), the entire annotated data was utilized to re-train the LinearSVC classifier with the same parameters. The advantage of re-training the classifier with the entire annotated data instead of using the best classifier from the cross-validation procedure, which is trained only on a subset of the annotated data, is that all the annotated data can be leveraged for the training, which supports better predictions.

5.2.4 Holding and placeholder prices

Previous studies about cryptomarkets sometimes encountered holding prices, which vendors use to mark out-of-stock listings, preventing their removal from the market (Soska & Christin, 2015; van Wegberg et al., 2018). Some holding prices are very high to prevent anyone from buying the product. The advantage of a holding price is that vendors can keep showing customers what was sold and what might be coming back in stock. However, when estimating price or sale volumes on markets, holding prices with very high values can distort the actual results. Therefore, we used a heuristic proposed and used by others (Soska & Christin, 2015; van Wegberg et al., 2018) to replace high holding prices ($\geq 10,000$ USD) with the original price (if available) or to remove it. In addition, we also looked at listings with very low prices (≤ 5 USD) and found that such prices were mainly not the actual selling price and seemed to function as placeholders too. For example, many listings with a price of 0 need further specifications by the customer (often instructed in the listing description), such as amounts, colours, or shipping, which affects the final price. However, during the data scraping process, the placeholder price is mostly that which is collected rather than the individual price variations. In some instances, vendors listed the variations of the products in separate listings and later merged them into a single listing with the option of making the wanted changes (colour, amount, etc.) or vice versa. In such cases, we can determine the average price of such a merged listing to get a more accurate representation of the product price. For listings with a holding and placeholder price, we searched for the same product from the same vendor to

find a replacement price. Table 5.4 shows the distribution of found and replaced holding and placeholder prices.¹⁷ Products with a high holding price for which we did not find replacements were excluded from further analyses of the value of the goods.

	n	% of all listings	Replaced	Avg price (USD)
Holding (≥10,000 USD)	120	0.08	83	140.67
Placeholder (≤5 USD)	6,106	3.87	1,040	178.64
Total/Avg	6,226	3.94	1,123	159.66

Table 5.4. Number of found and replaced holding and placeholder prices and the average price of all replacements.

5.3 Results

This section looks at the data for all products and counterfeits and their distribution across markets. We then focus on counterfeit product types and product origins and compare our measures with estimates from audits of goods seized by law enforcement at borders. Lastly, we evaluate the monetary value of offered and sold counterfeits and the generated sales volume of vendors.

5.3.1 Product offers and counterfeit prevalence

Figure 5.2 shows how many products (not just counterfeits) were offered across all markets over time. The volumes shown are monthly and contain all available products on the cryptomarkets. For most markets, the data range between January 2014 and September 2015, but the data for the market Alphasbay extends to January 2017. Evolution and Agora offered the most products, followed by Alphasbay, Abraxas, BlackBank Market, and Cloud 9. The remaining markets seem to have offered only a minimal number of products and for shorter periods. Reasons for this variation differ. For example, some markets were closed down by law enforcement (Cloud 9, Alphasbay), closed down voluntarily (The Marketplace, Agora), experienced an exit scam¹⁸ (Evolution, BlackBank Market, Andromeda, Middle Earth Marketplace, Abraxas), or were hacked (EMCDDA-Europol, 2017).¹⁹ However, scraping data from cryptomarkets can also be unstable, leading to gaps in the data (Ball et al., 2019; Du et al., 2018; Ghosh, Porras, et al., 2017; Van Buskirk et al., 2016). Thus, we can only capture a partial picture of overall events, probably leading to underestimating the availability of products on cryptomarkets and their value.

¹⁷ If several replacement prices were available, we took the average price as the replacement.
¹⁸ An exit scam describes a situation in which the platform (market) owners steal all cryptocurrencies from all customers. On many markets it is necessary to upload cryptocurrency to an account before making a purchase. Thus, market owners have full control over the customers deposits. In some instances, the market owners also control the implemented escrow service, allowing for an even bigger exit scam.
¹⁹ For detailed timeline of the market lifespans and their reasons for closing see (EMCDDA-Europol, 2017).

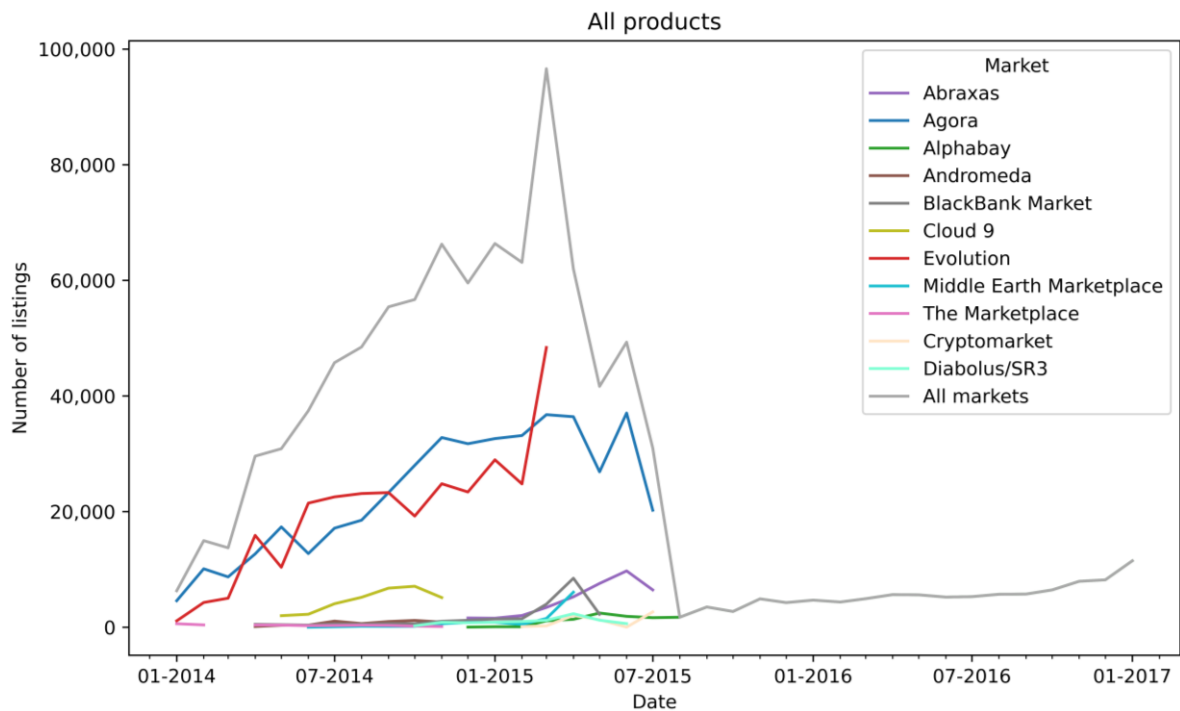


Figure 5.2. Monthly volume of products offered across markets.

Figure 5.2 also shows the monthly volume of products offered across all markets combined (grey line). Overall, product offerings seem to increase steadily, with a sharp peak at the beginning of 2015 with almost 100k listings. Offers then starkly declined, with only a few products on offer from mid to late 2015, followed by a slow increase for the remaining time, the latter solely attributed to Alphabay. To make comparisons and estimations of counterfeits across markets more comparable, we subsequently focus on the timeframe for which most markets had at least some listings on their platforms: January 2014 to September 2015.

Focusing on counterfeits (Figure 5.3), we see a similar overall trend (grey line). However, as expected, the overall number of offers is much lower, with counterfeits accounting for around 2.69% of all listings across markets. Interestingly, the observed proportion of counterfeits on cryptomarkets coincides well with the estimated overall proportion of counterfeits worldwide (3.3%) discussed above (OECD/EUIPO, 2019). Furthermore, only nine of the eleven markets seem to offer counterfeits, with Agora and Evolution offering the most, followed by BlackBank Market, Alphabay, and Middle Earth Marketplace. The remaining markets seem to harbour only a minimal number of counterfeits. Most offers seem to occur between the beginning- and mid-2015.

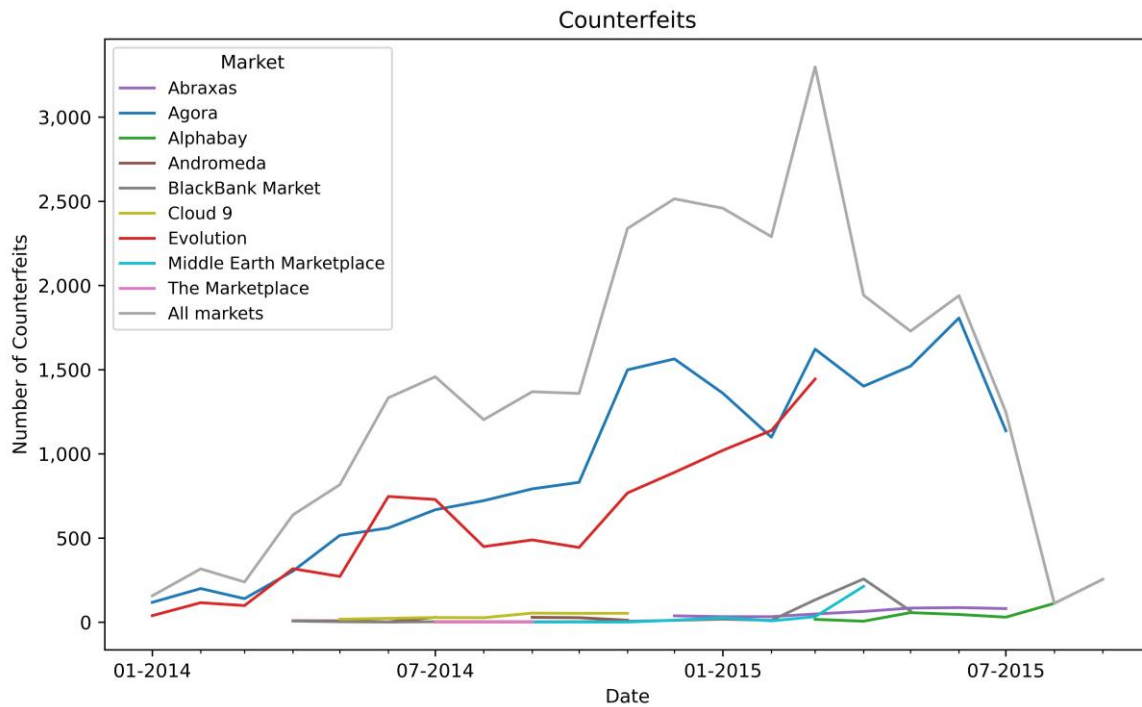


Figure 5.3. Monthly volume of counterfeit offers across markets.

5.3.2 Counterfeit product types and occurrences

Focusing on counterfeit product types (Table 5.5), we observe that watches make up most of all products (59%) listed on the markets, followed by four categories, each of which accounts for between 4-6% and collectively account for around 20% of all counterfeits. Most of the remaining categories contribute only a little, with most representing only 2% or less of all counterfeits. Thus, almost 80% of counterfeits listed were represented by only five (of the 16) categories of products.

By comparing our measures of the types of counterfeits to goods seized at borders, we can identify how products differ and discuss possible contributing factors to those differences. Based on a report by OECD/EUIPO (2019), which summarizes findings regarding seized counterfeits between 2014 and 2016, we see that not all categories represented on cryptomarkets are also present in seized goods (Table 5.5). Also, the distribution of counterfeits found on cryptomarkets and seized products varies greatly. In addition, sunglasses, handbags, and other accessories, which make up around 10% of counterfeits on cryptomarkets, are not listed individually in the report but are grouped within headgear (1.5%), miscellaneous (0.4%), and articles of leather (13.4%). The remaining categories show a similar distribution (OECD/EUIPO, 2019).

Another report by the Intellectual Property Office in the United Kingdom shows a different picture of IP and counterfeit-affected product categories (IP Crime Group, 2015). The report

summarizes independently reported IP crimes through Crimestoppers²⁰ and investigations of counterfeits by Trading Standards (TS)²¹ between 2014 and 2015. The top five reported and investigated IP crimes were tobacco, optical media, clothing, alcohol, and footwear. Although watches, jewellery, cosmetics, and electronics were also within the top 17 affected categories, they seem to be less prominent than on cryptomarkets and attracted fewer investigations by TS (Table 5.5). The differences observed for the categories tobacco, footwear, electronics, clothing, and watches, are further examined in the Discussion.

Category	Cryptomarkets	OECD/ EUIPO	IPO
Watches	59.27	5.70	1.19
Clothing	5.93	17.7	7.94
Other	5.41	-	7.41
Forgeries	4.55	-	-
Sunglasses	4.10	-	-
Electronics	3.78	12.25	0.80
Handbags	3.42	-	1.44
Other Acces.	2.93	-	-
Footwear	2.92	22.6	2.77
Jewelry	2.49	1.85	0.51
Services	1.90	-	-
Wallets	1.37	-	-
Metals	0.94	-	-
Pharma.	0.49	1.5	0.30
Cosmetics	0.26	3.5	1.10
Tobacco	0.24	-	28.15

Table 5.5. Percentage of counterfeit categories; not all categories are shared by the reports; see Appendix D6 and D7 for separate and complete lists of counterfeit categories by OECD/EUIPO and IPO.

5.3.3 Counterfeit origins

Next, we examine the shipping origins of products as indicated on the product listings. Figure 5.4 shows the percentage of shipping origins for all products and counterfeits across all markets. All countries that accounted for 1% or less are aggregated into the category other. While possible shipping destinations are included in the listing data, we did not analyse these as most destinations are listed as “Worldwide” or “Undeclared”, providing only limited information. The distribution of the shipping origins for all products seems to differ from counterfeits. However, the category undeclared takes up a considerable portion in both cases. While most products seem to originate from the USA, most counterfeits are from China, including Hong Kong. The category other contained mostly European countries (e.g., Italy, France, Poland, Portugal); it also contained a range of Asian countries (India, Thailand,

²⁰ Crimestoppers is a non-governmental organization, which allows citizens to anonymously report crimes and concerns (<https://crimestoppers-uk.org/>).

²¹ Trading Standards is the local law enforcement within the UK, investigating IP crimes and enforce consumer protection legislations (<https://www.tradingstandards.uk/>).

Singapore, Cambodia), and others (e.g., Afghanistan, Chile). The category EU (Europe) is not an aggregation we generated but was indicated on some products. Thus, for those products, we cannot say which European countries they originate from specifically.

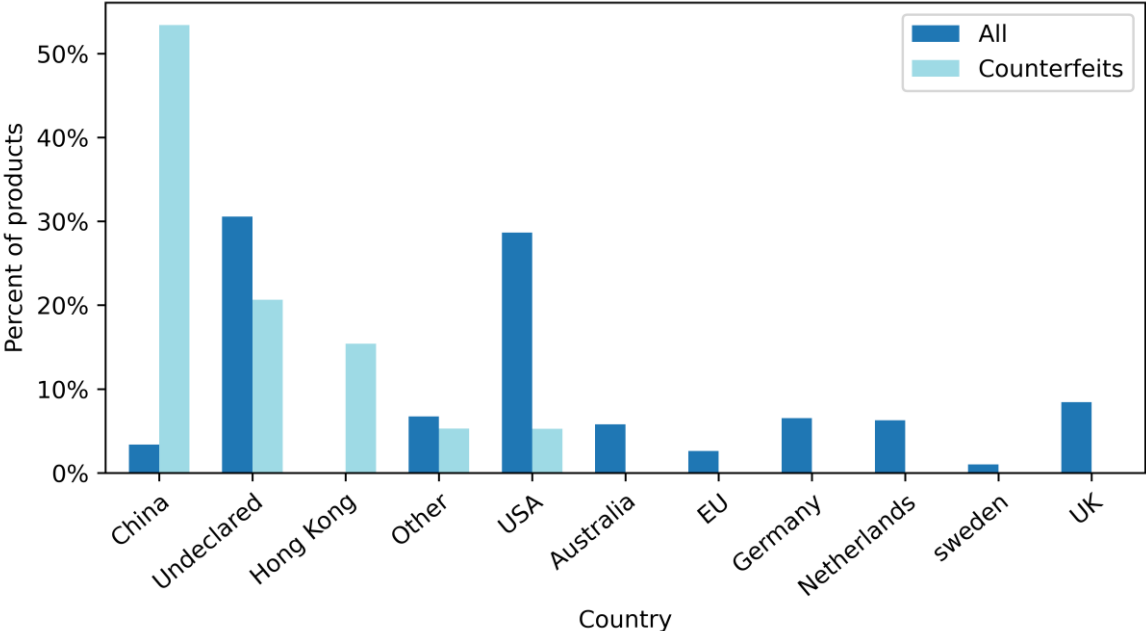


Figure 5.4. Percentage of shipping origins for all products and counterfeits.

Table 5.6 shows the association between particular types of counterfeit goods and the country they were listed as originating from. Countries that account for less than 10% of listings are aggregated into the category other. As previously indicated, China is well represented, contributing to many categories. For cosmetics, electronics, pharmaceuticals, and services, additional countries previously included in the category other are now visible and seem to specialize in supplying one particular type of counterfeit. However, for some counterfeits, the category other accounts for a substantial fraction of counterfeits indicating that in these cases, the products originate from a large number of countries. In addition, only six categories (Footwear, Clothing, Cosmetics, Pharma., Tobacco, and Watches) seem to have a rate of undeclared origins of below 20%, possibly indicating that many sellers are concerned about giving up too much information by indicating a product origin.

In contrast to the differences observed for counterfeit products seized at borders and offered on cryptomarkets, product origins seem to match better across data sources. For example, between 2014 and 2016, seized goods mainly originated from China (55%) and Hong Kong (26.2%) (EUIPO, 2019; OECD/EUIPO, 2019). However, seized goods also originated from the United Arab Emirates (3.8%), Turkey (3.1%), Singapore (2.8%), Thailand (1.4%), India (1%), and other countries (each with less than 1%) (OECD/EUIPO, 2019). In contrast, for the cryptomarkets, counterfeits were either not explicitly offered from these countries (e.g., Singapore, Thailand, India), or they accounted for less than 1% of the listings. Interestingly,

the USA seems to account for 5% of counterfeits on cryptomarkets while only accounting for 0.4% in seized goods.

Product category	China	Hong Kong	USA	AT	AU	TH	AF	BE	DE	EU	Other	Undc.
Footwear	74.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.08	0.00
Watches	69.75	22.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.23	0.00
Clothing	63.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	26.24	10.07
Jewelry	48.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.61	31.76
Sunglasses	40.81	0.00	17.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.76	20.53
Other Ac.	36.33	11.67	13.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	37.67
Electronics	23.06	0.00	25.39	0.00	0.00	12.44	0.00	0.00	0.00	0.00	14.25	24.87
Wallets	20.00	23.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.57	47.86
Handbags	11.71	0.00	14.00	0.00	0.00	0.00	11.71	0.00	0.00	0.00	10.29	52.29
Tobacco	0.00	0.00	76.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.00	0.00
Pharma.	0.00	0.00	54.00	0.00	18.00	0.00	0.00	0.00	10.00	0.00	18.00	0.00
Metals	0.00	0.00	35.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.67	47.92
Other	0.00	0.00	15.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24.95	60.04
Services	0.00	0.00	12.37	0.00	0.00	0.00	0.00	11.86	0.00	0.00	7.73	68.04
Forgeries	0.00	0.00	12.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.16	62.58
Cosmetics	0.00	0.00	0.00	59.26	0.00	0.00	0.00	0.00	0.00	11.11	14.81	14.81

Table 5.6. Percentage of counterfeit shipping origins by country and product category; percentages are split by countries and aggregate to 100% for each row; AT = Austria; AUS= Australia; TH = Thailand; AF = Afghanistan; BE = Belgium; DE = Germany; EU = Europe; Undc. = Undeclared.

5.3.4 Counterfeit prices, sales volume, and surface web prices

Lastly, we summarized counterfeit prices for each category (Table 5.7), estimated vendor sales volumes (Table 5.8 and Figure 5.5), and examined the price differences of products offered on cryptomarkets and the surface web (Table 5.9, Figure 5.6).

5.3.4.1 Observed counterfeit prices

Table 5.7 shows the prices for all counterfeit listings (offers) as customers can see them on the markets. Prices are expressed in USD and are based on all counterfeit listings at the time the listing was posted.²² The total price volume represents the accumulation of all prices from all unique counterfeits for each category (i.e., the total value if each listed item would have sold once, but only once). The total price volume of all unique counterfeits from January 2014 to September 2015 is around 1.8 million USD. Many maximum prices of each counterfeit category are high, often attributed to wholesales. The highest observed mean price is for metals, including collectible gold and silver coins or bullions, while the lowest is for sunglasses. With watches making up most listings, they also hold the highest volume, around 1 million USD. As previously discussed, minimum prices of 0.00 are mostly placeholders, and are not free products, often used to prompt the user to select an amount, colour, model, and so on (see above). The last two rows show the prices' mean, total, and weighted average.

²² All listings which contained only cryptocurrency prices were transformed into USD, utilizing the average conversion value from "https://coinmarketcap.com/currencies/bitcoin/historical-data/" on the day the listing was dated (scraping date).

Specifically, in the “Mean/Total” row, each USD column (Min, Max, Median, etc.) is averaged by dividing the sum of all product category prices by the number of product types, while solely the column “# Listings” is totalled. The weighted average is the result of taking the sum of the product of the category price and the number of listings of the same category, divided by the total number of listings. Thus, each average is weighted by the number of listings available in each product category.

Category	Price in USD (\$)					Volume	# Listings
	Min	Max	Median	Mean	SD		
Tobacco	0.00	1,401.25	0.00	110.82	315.07	2,770.52	25
Cosmetics	13.50	1,512.11	27.91	191.51	382.34	5,170.71	27
Wallets	0.00	285.04	88.69	100.69	61.86	14,096.17	140
Jewelry	0.00	856.27	38.34	57.55	114.00	14,503.22	255
Sunglasses	0.00	174.48	43.16	44.35	21.23	18,536.27	419
Pharma.	0.10	9,869.19	50.52	421.12	1,443.72	20,634.70	50
Footwear	0.00	310.01	74.22	90.46	45.82	26,322.84	299
Services	0.00	4,901.10	29.66	225.90	580.18	43,824.74	194
Handbags	0.00	1,570.34	96.04	127.10	145.00	44,358.32	350
Metals	0.00	5,413.99	161.68	524.47	1,029.88	49,824.93	96
Clothing	0.00	9,878.84	30.00	87.06	409.30	52,669.08	606
Electronics	0.00	2,988.94	84.96	165.07	271.64	63,220.44	386
Other Acc.	0.00	1,530.65	68.33	253.03	423.32	73,125.05	300
Forgeries	0.00	7,291.15	48.77	286.19	722.87	129,931.09	465
Other	0.00	5,516.25	118.51	321.20	675.91	225,805.96	553
Watches	0.00	3,957.15	100.91	171.03	178.79	1,017,309.52	6,060
Mean/Total	0.85	3,591.05	66.36	198.60	426.31	112,631.47	10,225
Weighted mean	0.04	4,013.13	87.22	174.24	262.70	633,721.53	-

Table 5.7. Summary counterfeit prices and volumes for each product category in USD; Mean = column mean prices.

5.3.4.2 Estimated counterfeit sales volumes

As in previous research (Soska & Christin, 2015; van Wegberg et al., 2018), we utilized the total number of feedback comments provided for each listing to estimate how often an item was sold.²³ That number was then multiplied by the product's listing price on the cryptomarket (PP_{CM}) to obtain an estimated sales volume in USD (Table 5.8). Table 5.8 shows that most sales were for watches, followed by “Other” (6.50%) and “Forgeries” (5.96%).

²³ Since the data was scraped recurrently, listings and their associated feedback is collected accumulative, adding old and new feedback every scrape to the data. Thus, to avoid an inflated feedback count, we only utilized unique feedbacks to every unique listing.

Category	Estimated Sales Volume (Based on feedback)		Total Feedback
	Total USD (\$)	Share (%)	
Cosmetics	16.98	0.01	1
Tobacco	52.90	0.04	2
Pharma.	579.60	0.45	20
Metals	675.71	0.53	5
Wallets	678.17	0.53	9
Footwear	1,250.46	0.98	16
Jewelry	1,464.83	1.15	31
Services	1,515.03	1.19	51
Other Acc.	1,948.04	1.53	40
Handbags	2,809.64	2.20	35
Clothing	2,936.27	2.30	35
Sunglasses	3,356.94	2.63	84
Electronics	7,176.75	5.63	55
Forgeries	7,604.31	5.96	31
Other	8,288.46	6.50	78
Watches	87,160.41	68.35	439
Total	127,514.50	100	932
Calculation	$\sum \frac{PP_{CM}}{F}$	$\frac{SV}{Total\ SV} \cdot 100$	Count

Table 5.8. Estimated sales volume (USD) for each category based on the number of feedbacks; PP_{DM} = Product Price on the cryptomarket; F = Number of feedback comments; SV = Sales volume.

Considering the monthly sales volume by category (Figure 5.5), we observed a similar trend for available listings (Figure 5.3), with most sales occurring between mid-2014 and mid-2015. We also observed two peaks in sales in mid-2014 and mid-2015. Again, watches are represented most, followed by forgeries and the category other.

Comparing these figures to seizures at borders, a report by the OECD/EUIPO (2019) found that the largest value share for goods seized at borders was for watches (22.9%), followed by leather articles (11.6%), electrical equipment & machinery (10.8%), footwear (10.5%), clothing (8.2%), jewelry (5.9%), cosmetics (4.9%), toys (4.6%), optical/photographic & medical instruments (4.1%), mechanical appliances (1.5%), vehicles (1.4%), and other products (less than 1%). Although watches seem to account for the most value in cryptomarkets *and* border seizures, the concentration of watches is much more pronounced for the cryptomarkets, with an estimated sales volume of over 68%. Thus, seized goods appear to show a more equally distributed range of values across products than is observed on cryptomarkets. Categories, such as machinery, toys, medical instruments, and appliances listed for seized goods, did not

appear to be explicitly sold on cryptomarkets, probably contributing to the skewed product distribution observed there.²⁴

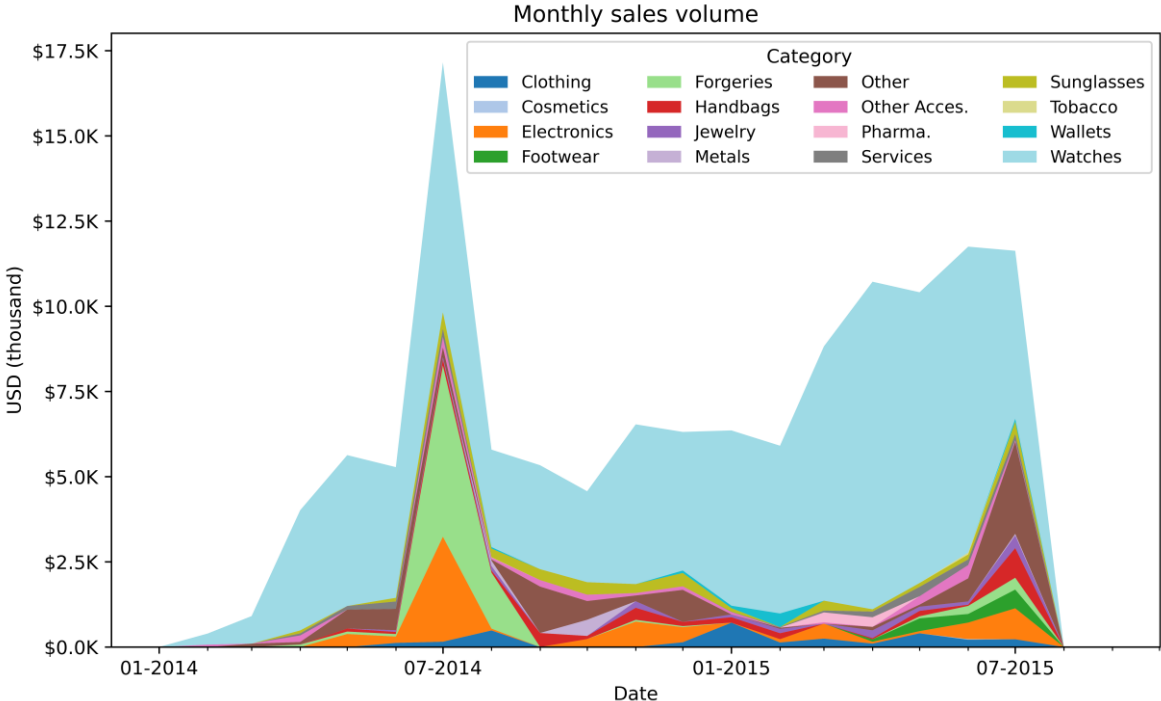


Figure 5.5. Monthly stacked sales volume (based on feedback) across categories.

5.3.4.3 Crypto and surface market prices

In addition, we sampled ten cryptomarket products from each category and determined their price on the surface web (Table 5.9). For 25 products, we determined the historical price on the surface web by utilizing a product price comparison site (geizhals.eu)²⁵ which records the complete price development over the product's lifespan.²⁶ We determined the current price using Google Shopping (shopping.google.com) for the remaining products.²⁷ When possible, we used the prices from original brand stores (e.g., Hermes, Louis Vuitton, Gucci, etc.) but selected prices from other shopping platforms if the products were not manufactured anymore or were not otherwise listed. For 46 cryptomarket products, we found the exact match on the surface web, while for the remaining listings, we selected the next best match from the same brand.²⁸ For the category metals, we adjust different indicated weights on the

²⁴ The category other contained a high variation of different products, including leather products (e.g., belts) and cars.

²⁵ Geizhals.eu collects prices from original, licensed vendors, and reselling platforms (e.g., eBay), capturing price developments after the products is not manufactured anymore.

²⁶ We determined the historic price by looking up the price of the product on the date it was listed on the cryptomarket. For four products, we found historic prices deviating from the listing date by 2-4 months.

²⁷ Google shopping shows products from original, licensed vendors, and reselling platforms (e.g., eBay).

²⁸ In some cases, the product titles from the cryptomarkets were not detailed enough (e.g., "LV wallet") to find the exact products on the surface web. Three products in the categories Cosmetics and five in Tobacco could not be found on the surface web.

listings (e.g., 10 ounces, 1 gram, 1Kg) by extrapolating the cost for 1 ounce for each listing, making a comparison possible. We excluded products from the categories services, forgeries, pharmaceuticals, and other since most of these products cannot be purchased on the surface web.²⁹ Product prices in Euro were converted into USD based on the conversion rate present on the price date. Since we selected only ten random samples for each product category, the estimated price differences are only intended to illustrate the observed trend and should not be regarded as a complete analysis.

Category	Mean [SD] USD
Cosmetics	112.82 [79.94]
Sunglasses	195.40 [116.98]
Tobacco	331.21 [433.75]
Electronics	386.07 [448.00]
Wallets	867.22 [966.92]
Metals (1oz.)	1,084.08 [755.42]
Other Acces.	1,167.03 [1,654.88]
Footwear	1,263.27 [2,767.35]
Handbags	1,351.87 [856.22]
Jewelry	1,712.02 [1,318.34]
Clothing	2,391.83 [6,712.34]
Watches	25,338.95 [52,631.70]
Mean	3,464.48 [6733.06]

Table 5.9. Mean [SD] USD of 10 sample products for each category on the surface web markets.

To better understand the relationship between cryptomarkets and surface web prices, we plot one against the other in Figure 5.6. Across all product categories, products are more expensive on the surface web, but prices between and within categories vary considerably. Prices between cryptomarkets and the surface web are closest for cosmetics (for which the mean ratio was 2.22) and most different for watches, which were, on average, 147.23 times more expensive on the open web than on cryptomarkets.

²⁹ While randomly sampling products from the category other, we encountered products such as a car code grabber, a mail list for spam, and other digital services, which are not sold on the surface web. We also encountered a counterfeit of a Picasso painting, “Seated Woman (Marie-Therese)”, priced at over USD 60 million.

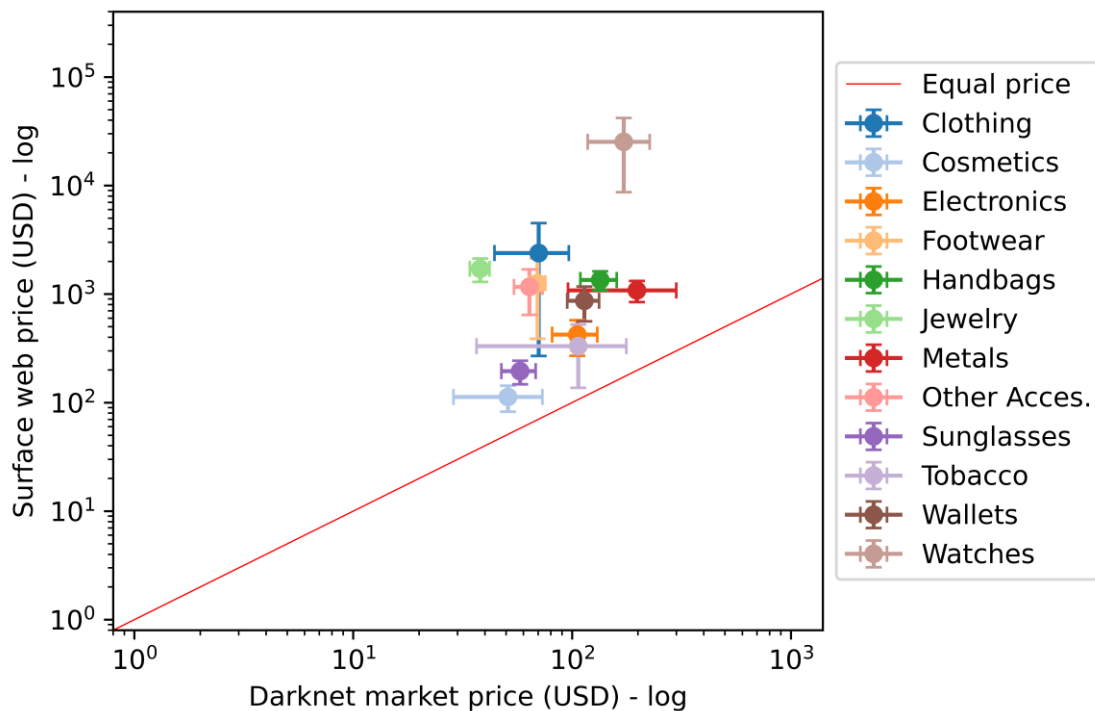


Figure 5.6. Mean (SE) price differences for each product category between Cryptomarkets and the surface web (See Appendix D8 for price differences of each product).

5.4 Discussion

Insights about counterfeits typically originate from data on goods seized at borders by law enforcement agencies. As discussed, these data are not collected through random sampling or other approaches that would ensure that the findings are representative of the ground truth. Instead, they are subject to various biases associated with the intelligence that law enforcement agencies collect or have access to, or the policies followed at borders. This means that our understanding of what is counterfeited is likely to be biased. Further indications of possible biases can be found in the prevalence estimation differences for various agencies (IP Crime Group, 2015; OECD/EUIPO, 2019). Given that IP crime is known to be increasing (Federal Bureau of Investigation, 2014, 2015, 2016; OECD/EUIPO, 2019), it is important to understand the counterfeit economy better, and in this study, we examine what insights the analysis of data regarding the availability and price of counterfeits on cryptomarkets might provide.

5.4.1 Product categories

The current study suggests that the share of counterfeits on cryptomarkets (2.69%) seems to be slightly above previous expectations, which were around 1.5-2.5% (Europol, 2017). We also see differences in some product categories observed during seizures and counterfeits offered on cryptomarkets. As already described, seized products are most likely biased through the activities and procedures adopted by authorities affecting estimations on which product types are affected. Examining the counterfeit categories, we see that watches account for most of the value in both cases but are more prominent on cryptomarkets overall. Watches might be

more challenging to identify or detect as counterfeits as other products (e.g., shoes, clothes, tobacco) in seizures, perhaps due to very high-profit margins, an increased effort is put into making fake watches more difficult to identify. Alternatively, watches might be less suitable for bulk shipments and make their way through borders differently than other items (e.g., single parcel shipments through the air versus containers at ports). Hence, watches might be shipped more diversely, possibly going through different security measures and being more difficult to catch overall. However, single parcel shipments might only be worthwhile for high-value items, such as watches, but less profitable for items that need high-volume sales.

Other strong prevalence estimation differences, such as for tobacco, footwear, electronics, and clothing, are also interesting. For some product groups, estimations from authorities are missing entirely (e.g., sunglasses, handbags, accessories, wallets, metals, tobacco). “Tobacco” in particular seems to make up only 0.24% of counterfeits on cryptomarkets, which is missing in estimations by OECD/EUIPO (2019) but is highly representative in measures by IP Crime Group (2015). Vendors on cryptomarkets might favour high-value products, possibly tailoring more towards end-consumers than other businesses. Thus, tobacco might be more difficult to sell in high volumes on cryptomarkets. Similarly, OECD/EUIPO (2019) measured relatively high ratios of footwear, clothing, and electronics, which are far less prevalent on cryptomarkets. Again, such differences might originate from biases in selecting shipments for inspections but also illustrate the current issue of inconsistent measurements capturing what is being counterfeited. Important to note is that the authorities also seized counterfeits that are missing on cryptomarkets, such as vehicles, furniture, or alcohol, which can distort the ratio of product groups (see Appendix D6 and D7 for a full lists of seized goods).

Comparing counterfeit product types on cryptomarkets and across measures (seizures, complaints, etc.) might be further complicated since vendors might tailor their counterfeits to customers who mainly purchase non-counterfeits (e.g., drugs). As a result, products might be skewed towards prestigious products, which would offer some explanations for the high number of watches, clothes and sunglasses offered on cryptomarkets. However, other relationships to surface web trends (e.g., product releases, specific product demands) could also provide some explanation for the product portfolio of counterfeits on cryptomarkets.

5.4.2 Product origins

Seized and cryptomarket counterfeits mostly seem to originate from China and Hong Kong. However, some uncertainty surrounds the information about the origins of cryptomarket counterfeits since providing this information is voluntary, and a large portion is undeclared (see Limitations). Nonetheless, the stark outlier in product origins of seized goods and product offers on cryptomarkets is the US. Around 5% of cryptomarket counterfeits were listed as originating from the USA, while only 0.4% of goods seized at borders come from the US. Again, such a discrepancy might be due to biased expectations by law enforcement, as searches are sometimes based on shipment origins (Männistö et al., 2021). Thus, cryptomarket measures suggest that border seizures might overlook counterfeits originating from certain countries,

such as the US. For example, tobacco, pharmaceuticals, metals, electronics, and accessories (e.g., sunglasses) could be scanned for counterfeits when originating from the US. Similarly, cosmetics seem to originate from Austria more frequently, and pharmaceuticals from Australia. Alternatively, counterfeits from the US might be more heavily purchased domestically, leading to limited exportation, which would avoid border controls. Moreover, cryptomarket listings represent the availability of a product rather than the actual supply of them. Although knowing which country counterfeits are available is helpful, products must be purchased first and subsequently shipped to be found at a border. Thus, estimation of product origins from cryptomarkets and measures of seized goods might also vary because they capture products at different supply chain stages.

5.4.3 Vendor sales volume and product values

Similarly, estimating the sales volume and monetary value of counterfeits on cryptomarkets is accompanied by uncertainty, which is further addressed in the next section (Limitations). However, we can see that the estimated sales volume generated for counterfeits on cryptomarkets seems very small compared to the possible value of the items on the surface web. Europol (2017) estimated that most physical counterfeits on cryptomarkets are sold for one-third of the actual price. Based on the current study, the discrepancy between counterfeit prices and their actual values on the surface web are more diverse and can be twenty times larger (e.g., for watches). Such differences suggest that the prices and possible sales volumes depend highly on the product category. However, the current price differences illustrate that purchasing cryptomarket counterfeits and selling them on the surface web could lead to considerable profits. Thus, it might be helpful to focus the attention of authorities on highly valuable counterfeits, such as watches, clothing, or jewelry, as they seem to generate the biggest profits. Notably, relative to the patterns observed for cryptomarkets, watches were underrepresented in the estimates based on seizures, and metals were not featured at all.

We can also see greater differences between anonymity network and surface web prices for higher-value products, such as watches, clothes, and jewelry. Cryptomarket vendors might prioritize higher-valued products, which can generate profits faster than products with lower profit margins (e.g., accessories, tobacco). Such a strategy would support previous ideas that cryptomarket vendors might tailor their products more towards end-consumers (Europol, 2017), who would purchase at a lower volume but with higher frequency rather than businesses, which could purchase items in high volumes with the purpose of re-selling them. In other words, lower profit margin products need higher turnovers for high profits, which is facilitated by business-to-business transactions.

5.4.4 Possible preventative measures

Since the counterfeits identified here were fully manufactured consumer products, for them to enter the supply chains of legitimate retailers, the latter would either have to sell them knowingly or they would have to be introduced during distributional processes so that they

are mixed with genuine products without the retailer's knowledge. Counterfeits could be introduced during packaging, distribution to wholesalers, retailers, or any other transportation process. As Hollis & Wilson (2014) discuss, addressing the problem in cases where companies have been misled would involve improvements to guardianship in risky parts of the supply chain. Companies could be provided with information about which products are affected and from which country they originate to facilitate their efforts to identify risks in their supply chain. For example, informing personnel who are responsible for overseeing the distribution of an affected product (which is being counterfeited) could help them to implement or re-evaluate their internal working processes to reduce the risk of counterfeits entering their supply chain and increasing the risk of discovery for the counterfeiters (Hollis & Wilson, 2014). Such implementations could include raising employee awareness of the affected products, implementing reporting mechanisms, or introducing additional validation checks for particular product types for specified periods of time. To aid in this activity, cryptomarket data – searchable by brand – could be made accessible to companies. Since product information is quite detailed, an implementation with up-to-date cryptomarket data is feasible.

Another issue concerns the leaking of product designs. One approach to help address this would involve the identification of products that are found to be offered on cryptomarkets before their official release on the surface web. Knowing that plans were shared would help companies narrow down which processes would have to be reviewed and where measures should be put in place to ensure adequate guardianship. Such measures might involve limiting access to project plans to only those who need to know about them (to minimise insider threats) and ensuring that all data are secure (to minimise external threats). While some cyber security and brand protection organizations advertise anonymity network monitoring to detect data leakages, such as personal data, to what extent they track counterfeits is unclear (Corsearch, 2023; Lenaerts-Bergmans, 2023).

Other approaches to counterfeiting might involve one or more of the 25 techniques of situational crime prevention (Clarke, 1995; Freilich & Newman, 2018) discussed in Chapter Chapter 2. One such technique is target hardening, which aims to make the target of an offence (e.g. counterfeiting a product) less viable for the offender. Knowing which counterfeits are offered on cryptomarkets could help companies to make those products more difficult to counterfeit. For example, companies could change the materials used or the manufacturing process to increase the efforts of imitating the product. Traceability of genuine products within a supply chain would also fall within that category, as it increases the efforts needed to counterfeit them, which could be technologically facilitated (Gayialis et al., 2022). Alternatively, the offenders' rationalisation for committing a crime could be challenged by removing possible excuses for their actions. Removing excuses includes approaches such as setting up rules or posting instructions to reduce ambiguity in situations that can be exploited. Such strategies could be helpful to deter employees in situations in which they could act maliciously (stealing plans, reintroducing counterfeits, sharing manufacturing or packaging

plans, etc.) by reminding them what actions are disallowed or how specific work tasks should be performed (Freilich & Newman, 2018).

5.4.5 Future studies

Given the results of this study, it would be interesting to examine if and how such information about counterfeits on cryptomarkets can be utilized as intelligence for law enforcement activities or policymakers. Besides validating findings from seized goods, cryptomarkets could serve as indicators of early trends for the onset of activities on the surface web. For example, future work could establish a monitoring system that collects counterfeit data from cryptomarkets. Such a dataset could be used to search for cryptomarket products on the surface web (e.g., Amazon, eBay) to establish if the same or similar products are sold across platforms. Furthermore, a longitudinal study could explore temporal trends, particularly if products tend to appear first on anonymity networks and subsequently find their way to the surface web. A common problem with research concerning cryptomarkets is the accurate estimation of sales value. While utilizing customer feedback to make informed estimations is helpful, future work could explore if it is possible to exploit transactional data associated with cryptocurrencies with users from cryptomarkets (Chen et al., 2021; ElBahrawy et al., 2020; Nadini et al., 2021) to complement sales assessments of specific products. Future work could also examine vendors and analyse whether they operate across markets (e.g., through PGP keys), how their profile changes over time (e.g., number of listings), and which non-fraud related listings they offer. Thus, exploring to what extent vendors specialize or diversify their portfolio.

5.4.6 Limitations

This study used a large set of historical data on cryptomarkets. Although the included data covers around two years, some bigger markets, such as the first Silkroad, Hydra, Empire, Hansa, Wall Street, and Sheep, were missing. The reasons for their exclusion were that they were not included in the data archive or lacked sufficient product categorization needed for the current analyses. The data also does not cover possible user-to-user transactions, which bypasses the markets altogether (Nadini et al., 2021). Thus, the findings reported here do not reflect the entire cryptomarket economy, just the activity recorded for those markets sampled. Furthermore, the present analyses utilized historical data without newer scrapes (see ElBahrawy et al., 2020), potentially limiting the current policy or prevention implications. However, previous work has not provided us with an understanding of how extensive counterfeits are present on cryptomarkets and re-using existing data in the current study serves as a proof of concept, showing that cryptomarket data can be valuable in understanding the counterfeit economy better. Thus, showing how newer data could be utilized for counterfeit research.

Furthermore, a general problem related to research with cryptomarkets is that the data collection procedure is constrained by the scraping process and individual platform closures,

which can lead to gaps in the available data, which can make exact measurements of cryptomarket activity difficult. The scraping process can be disrupted due to the slow connection of the Tor network, security measures of the website that are implemented to hinder automated data collection (e.g., required log-ins due to set session time-outs or solving recurring captchas), or temporary website closures. Thus, the overall number of observed listings and associated estimations will be more uncertain, making general conclusions more difficult.

While comparing seized counterfeits to cryptomarkets counterfeits can help us understand how the two areas relate to each other, the comparison is only partly applicable. Cryptomarket listings are offers, while seized products may already have been sold. Although seized products can also inform us about offers, they are only a subset of sold counterfeits from the overall market. Thus, comparisons of cryptomarket listings with seized goods are informative, but they do not always encompass the same measures.

Similarly, uncertainties are present with shipping information and feedback associated with cryptomarket listings. It is voluntary for vendors to make product origin declarations, and many choose not to do so. Nonetheless, many declared origins are in line with the origins of seized goods, providing us with some confidence in our measures. Also, information on postage times and possible tracking numbers is highly valued amongst customers, often referred to in feedback, making a genuine declaration of origin more attractive to vendors. Therefore, we cannot say how accurate product origin declarations are, but some incentives exist for vendors to make truthful indications.

As for product feedback, we cannot always know whether they are mandatory and whether the feedback is for a single or bulk purchase. Thus, the calculated sale volumes are approximations and will come with a general uncertainty because not all purchases will have produced feedback, one instance of feedback might be counted as a single purchase, or feedback could be artificially created to generate trust (Dellarocas, 2006). Similarly, our value estimation process should be taken with caution. Taking ten random samples for each product category will produce only rough estimates and was only intended to illustrate the estimated difference between prices on cryptomarkets and the surface web. Furthermore, a historic price could not be obtained for all product samples, and prices can vary considerably over time (e.g., original soccer shirts or Nike shoes), influencing estimations.

5.5 Conclusion

Based on the analyzed cryptomarket data, we can say that counterfeit goods are rare (2.69% of all products) on cryptomarkets and are often included in miscellaneous categories. Thus, accurately measuring the prevalence of counterfeits across anonymity networks is difficult. However, we disentangled product categories using a classification model, allowing for a more in-depth analysis. We showed that some product types exhibit a strong prevalence discrepancy between cryptomarkets and seized goods. Specifically, watches are more

prominent on cryptomarkets, while electronics, shoes, clothes, and tobacco are more prevalent among seized goods. Furthermore, vendors seem to favor high-value products with big profit margins (e.g., watches) instead of products for which higher turnovers are necessary (e.g., tobacco) to obtain the same revenues. Interestingly, we found some similarities in shipping origins between cryptomarkets and seized goods, with some exceptions, such as relatively high origin shares from the US in cryptomarket counterfeits.

While the study is based on historical data, we showed that examining cryptomarket counterfeits in more detail can contribute to our understanding of the counterfeit market. Thus, looking at current cryptomarket data would be valuable in future analyses of IP crime, which would provide us with more up-to-date insights. Collecting data from cryptomarkets to gather intelligence could be done manually and automatically and would probably be very cost-effective compared to (border) seizures. Once implemented, prolonged data collection could be easily maintained, providing us with regular details on counterfeits. Such information would be usable by authorities and businesses, informing them which products are currently affected.

Chapter 6: From anonymity networks to the surface web: Scouting eBay for counterfeits

Fabian Plum contributed to this chapter by conducting all image related analyses, which are described in section 6.4 (Image similarities).

6.1 Introduction

Big online shopping platforms, such as eBay, Amazon, or Alibaba, as well as social media platforms, including Instagram and Facebook, struggle to deal with the increase in counterfeit sales on their platforms (BBC, 2015; Conlon, 2017; Ihaza, 2017; Mooij, 2018; Scheck, 2019; Suthivarakom, 2020). Counterfeits are physical or digital goods that violate intellectual property (IP) rights (e.g., copyrights, trademarks, patents) (OECD/EUIPO, 2019; WTO, 1994), and current measures seem insufficient to deter counterfeit sales (Duhigg, 2019; Zimmerman, 2020). With the increase in counterfeit sales, which are detrimental to brand values and can hurt customers financially and physically (EMCDDA-Europol, 2017), the reliable and efficient detection of counterfeits on online shopping platforms has become increasingly important.

Aside from surface web markets (i.e., eBay, Amazon), counterfeits are also sold on cryptomarkets – online shopping platforms on the deep web – which have received increasing attention from the research community and law enforcement (Baravalle & Lee, 2018; Christin, 2013; Europol, 2017; Ghosh, Porras, et al., 2017; Van Buskirk et al., 2016). Most markets utilize The Onion Router (Tor) network, which directs internet traffic through a relay network, ensuring a high degree of anonymity for both vendors and customers (Çalışkan et al., 2015; Gehl, 2018; The Tor Project, Inc., 2020). The most commonly sold goods are drugs (Rhumorbarbe et al., 2016; Soska & Christin, 2015), but cryptomarkets also offer weapons, phishing information, hacking services, counterfeits, and more (Douglas, 2015; D. L. Roberts & Hernandez-Castro, 2017; van Wegberg et al., 2018). Apart from counterfeits, almost none of the products and services offered on cryptomarkets would be offered on online shopping platforms on the surface web (e.g., Amazon, eBay). While counterfeits are typically sold (deceptively) as genuine products on the surface web, they are sold openly as counterfeits on anonymity networks. Given that counterfeits are present on both cryptomarkets and surface web markets, it seems plausible that activity on both markets might be interdependent. Although previous reports on counterfeits on cryptomarkets have noted that sellers seem to operate across anonymity networks and the surface web (EMCDDA-Europol, 2017; Europol, 2017), the extent to which this is the case and for which products this applies is unknown. Similarly, individuals purchasing counterfeits on cryptomarkets might resell the items for profit on surface web platforms, or vendors on cryptomarkets might conduct market research on surface web platforms (e.g., forums, shopping platforms) to find out which products are in high demand to determine which products they should offer.

Because counterfeits on cryptomarkets are sold openly, this presents an opportunity to use the available information (product names, descriptions, pictures, etc.) to search for matching listings on the surface web. In this chapter, we explore how an automated search for counterfeits on eBay, based on current cryptomarket counterfeits, might work. By utilizing text and image similarity metrics between anonymity network and surface web listings, as well as a ranking system of the similarity scores, we determine the best matches, which could subsequently be prioritized for manual inspection. Although the methods to produce single similarity scores used in this study are not novel, we are not aware of any work utilizing them to process information from cryptomarket counterfeits to search for potential counterfeits on the surface web. Since authorities can often only react to incidents of fraud or are faced with intensive manual investigative web searches to find counterfeits (FBI, 2018), the goal of the proposed system is to demonstrate how manual searches for counterfeits could be supported with automated approaches to become more efficient.

In what follows, we collect eBay product information with cryptomarket counterfeit product names as search queries at two separate points in time that are four months apart to enable an examination of how offers change over time. We then determine similarities between cryptomarket and eBay listings by merging automatically generated image and text similarities and human-annotated similarity scores. Product matches are then ranked by similarity and manually inspected to determine if we can find the same products across anonymity networks and surface web. Thus, this study aims to explore how manual time-intensive investigations for finding counterfeits on the surface web could be supported through partial automation of the search process.

6.1.1 Motivation for the current system

Many fraud and deception detection methods use supervised classification systems (Abdallah et al., 2016; Almendra & Enachescu, 2012; Hernandez-Castro & Roberts, 2015; Sahingoz et al., 2019). Supervised classifiers automatically infer from labelled data how to distinguish the features or attributes of the data into pre-defined classes (Murphy, 2012). This would include, for example, learning how to differentiate fake and genuine product reviews based on the words used in descriptions of the products. However, for several reasons, a classical supervised approach would not work.

In our use case, we have no precise knowledge about the listing's attributes or features (product name, descriptions, etc.) of the products we are looking for. In other words, we have no ground truth data we could use to train a supervised model. With that, we also do not know how many or if any products on cryptomarkets can be found on the surface web. Lastly, supervised classification systems tend to only work well within the domain they are trained in and do not generalise well (Geirhos et al., 2020). Thus, the performance of a classifier, which is trained to detect fake watches, might not be transferable to other products, such as shoes, mobiles, or clothes. Given the reasons above, we devise a method that does not employ a

supervised classifier but can work across various products, is not reliant on annotated data and circumvents the problem of not knowing the true distribution of counterfeits.

6.2 Data

To test whether we can find potential counterfeits on eBay, we first collected cryptomarket (CM) listings of counterfeits, which we subsequently searched for on eBay. We manually collected 453 CM listings on 25 June 2021, while eBay data was collected twice with a time delay of around four months. This delay was used to examine if we could identify a change in product listing occurrences over time. 66,430 eBay listings were collected automatically between 28 June and 1 July 2021 (period 1), and another 68,532 listings were collected between 5 and 9 November 2021 (period 2). During each eBay data scrape, we obtained text data and image links, which we used to collect high-resolution images separately. Images were collected after all text data were fully scraped, which took around ten days for each period. Specific collection procedures and descriptive statistics are detailed in sections 6.2.1-6.2.3 below, and all of the scraped data are available upon request. The study and data collection were approved by the Ethics committee of the Department of Security and Crime Science, University College London.

6.2.1 Cryptomarket data

Using the Tor browser³⁰, we visited 12 cryptomarkets listed on “www.dark.fail”³¹ and determined if they contained counterfeits. From all the visited markets, we chose four markets (Darkode, Torrez, White House Market, World Market) because they contained the most counterfeits, which were also explicitly categorized as counterfeits. Most markets required registration with a username and password, and some required an additional set of PGP (Pretty Good Privacy) keys³². The information on cryptomarket counterfeits was collected manually by saving the web page and the associated images to a local machine. Automated data collection from anonymity networks can be challenging and is known to lead to data gaps (Ball et al., 2019; Du et al., 2018; Van Buskirk et al., 2015). Since we only needed a small amount of cryptomarket data, we favoured a manual approach to ensure reliability. We were regularly prompted to solve Captchas during logins and data collection, indicating that many markets implement anti-scraping measures.³³ The saved HTML pages were later parsed using the Python package “selectorlib” (Rajeev, 2019) to capture the listing information in a structured way, which is usable by data science methods. To increase the likelihood of finding

³⁰ We changed the settings to “most secure” and disabled JavaScript to limit our vulnerabilities against potential malicious attacks (e.g., phishing). Before starting the browser, we enabled a VPN to further secure our privacy.

³¹ A platform listing verified onion pages and displaying if they are currently online.

³² PGP keys enable secure communications between two individuals and are often required for any transactions on anonymity networks as well as determining the validity of the webpage you are visiting. It seems that some onion domains from markets are compromised and are utilized as phishing platforms.

³³ Some Captchas were already prompted by opening too many browser tabs or clicking through the webpage too fast.

current matches on the surface web, we collected information from the first 50 products (when available) from all counterfeit categories on the cryptomarkets, sorted by newest to oldest whenever possible. Some categories contained less than 50 products. As shown in Table 6.1, across the four markets, we collected information for 453 counterfeits, which covered five categories of products.

Category	Cryptomarket				Total	%
	Darkode	Torrez	W. H. Market	World Market		
Clothes	-	44	-	-	44	9.71
Electronics	-	1	28	-	29	6.40
Jewellery	-	44	50	50	144	31.79
Other	34	-	35	29	98	21.63
Watches	50	39	49	-	138	30.46
Total	84	128	162	79	453	100.00

Table 6.1. Counterfeits for each market across categories. The category “Other” contains products such as wallets, (hand) bags, and hats, but items that are also found in existing categories, such as shirts, shoes, or jewellery, as categorization varied by markets.

For each CM product, we collected the vendor’s name and additional vendor details (when available), as well as the product title, description, price (USD), origin, and (possible) shipping destination(s). Furthermore, we collected 1,488 images from these cryptomarket counterfeit listings, with the majority containing around four images (Mean=3.54, Median=4, SD=1.25, range=1-5). Figure 6.1 shows an example listing scraped from Darkode.

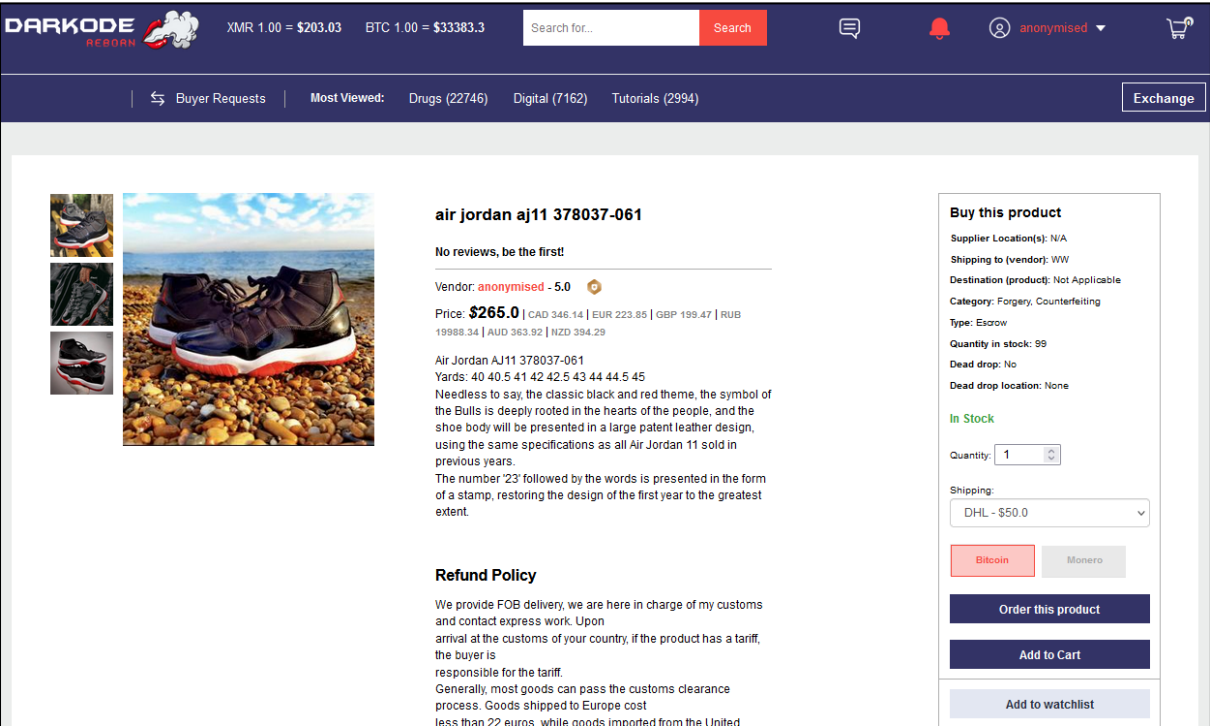


Figure 6.1 Screenshot of a counterfeit listing on the CM Darkode.

6.2.2 eBay data

To collect eBay data, we automated a product search based on the cryptomarket counterfeit product names and scraped the first page of the results. This equated to approximately 200 listings per search.³⁴ When generating the search terms, we manually removed any words that would indicate that a product was a counterfeit (e.g., “fake”, “replica”, “counterfeit”, and “forgery”) as we would not expect such terms to be included in the open web adverts. To scrape eBay product information, we used the Python package “selectorlib” (Rajeev, 2019). We used “www.eBay.com” to search for and collect product information. The local machine we used was in Europe, but we used a VPN (Virtual Private Network) set to the USA to prioritize listings in the USA. All scraped information was publicly accessible, obtained without an account, and is in line with eBay’s Terms of Service, ensuring that no information was collected outside eBay’s guidelines. The data collection procedure involved three steps, which were the same for the first and second eBay scrape:

- I. Searching for products on eBay using the CM counterfeit product names and scraping the product links of the first 200 results.³⁵
- II. Scraping detailed product information from the previously obtained product links and obtaining the associated image links.
- III. Scraping all images using their associated image links but altering them to obtain images in their native resolution. The number of images was limited to a maximum of 10 per listing.

For the first eBay scrape (period 1), we collected 66,430 listings and found, on average, 156 listings for each search (i.e., for every CM listing product name) (Median=200, SD=79, range=1-221). For the second eBay scrape (period 2), we collected 68,532 listings and found, on average, 162 listings for each search (Median=200, SD=69.85, range=2-252). For period 1, for 27 CM products, no eBay products were found, and for period 2, no eBay products were found for 29 CM products.

For each eBay listing, we collected the product title, specifics (e.g., height, weight, etc.), descriptions, price (USD)³⁶, origin, (possible) shipping destination(s), return policy, condition, assurances³⁷, as well as the vendor’s name, feedback score, and the positive percentage of the feedback score (aggregated by eBay). On average, we scraped 7 images for each eBay listing (median=7, SD=2.72, range=1-10) and obtained a total of 935,100 unique images across both scrapes.

³⁴ The eBay search result page was set to show the maximum of 200 items per page.

³⁵ When eBay could not find exact matches for the search query, it reduces the number of words for the query and searches again.

³⁶ The price was in some cases the starting bid.

³⁷ Policies such as a money back guarantee, usage of escrow, authenticity check etc.

6.2.3 Cryptomarkets and eBay descriptive statistics

6.2.3.1 Product origins and destinations

Figure 6.2 shows the product origins for CM and eBay products for the first and second scrape periods. In line with other findings (Europol, 2017), most counterfeits seem to have originated in Hong Kong and China. Interestingly, the UK is the third most frequently identified country. In contrast to the trends observed for cryptomarket products, most eBay listings indicated the USA as the origin, which is expected due to the VPN settings, followed by Japan. China was far less represented in eBay listings (6%). Most product destinations were indicated as “worldwide” or were undeclared (CMs), providing only limited information.

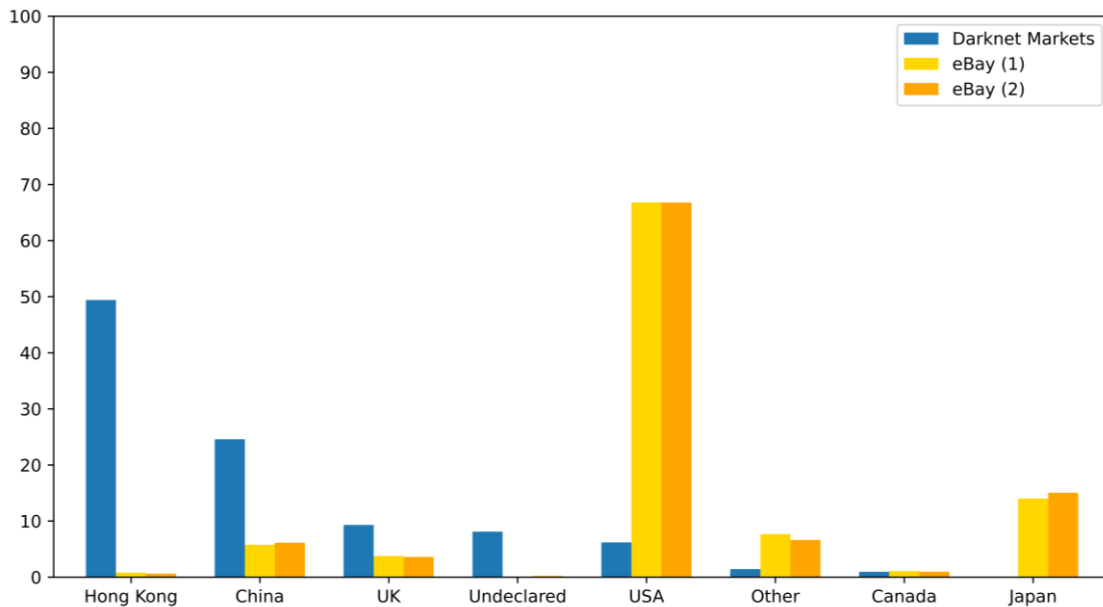


Figure 6.2 Percentage of product origins in percentage. Origins contributing less than 1% of the total number of products are aggregated into “Other”.

6.2.3.2 Prices

Table 6.2 shows descriptive statistics for the price of CM and eBay products. CM product prices vary considerably, ranging from 15 to over 6,000 USD, with electronic products having the highest average prices. Most eBay listings were for immediate purchase. However, customers could occasionally bid on them for which a minimum price was advertised (e.g., 1 cent). Similar to CM products, most categories contained some eBay products with extremely high prices of up to 999k USD for particular watches.

Category	Price (USD)										Listings	
	min		max		median		mean		SD		CM	eBay
	CM	eBay	CM	eBay	CM	eBay	CM	eBay	CM	eBay		
Clothes	67	0.01	290	124,999	119	35	145	175	63	1,488	44	15,043
Electronics	140	0.01	6,250	13,000	1,275	550	1,687	689	1,837	752	29	7,807
Jewellery	40	0.01	3,215	650,000	75	350	198	5,613	387	1,9627	144	42,355
Other	15	0.01	5,000	45,000	150	196	279	464	625	982	98	33,493
Watches	230	1.63	1,480	999,999	410	3,703	488	17,829	242	55,456	138	36,264
Average/Total	98	0.33	3,2467	366,600	406	967	559	4,954	631	15,661	453	134,962

Table 6.2. Product price distributions across markets and categories.

6.3 Similarity metrics

To identify products on eBay that resemble the counterfeits sold on the CM, we calculated similarity scores between the listings. Specifically, we calculated text and image similarity scores between the CM listings and their associated eBay search results to capture different aspects of similarity.

6.3.1 Text similarities

To compare the title and product descriptions of CM and eBay listings, we calculated four different similarity scores between the titles and the descriptions (see Appendix E1 for descriptive statistics on word occurrences in the texts). Specifically, the *Word Mover Distance* (Kusner et al., 2015) and three *cosine similarities*.³⁸ Since eBay listings contained three types of product descriptions (item specifics, description by eBay (1), and by the seller (2)), we merged them into one text before calculating any similarity score. For all text similarity metrics and any text pre-processing steps, we used the Python package “spaCy” (Honnibal & Montani, 2017).

6.3.1.1 Word Mover Distance

Calculation of the Word Mover Distance (WMD) requires the use of word embeddings. These represent words in a vector space, in which semantically similar words are closer to each other than semantically dissimilar words (Jurafsky & Martin, 2019). We used the pre-trained Word2Vec embedding space trained on the Google News dataset to create word embeddings for each document (e.g., title, description) and used the Python package “genism” to calculate the WMD (Kusner et al., 2015; Pele & Werman, 2008, 2009). The WMD score indicates the minimum cumulative (Euclidian) distance the word embeddings of document A have to travel to the word embeddings of document B within the embedding space (Kusner et al., 2015). A WMD score of 0 would indicate no distance between the compared documents, indicating the highest similarity. Any WMD score greater than 0 indicates a distance, with greater scores indicating a larger distance between the documents and hence less similarity between them. Before calculating the WMD, we removed all stop words³⁹ from each document and made all text lower case.

6.3.1.2 Cosine Similarity

Q-grams: Q-grams are character-based strings of length q (Ukkonen, 1992). In our case, we decided to split each document into character lengths of 3 (e.g., words such as “good” into

³⁸ We also generated additional similarity metrics, such as the Levenshtein distance (Levenshtein & others, 1966) or the Jaccard index (Jaccard, 1912; Rissola et al., 2020), based on q-grams and document embeddings, but removed them since they showed strong correlations (>0.7) with other similarity measures.

³⁹ We use the English stopwords provided by “spaCy”.

“goo” and “ood”). By creating q-gram frequency vectors for each document, we can calculate the cosine distance between the documents.

S-BERT: Next to the Euclidian distance in the WMD, we also calculated the cosine similarity between document embeddings, instead of word embeddings, using the neural network language model Sentence-BERT (Reimers & Gurevych, 2019), a modification from BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). For our use-case, we choose the model instance “paraphrase-MiniLM-L6-v2”⁴⁰, which is finetuned with a collection of 12 datasets⁴¹, because it scores highly on several benchmark datasets⁴² and prioritizes fast processing. Before creating the document embeddings, we made the text lowercase but omitted other pre-processing steps, such as stop word removal, as the model is trained on unchanged texts.

Universal Sentence Encoder: Lastly, we calculated the cosine similarity between document embeddings generated with the Universal Sentence Encoder (Cer et al., 2018). For our use case, we utilized a pre-trained model that used the deep averaging network (DAN) architecture. The model was trained on various texts from Wikipedia, other web resources⁴³, and the Stanford Natural Language Inference corpus (Bowman et al., 2015). We implemented version 4 of the model with TensorFlow (Abadi et al., 2016).⁴⁴ Similar to generating S-BERT embeddings, we first made all text lowercase but omitted other pre-processing steps, such as stop word removal.

6.4 Image similarities

We combined several comparison methods to compute image similarity metrics, including colour histogram correlations with noise removal, different feature extractor and matching algorithms, and a custom-built and trained Siamese deep neural network (discussed in detail below). We produced similarity scores for every image of each cryptomarket listing (1588 cryptomarket images in total) compared to every image of each surface web listing of the respective search query, resulting in a total of ~3.5 million comparisons with five similarity scores each. The majority of scores are computed with functions native to the “OpenCV” Python package (Bradski, 2000). We decided to select only the maximum scores obtained for any of the image comparisons for each metric.

⁴⁰ Model card for all parameters: <https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

⁴¹ The authors of “sentence-transformers” provide a range of models, finetuned to specific use cases. For more detailed information about model types and their associated performances, see: https://www.sbert.net/docs/pretrained_models.html

⁴² Benchmark comparisons: https://www.sbert.net/_static/html/models_en_sentence_embeddings.html

⁴³ See Cer et al. (2018) for details on the training corpus

⁴⁴ For full details on the model and its benchmark performances see: <https://tfhub.dev/google/universal-sentence-encoder/4>

6.4.1.1 Image pre-processing

Images from cryptomarkets came in various resolutions and file formats and were therefore rescaled to a maximum resolution of 1024 x 1024 pixels to preserve aspect ratios, facilitate faster processing, and encourage comparable feature sizes (see Appendix E2 for more details). Some of the methods employed to compare image content are susceptible to changes in aspect ratio. Therefore, non-square images were padded (i.e., adding black borders either vertically or horizontally).⁴⁵ All files were stored in .jpg format with lossless compression to minimize the influence of downsampling and compression artefacts.

The images associated with the surface web (eBay) were relatively small and were, therefore, not downsampled for low-level comparisons (i.e., colour histogram comparison and feature matching of various descriptors), as detailed below. However, the employed mid-level comparison method using a custom-built Siamese network architecture, implemented in Keras, required equisized width and height of images for both inputs. Therefore, we downsampled and padded surface web images to 299 x 299 pixels, using the same method as for cryptomarket images.

6.4.1.2 Colour histograms

As a first low-level measure of similarity, we compared colour histograms of the cryptomarket and eBay images, based on the colour histograms of bilaterally blurred images (Tomasi & Manduchi, 1998) with a kernel size of 5×5 to counter the influence of image noise and compression artifacts on resulting similarity scores. Histograms are produced for each RGB colour channel and normalized before comparison via histogram correlation.

6.4.1.3 Feature detection & matching

We compared the cryptomarket and eBay images by matching feature descriptors generated by several common feature detectors, namely *Scale Invariant Feature Transform* (SIFT) (Lowe, 1999, 2004), *Speeded Up and Robust Features* (SURF) (Bay et al., 2008) and *Oriented fast and Rotated Brief* (ORB) (Rublee et al., 2011). These detectors are used to extract *key points*, also known as *features*, consisting of visually distinct image elements, such as corners, blobs, and edges, and are described in their relation to neighbouring pixels. Such features can be subsequently matched between different images, where different feature detectors are distinguished mainly by their processing speed and their ability to handle changes in scale, rotation, distortion, and illumination (Karami et al., 2017; Tareen & Saleem, 2018).

In the performed similarity comparisons, 1000 features were extracted from each image and brute-force matched by L2 distance, which produced the two highest-ranking matches per feature. Likely matches were then filtered, as suggested by Lowe (2004). To rank extracted descriptors by their likelihood, we used the *Fast Library for Approximate Nearest Neighbors*

⁴⁵ Since the padded space consists of null values, they will not affect the calculation of any image metrics.

(FLANN) (Muja & Lowe, 2011). We then divided the number of likely matches by the number of extracted features to normalize similarity scores for each image pair. In some rare cases of highly repetitive textured backgrounds, the number of repeatedly matched features can exceed the number of extracted features, resulting in scores higher than 1.

6.4.1.4 Siamese neural network

A Siamese neural network is a type of deep neural network architecture, usually trained as a classifier. It consists of two convolutional network pathways with shared weights and a fully connected network head to produce a binary classification, inferring whether the contents presented in both input images share the same identity. While functionally similar, our architecture, designed to produce similarity scores, slightly diverges from this convention in the sense that the convolutional pathways are two identical Inception v3 networks (Szegedy et al., 2015, 2016) with frozen weights, pre-trained on the ImageNet (Deng et al., 2009) dataset. They function as feature extractors, and their output is fed into a set of fully connected layers trained on our custom dataset. A depiction of the information flow within the architecture can be found in Appendix E3.

An additional challenge of our dataset is that there is no ground truth for likely matches between cryptomarket listings and corresponding eBay listings. However, we require an estimate of truly matching product image pairs to train the Siamese neural network. Therefore, the network was instead trained by associating all images of the same listing as a likely match and all other images as unlikely matches. So-called “triplets” were formed from two images, sampled from cryptomarket and eBay listings, and a binary label indicating whether they belonged to the same listing. We define a positive triplet as any two images of the same product and a negative sample as any two images of different products across all groups. As we cannot anticipate the true distribution of matches between known counterfeit and surface web listings, to train the classifier in a balanced manner, we extract the maximum number of positive samples and an equal number of negative images by randomly drawing image pairs from both datasets. This results in a total of 3,393,154 triplets using 80% to train the model, withholding 20% for validation purposes. The fully connected network head of the Siamese architecture was then trained using the ADAM optimizer (Kingma & Ba, 2014) over a total of one million iterations with a batch size of 16, resulting in a classification accuracy of 82.95%. To compute the final similarity scores, instead of noting the binarized output of the network, we use the SoftMax output of both output nodes as normalized measures for similarity. Thus, we received two output values in the form of the activation of the output nodes, indicating the network’s prediction and confidence regarding whether the input images were of the same identity. On the one hand, a high activation (values close to 1) of the first output node and a low activation (values close to 0) of the second node indicate high confidence in the images being of the same identity. On the other hand, a low activation of the first node and a high activation of the second node indicates the opposite. If the activation of either or both nodes is close to 0.5, the confidence in the prediction is low. To simplify the integration of the two scores with the other metrics, we combined them by subtracting the

dissimilarity from the similarity score leading to a single score ranging from -1 (low similarity, high confidence), over 0 (low confidence), to 1 (high similarity, high confidence).

6.5 Determining similarity between product pairs

While generating multiple metrics helps capture different aspects of similarities, incorporating them into a single meaningful score poses an additional challenge. Instead of simply adding all similarity scores together, we determined for each metric an individual weight. To do so, we asked crowd workers to annotate 1000 listing pairs, rating their overall similarity (rating 1 to 7), and ran a regression analysis with the similarity metrics as the independent variables and the human-annotated similarity score as the dependent variable. We inferred weights for each similarity metric, which we used to generate the final similarity scores for all unannotated listing pairs. The annotation task was reviewed by the ethics committee of the UCL Department of Security and Crime Science and was exempted from requiring approval by the central UCL Research Ethics Committee. Participants provided informed consent before taking part in the study.

6.5.1 Sampling annotation data

Since we expect only a low number of the same or highly similar cryptomarket and eBay listing pairs, randomly sampling (say) 1000 listings would most likely not yield a sufficient variety of low and high-human-rated similarity product pairs. Thus, we generated a preliminary manual scoring procedure to provide a crude ranking of product pairs from which we could sample, aiming to include product pairs with a broader range of similarity scores. We scored product pairs by weighting image and text similarity equally and merging them by cumulating unusual score distribution counts, indicated by scores of larger or lower than two standard deviations from the metric mean (for a detailed procedure description, see Appendix E4). After standardizing the score counts, we ranked the product pairs from highest to lowest similarity and sampled the first and last 250 products as well as a random sample of 500 from the remaining products, resulting in a sample of 1000 product pairs.

6.5.2 Annotating similarity scores

We recruited 220 participants from the crowdsourcing platform Prolific⁴⁶ and redirected them to a Qualtrics survey in which each annotated 10 listing pairs, enabling us to obtain at least two similarity ratings for each pair. For each product pair, participants were presented with the HTML page of a cryptomarket and eBay listing that we hosted locally. Distracting information that was irrelevant to the product itself (e.g., advertising, recommended other listings) was omitted and usernames were anonymized. For each viewed pair, participants were presented with three questions:

⁴⁶ www.prolific.co

- “Based solely on the images, how likely do they show the exact same product?”
- “Based solely on the texts, how likely do they describe the same product? (Title, Description, Specifics)”
- “Based on the combination of images and texts, how likely are the listings about the exact same product?”

Participants rated each question on a 7-point Likert scale with the labels 1: “Not at all”, 4: “Somewhat”, and 7: “Completely”. Figure 6.3 shows the distribution of similarity ratings, including the first and second ratings of the same product matches. Most product matches were rated as not at all similar or somewhat similar, while around 5% were rated as completely similar.

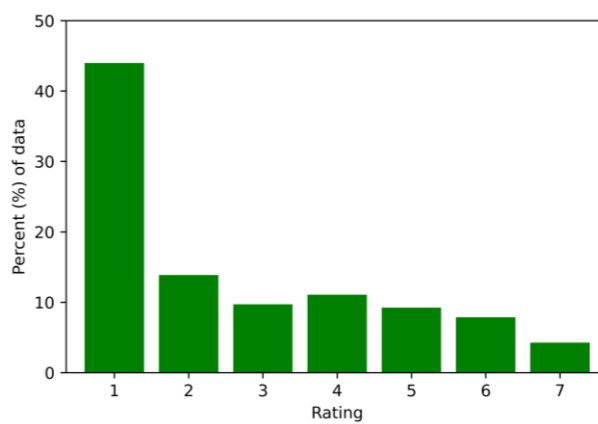


Figure 6.3. Distribution of annotated similarity ratings.

To examine the extent to which participants agreed on their similarity ratings between cryptomarket and eBay listings, we calculated the associated linear weighted Cohen’s Kappa score for the image, text, and overall similarity ratings (Table 6.3). Annotators seem to agree only slightly (0.01-0.20) or fairly (0.21-0.40) with each other (J. Cohen, 1960), indicating that the annotations might not be as reliable as hoped.

Similarity rating	Product group (N)					
	Clothes (94)	Watches (136)	Other (259)	Jewellery (394)	Electronics (65)	All (948)
Images	0.07	0.22	0.27	0.15	0.21	0.20
Texts	0.06	0.12	0.15	0.10	0.36	0.14
Overall	0.11	0.07	0.25	0.11	0.24	0.16

Table 6.3. Similarity rating agreements between listings for each product group and overall.

6.5.3 Predicting similarity scores

Although the agreements between annotators were low, we decided to examine the ratings and compare them to the hand-crafted rating system we used to sample the annotation data. We were interested in whether human-annotated similarities could facilitate a better ranking

system, even with low agreements. We averaged the similarity ratings for each product pair and performed a regression analysis with the averaged human ratings as the dependent variable and the automated generated similarity metrics as independent variables. Before running the analyses, we tested whether any independent variables showed strong multicollinearity by calculating the variance inflation factors (VIF). Excluding the word mover distance of the product names ensured that all VIF scores were below 5, which is an acceptable level of linear intercorrelation between the independent variables (Craney & Surles, 2002; Kim, 2019).

We then trained and tested a Random Forest regressor (RFR), an ordinary least squares linear regression model (OLS), and a linear Support Vector regression (SVR) model, each with a 10-fold cross-validation procedure. Model performance was measured with the mean absolute percentage error (MAPE) listed in Table 6.4. The closer the MAPE score is to 0, the better the prediction. However, all MAPE scores are high, with the SVR model performing slightly better. For the regression analyses, we used the Python package “scikit-learn” (Pedregosa et al., 2011) and the default settings for each model (see Appendix E5 for all model settings). We also considered a linear mixed-effects model with the product groups as a random effect, but such an approach seemed to perform worse than integrating the product categories as a dummy variable. For a detailed model comparison, see Appendix E6.

Model	MAPE (Std)
RFR	56.77 (5.23)
OLS	56.56 (5.11)
SVR	49.48 (4.02)

Table 6.4. Average model performances (10-fold).

We used the product pair ranks obtained with the SVR model for any further analyses since it showed the best performance. Table 6.5 shows the individual coefficients of the SVR model for each metric, which can show us how strong and in which direction the individual metrics influence the final similarity score. However, only USE (product names), q-grams (descriptions), and product types were significant at the 0.05 level.

Category	Metric	Coef.
Text (Product names)	Cosine (dist) Q-grams (3)	-0.864
	Cosine S-BERT	-0.148
	Cosine USE	0.957*
Text (Descriptions)	Cosine (dist) Q-grams (3)	-1.802*
	Cosine S-BERT	0.002
	Cosine USE	0.932
	WMD	-1.540
Images	Histogram (blurred)	-0.106
	ORB	-1.580
	SIFT	0.115
	SURF	-0.498
	Siamese (merged score)	0.284
Product type	Dummy variable	0.132*

Table 6.5. SVR coefficients; Significance level: * = $p < 0.05$.

6.6 Examining similarity scores and product matches

Before examining the product matches more closely, and to ensure a fair assessment, we excluded 643 pairs of the first and 647 pairs of the second scrape period since they contained no image similarity scores due to missing images in either the cryptomarket or eBay listings. In the coming sections, we present findings as to how the similarity scores varied across product categories, product origins, and eBay scrape periods to understand their importance in affecting similarity scores, which could be valuable in determining which eBay products are more likely to be found identical to cryptomarket counterfeits. We also manually inspected 200 product matches to assess how well the matching and ranking procedures work in finding identical products across anonymity networks and surface web.

6.6.1 Changes in listing similarities across categories, origins, and scrape periods

To examine whether the similarity scores varied over product categories, product origins (eBay), and the two scrape periods, and whether there were any interactions, we ran a three-way ANOVA with the similarity metric as the dependent variable. All main effects of product categories ($F(4) = 5489.36, p < 0.001$), product origins ($F(5) = 1259.25, p < 0.001$), and scrape periods ($F(1) = 1087.55, p < 0.001$) were statistically significant. Two-way interactions between product categories and product origins ($F(20) = 203.01, p < 0.001$), product categories and scrape periods ($F(4) = 78.98, p < 0.001$), as well as product origins and scrape periods ($F(5) = 16.29, p < 0.001$), were also statistically significant, as was the three-way interaction between all factors ($F(20) = 7.30, p < 0.001$).

Since we were interested in how product categories and product origins differ, as well as how they change across scrapes, we performed post-hoc t-tests between all product categories and product origins for each scrape period as well as between scrape periods alone. Table 6.6 and Table 6.7 show the Cohen's d effect sizes between all product categories and product origins. For both tables, the lower-left diagonal half of the table represents the differences within the first scrape period, and the upper-right diagonal half within the second. The direction of Cohen's d values can be read from row to column. Specifically, each table cell indicates the similarity difference from the reference category (row name) to the target category (column name). For example, the cell with the value of $d = -0.31$ for electronics (row) and watches (column) indicates the similarity difference between electronics to watches within the first scrape period. Looking at the similarity difference between the same categories within the second scrape period, we look at the electronics column and watches row with $d = 0.37$. Since the column and row labels are flipped, the value of $d = 0.37$ indicates the similarity difference between watches (row) to electronics (column). Thus, if the effect size sign (+, -) flipped from the first to the second scrape period (or vice versa) for the same categories (e.g., electronics and watches), the direction of similarity difference is the same in both periods. We used the Bonferroni alpha level correction to account for multiple comparisons with a starting alpha level of 0.05, resulting in an adjusted alpha level of 0.0016.

	Watches	Other	Electronics	Jewellery	Clothes
Watches	-	-0.68*	0.37*	-0.14*	-1.73*
Other	0.48*	-	0.93*	0.45*	-0.85*
Electronics	-0.31*	-0.62*	-	-0.42*	-1.91*
Jewellery	0.24*	-0.19*	0.44*	-	-1.22*
Clothes	1.67*	0.76*	1.55*	0.96*	-

Table 6.6. Cohen's *d* effect sizes for product category comparisons (row to column) for the first (lower-left diagonal half) and second (upper-right diagonal half) scrape period. Significance level: * = $p < 0.05$ (Bonferroni corrected).

Similarity changes between product categories were all significant across both scrape periods (Table 6.6). In the first scrape period, we can see the strongest (positive) difference in Cohen's *d* from Clothes to Watches and from Clothes to Electronics. Within the second scrape period, we can see the biggest (negative) changes in Cohen's *d* from Electronics to Clothes and from Watches to Clothes, mirroring the same differences as in the first scrape, but stronger. Cohen's *d* differences from the first to the second scrape period mostly became stronger, with only weaker (but in the same direction) differences between Jewellery and Watches as well as Jewellery and Electronics. The direction of differences did not change across scrape periods. Clothes seemed to show the lowest while Electronics had the highest similarities across scrape periods.

Similarity differences between product origins were almost all significant, and only the difference between Europe and USA within the first scrape period is not significant (Table 6.7). The biggest differences within the first scrape period were between the UK and China, as well as the UK and Other. Within the second scrape, we can see the biggest difference between Other and the UK, as well as China and the UK, reflecting a similar trend as in the first period. The direction of differences in the first and in the second scrape period remained the same. The UK seems to have the lowest similarity scores across scrape periods, while Other has the highest. Product categories and origins significantly changed similarity scores, showing that similarity scores behave differently for each property. Thus, some properties might be worth focusing on (e.g., Electronics and products originating from China) more than others (e.g., Clothes and products originating from the UK).

	China	Undec.	USA	H. K.	Other	Europe	UK
China	-	-0.82*	-0.11*	-1.29*	0.17*	-0.33*	-1.56*
Undec.	0.51*	-	0.65*	-0.49*	1.08*	0.51*	-0.80*
USA	0.14*	-0.29*	-	-1.14*	0.25*	-0.17*	-1.24*
H. K.	1.45*	0.88*	1.21*	-	1.47*	0.96*	-0.35*
Other	-0.32*	-0.73*	-0.33*	-1.72*	-	-0.56*	-1.58*
Europe	0.34*	-0.17*	0.12	-1.10*	0.71*	-	-1.11*
UK	1.81*	1.13*	1.28*	0.47*	1.76*	1.21*	-

Table 6.7. Cohen's *d* effect sizes for product origins comparisons (row to column) for the first (lower-left diagonal half) and second (upper-right diagonal half) scrape period; Undec. = Undeclared; H. K. = Hong Kong; Significance level: * = $p < 0.05$ (Bonferroni corrected).

Lastly, we compare similarity differences between scrape periods which showed a significant Cohen’s *d* effect size of $d = -0.16$, indicating a slight decrease in similarity scores from the first to the second scrape period. Thus, suggesting fewer good matches between cryptomarket and eBay products over time.

6.6.2 Finding the same products from cryptomarkets on eBay

To examine if the ranking system was able to find highly similar or the same products, we manually examined the top 50 matches of cryptomarket and eBay product pairs, which were ranked based on the trained SVR model and pairings selected at random (excluding the top 50). We inspected both datasets, consisting of product matches from the first and second eBay scrape periods.

6.6.2.1 How well do products match in the first scrape period?

Within the top 50 ranked product pairs of the first eBay scrape period, we found 13 unique cryptomarket products (i.e., cryptomarket products that matched with 50 eBay listings) distributed across World Market (7), Darkode (3), White House Market (2), and Torrez (1). From the random sample, we found 47 unique cryptomarket products distributed across World Market (10), Darkode (12), White House Market (7), and Torrez (18). Table 6.8 shows the distribution of product types for each sample of data.

Category	Top 50	Random 50
Clothes	0	8
Watches	2	9
Electronics	3	1
Jewellery	14	18
Other	31	14

Table 6.8. Product categories within the top 50 ranked and 50 randomly selected product pairs of the first scrape period.

Electronic products were either Apple smartphones or (Apple) headphones. Jewellery products were mostly watches in the top 50, but for the random sample, these also included necklaces, earrings, rings, handbags, and bars of gold or silver. Products categorized as “Other” were mostly shoes, specifically Nike shoes, but also contained one sweatshirt in the top 50. For the random sample, there were also wallets, earrings, wristbands, handbags, caps, and gold bars. Clothes were predominantly shirts but also contained jackets and a hoodie in the random sample. Table 6.9 shows the distribution of similarity ratings for the normalized SVR scores. The random sample contains ranks in the range of 502-65,342 (out of 66,430 products) with a median rank of 41,778.

Sample	Min	Max	Median	Mean	Std
Top 50	0.88	1.00	0.88	0.90	0.02
Random	0.37	0.83	0.59	0.60	0.10

Table 6.9. Similarity score distribution within both product samples of the first scrape period.

Based on manual inspections of the image and text for the top 50 ranked product pairs, we found 8 product pairs that seemed identical (the other pairings are discussed below). These were ranked 3, 4, 8, 14, 23, 33, 34, and 36. Specifically, we found five Nike shoes on eBay, which all resembled the Nike shoes found on Darkode (Figure 6.4: CM-1); two Apple smartphones resembling smartphones found on White House Market (Figure 6.4: CM-2), and one watch found on World Market (Figure 6.4: CM-3). Figure 6.4: eBay-1,2,3 shows image examples of the matching eBay products.

Product titles of cryptomarket and eBay pairs also exhibited high resemblance (Table 6.10), with only slight variations in word usage. The matching Nike shoe also shows the same brand or product identification number (Table 6.10: A). Prices are (substantially) lower for these cryptomarket products than for the matching eBay products.

Looking at the content of the descriptions for the cryptomarket products and the matching eBay products, we can see that a large portion of the text often differed, covering different aspects of the product or warranty and shipment (See Appendix E7 for complete example descriptions). For example, cryptomarket descriptions often explain how to order, how long the shipment will take, what measures are in place to avoid detection and how detections or complaints are handled. Besides shipping and warranty information, such aspects are mostly missing in eBay descriptions, as they are irrelevant concerns. However, both descriptions often contained additional product information, such as weight, height, colour options, sizes, etc., which can be valuable in determining the similarity between the products. Given the images, titles, and descriptions, we can say that seven eBay products out of the initial eight identified could be the same products as those sold on the cryptomarket. One of the identified Nike shoes on eBay was only available in a shoe size not sold on the anonymity network. Thus, 14% of the top 50 ranked product pairs might be the same products.

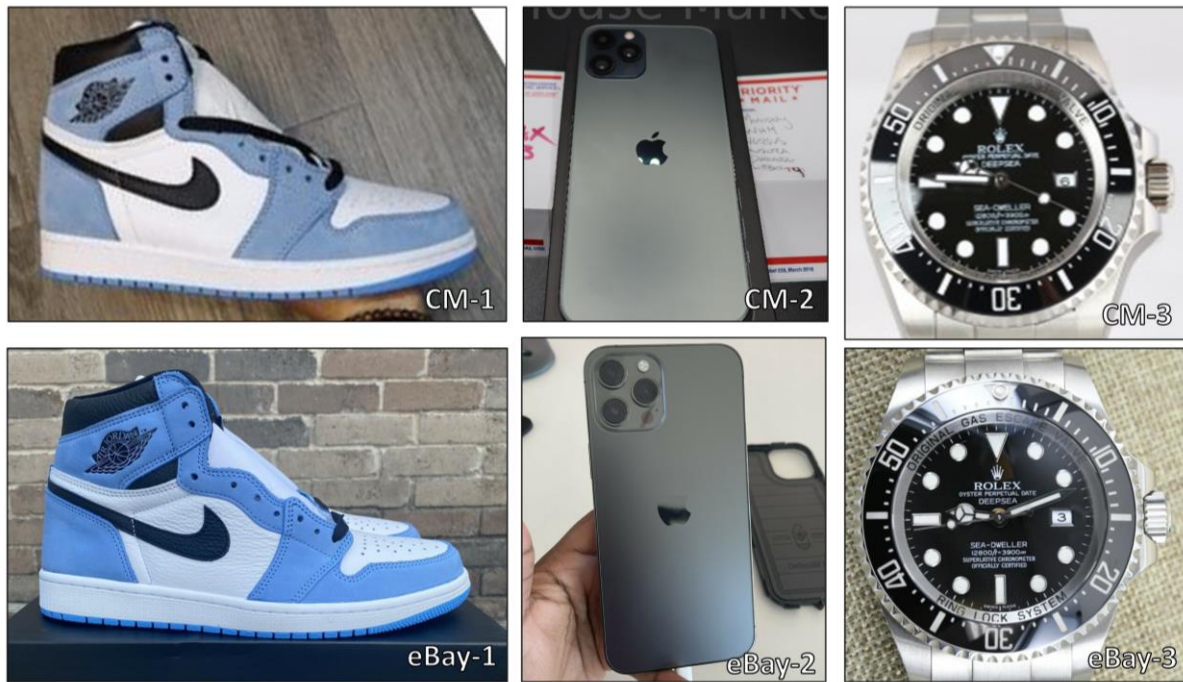


Figure 6.4. Examples of product images from the first eBay scrape from cryptomarkets (CM) and corresponding matching eBay products (eBay) of Shoes (1), Smartphones (2), and Watches (3). Some images are rotated and/or cropped.

		Product type (figure reference)		
		Shoes (1)	Phones (2)	Watches (3)
Platform	CM	nike air jordan 1 retro high og 555088-134	GRAPHITE - 512GB iPhone 12 Pro Max Sealed in Box - EZ BURN SERIES	Rolex - DEEPSEA SEA-DWELLER N V5S SAB UltimateAAA+
	Price	\$238	\$300	\$300
eBay	Title	Nike Air Jordan 1 Retro OG High White University Blue 555088-134 Men's Size 9.5	iPhone 12 Pro Max - Verizon - 512GB - Graphite - Open Box	Rolex Deepsea Sea-Dweller 116660 44mm Watch
	Price	\$355	\$1,339	\$12,500

Table 6.10. Examples of titles and prices of matching product pairs corresponding to the products shown in Figure 6.4; Words indicated with grey background were removed for the automated eBay web search.

Examining the random sample of 50 matches of the remaining ranks, we found two eBay products, with the ranks 4,617 and 13,149, that seemed identical to cryptomarket products. Specifically, a Louis Vuitton wallet and an iPhone. Although the images of the iPhone only contain the sealed box, making a visual comparison more difficult, the titles and descriptions match the phone model (version, memory specification, etc.). Based on the identical products found in both samples with seven (top 50) and two (random) identical matches, we found $3.5 \left(\frac{7}{2}\right)$ times more identical products within the top 50 than within the random sample.

6.6.2.2 How well do products match in the second scrape period?

In the second eBay scrape period, we found 27 unique cryptomarket products (i.e., cryptomarket products that matched 50 eBay listings) originating from White House Market (13), Torrez (7), World Market (6), and Darkode (1). From the random sample of 50, we found 46 unique cryptomarket products distributed across World Market (6), Darkode (13), White House Market (8), and Torrez (19). Table 6.11 shows the distribution of product types for each data sample.

Category	Top 50	Random 50
Clothes	0	6
Watches	1	12
Electronics	5	1
Jewellery	40	17
Other	4	14

Table 6.11. Product categories within the top 50 ranked and 50 randomly selected product pairs of the second scrape period.

Electronic products were either Apple smartphones or (Apple) headphones. For the top 50, jewellery products included wristbands, rings, necklaces, earrings, and watches. The random sample also included rings, handbags, and bars of silver or gold. Products categorized as “Other” were mostly shoes, specifically Nike shoes and sweatshirts in the top 50, as well as pants, handbags, sunglasses, caps, and slippers in the random sample. Table 6.12 shows the distribution of similarity ratings for the normalized SVR scores. The random sample contains ranks in the range of 1206-67,638 (out of 68,532 products) with a median rank of 33,417.

Sample	Min	Max	Median	Mean	Std
Top 50	0.85	0.98	0.87	0.88	0.03
Random 50	0.35	0.78	0.61	0.61	0.09

Table 6.12. Similarity score distribution within both product samples of the second scrape period.

Based on the manual image and text inspections, we found that three of the top 50 ranked product pairs seemed to be identical (Figure 6.5). Again, we found a match for the same Nike shoes (Figure 6.5: CM-1) previously identified in the first scrape period, but also a match of a bag charm sold on White House Market (Figure 6.5: CM-2), and a match of a watch sold on World Market (Figure 6.5: CM-3). All three corresponding eBay matches are seen in Figure 6.5: eBay-1,2,3. For the watch found on White House Market, we can see a sticker is placed on the glass, which is absent on the watch sold on eBay. However, such a sticker can most likely be removed for further sales.

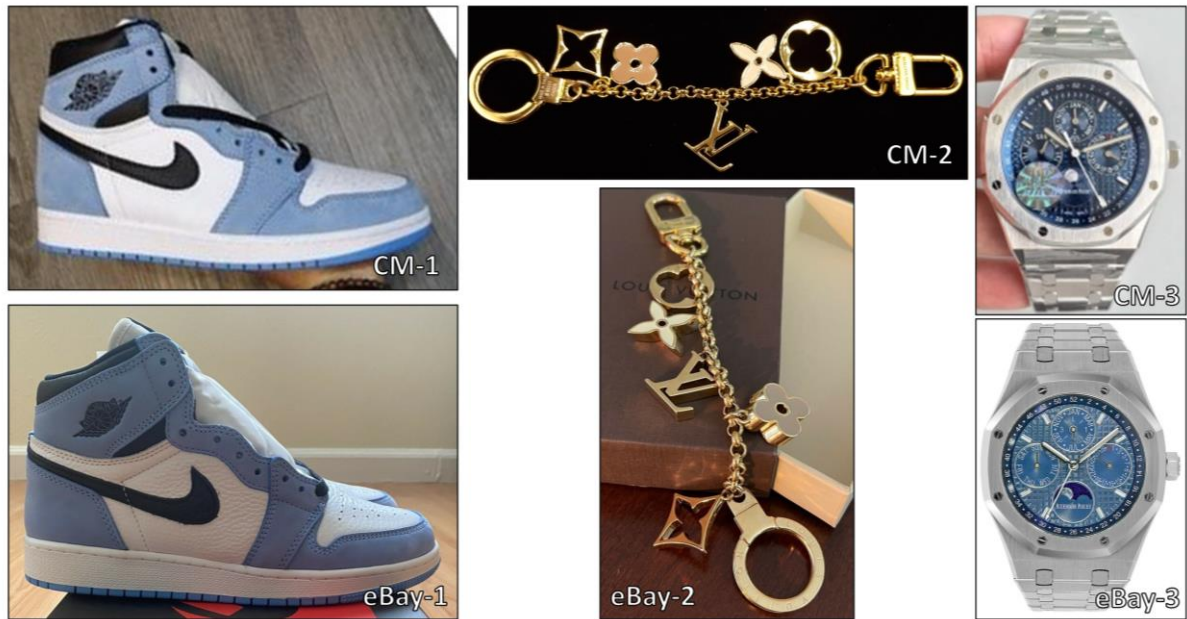


Figure 6.5. Examples of product images from the second eBay scrape from cryptomarkets (CM) and corresponding matching eBay products (eBay) of Shoes (1), bag charms (2), and Watches (3). Some images are rotated and/or cropped.

Again, product titles are very similar, and product prices are consistently lower on cryptomarkets (Table 6.13). However, the identification number of the Nike shoes within the titles does not match, but that could be related to the indicated children's shoe size, which was also available on the anonymity network.

		Product type (figure reference)		
		Shoes (1)	Bag charms (2)	Watches (3)
Platform	CM	nike air jordan 1 retro high og 555088-134	Louis Vuitton Bag Charm Chain Fleur de Monogram - RETAIL \$830 - UNDETECTABLE	Audemars Piguet - ROYAL OAK PERPETUAL CALENDAR B [UltimateAAA+]
	Price	\$238	\$300	\$450
eBay	Title	Nike Air Jordan 1 Retro High OG GS University Blue 575441-134 Size 7Y	LOUIS VUITTON Bag Charm Chain Fleur de Monogram Key Ring (104623)	Audemars Piguet Royal Oak Perpetual Calendar Watch 26574OR.OO.122 0OR.02
	Price	\$350	\$750	\$215,997

Table 6.13. Examples of titles and prices of matching product pairs corresponding to the products shown in Figure 6.6; Words indicated with grey background were removed for the automated eBay web search.

The corresponding product descriptions (Appendix E8) show a similar trend as previously identified but are more limited to factual details about the products (e.g., colour, shape, size) and their shipment or warranty. Thus, they seem more similar from a qualitative perspective than the inspected product description in the first scrape period. Based on the images, titles,

and descriptions, all three eBay products could be the same as advertised on the cryptomarket.

Examining the random sample of 50 matches of the remaining ranks, we found one eBay product, with the rank 1,245, that seemed identical to a cryptomarket product: a Rolex watch. The images, as well as the title and descriptions, match on both listings. Based on the identical products found in both samples with three (top 50) and one (random) identical match, we found $3 \binom{3}{1}$ times more identical products within top similarity ranks than in a random sample.

6.6.3 Highly ranked and similar product pairs that were not identical

32 products in the first scrape period and 12 in the second scrape period of the top 50 did not appear to be the exact same product but were highly similar and mostly varied only in their colour scheme. For example, corresponding cryptomarket Nike Shoes, seen in Figure 6.4: CM-1 and Figure 6.5: CM-1, were often matched with eBay Nike shoes, such as seen in Figure 6.6. We could also observe similar matching behaviour for other products, such as the Audemars Piguet watch seen in Figure 6.5: CM-1 matching with the watch seen in Figure 6.6. In many cases, the wrist bands of the watches were different, but these could easily be swapped for further resales. Thus, detecting the same product type with the same shapes or geometrical features seems to work better than accurate colour detection. A possible reason for such colour mismatches could be due to the colour histogram comparisons. Since the colour histograms account for the entire image, including the background of the products, the distribution of the histogram can be easily affected, possibly leading to inaccurate colour comparisons of the products.



Figure 6.6. Highly similar eBay products resembling counterfeits on cryptomarkets (Figure 6.4 and Figure 6.5: CM-1; Figure 6.5: CM-3) but with different colour schemes; Images are cropped.

6.6.4 Top 50 ranked product matches without image scores

Of the product matches that were excluded due to missing image comparison scores across both scrapes, 41 were initially ranked within the top 50. Here, we examine those excluded matches manually to determine why they were ranked so highly and if the ranking procedure could find identical products without image scores. By examining the raw webpage data and texts from those excluded matches, we can see that most of the product pairs without image scores originated from two cryptomarket products, specifically from “New In Box NEVER WORN Air Jordan 11 Space Jams Mens Sz 11” for which we found 31 matches in the first and

36 matches in the second scrape period. In addition, eight matches were found in the first and two in the second scrape period from “New In Box NEVER WORN Air Jordan 4 Supermans Mens Sz 11”. Based on the titles and descriptions of the two cryptomarket products, many of the associated eBay products seemed very good matches (Table 6.14). Interestingly, the raw webpage of the Air Jordan Space Jams Shoes contained small, low-resolution images, which were not properly extracted in high-resolution for the image analyses. However, it was still possible to compare the images manually with matches from eBay. With the help of the images, we identified 10 eBay product matches in the first and 12 in the second scrape period, which seemed to be identical products, but based on the titles, only three of them appeared to be of the same size.

eBay product titles
CLEAN Nike Air Jordan Retro 11 Space Jam 2016 Size 9 Men’s 378037-003 XI
Nike Air Jordan 11 Space Jam Men’s Size 10.5 - 2016 OG All
Air Jordan 11 Space Jam 2016 size 4 mens/5.5 womens with box, great condition.
DS 2016 Air Jordan 11 Retro men size 10 “SPACE JAM” 378037-003 BRAND NEW
Jordan 11 Retro Space Jam (2016) - Size 11 Men’s - New, Authentic
Men’s Nike Air Jordan 11 Retro Space Jam 2016 Release Size 10.5
Air Jordan 11 Retro Space Jam Men’s Size 7.5

Table 6.14. Examples of eBay product titles matched to the CM product “New In Box NEVER WORN Air Jordan 11 Space Jams Mens Sz 11”.

Furthermore, the Air Jordan Space Jam shoes from White House Market were sold for \$1,000, which exceeds all prices from the identical eBay matches by around \$400-\$800. Thus, the product matches are less likely to be the exact same products or re-sales from cryptomarkets on eBay.

6.7 Discussion

The current study uses information about cryptomarket counterfeits to find the same products on eBay. However, the system we tested cannot – and was not intended to – validate if the products found on eBay are, in fact, counterfeits. Instead, the purpose was to examine whether an automated approach could potentially alleviate some of the workload current law enforcement agencies face and facilitate a better understanding of current trends in counterfeit sales. By partially automating an otherwise tedious web search, we can speed up the gathering of intelligence, which can be used further by manually inspecting highly similar products. Thus, the current system should not be regarded as a stand-alone solution to finding surface web counterfeits but rather as a partial automation of otherwise manual web searches. Next, we will discuss some of the methods used, the results, limitations, and future avenues of the approaches employed.

6.7.1 How do product categories and origins affect product matching?

Based on the ANOVA analyses, we found that product similarities vary between product categories and product origins as well as the two combined. Specifically, post-hoc t-tests showed that similarities differed significantly between all product categories and almost all product origins. Finding the same products seemed more likely for Electronics and for products originating from countries that contributed less than 1% of all origins (“Other”). In turn, finding the same products seemed least likely for clothes and products originating from the UK. A possible reason for good matches within electronics might be the detailed descriptions of specifications that are often not present in other product types (e.g., memory, model number, colour, camera). An explanation as to why clothes show the least increase of similarity might be due to the high variability within that category (e.g., shirts, jackets, sweatshirts), making matching more difficult. We also observed that products originating from China had higher similarity scores than almost all other origins. Previous work shows that China is a predominant exporter of counterfeits, making the country a likely candidate to find very similar products (EUIPO, 2019; OECD/EUIPO, 2019), which is supported by our findings. Although unclear as to why finding highly similar or the same products differs for every product type and where they are sent from. Therefore, future research could investigate whether similarity cut-off scores, indicating which product matches should be manually examined, should be adjusted by product category or origins.

6.7.2 How does product matching change over time?

From our findings, we can see that the overall similarity scores slightly decreased from the first to the second scrape period. Those results are also reflected in our manual inspections, in which we found more identical products in the first period than in the second. Although we found indications that eBay listings seemed to become less similar to cryptomarket listings over time, it is important to note that most recent cryptomarket listings were collected (i.e., top 50 in each category, sorted from newest to oldest), but without exact dates. Thus, we do not know how long products were online exactly, making a more accurate assessment of time effects difficult. For example, if we assume that counterfeits will appear first on cryptomarkets and then on the surface web, our current observations seem to be contradictory. However, our data collection could have captured a later stage of the product offer cycle, with products present for some time already and offers slowly decreasing over time (e.g., due to market saturation of specific products or products being sold out).

We also see scrape periods interacting with product categories and product origins. Specifically, we observed that compared similarities between most product categories diverge more from each other from the first to the second scrape period, indicating greater differences between categories over time. However, comparing similarities between product origins, we observed for most comparisons a convergence across scrape periods, indicating more closer similarities from the first to the second scrape period. Thus, we see opposing trends between product categories and product origins over time. Although the current exploratory observations of product changes over time are preliminary and need to be tested

further, they give some indications that similarity scores behave differently over time depending on product groups and product origins. Thus, some specific products might show higher similarity for longer than other product types and might be worth tracking longer to find potential counterfeits.

6.7.3 How well does the matching and ranking procedure work?

Giving a clear estimation of how many cryptomarket products can also be found on eBay is difficult. However, with the comparison of the top 50 ranked product matches to a random sample of 50, we roughly estimated that there were about 3-3.5 times more identical matches within the high-ranked products than in a random sample. Although the samples for that manual inspection were small, their comparison may provide an indication that the ranking procedure works in ranking more likely matches higher. However, only seven out of the 50 top-ranked products were identical, and some identical products were also found in the random sample, showing that improvements are still needed. Ideally, we would determine a similarity cut-off score that would increase the likelihood that identical products would be found at an acceptable rate. As we asked individuals to annotate similarity scores of product pairs, we could consult practitioners for an acceptable cut-off score or crowdsource the task of finding such a cut-off score experimentally. In addition, a graphical interface connected to the ranking system, in which the product pairs are presented, would be helpful to speed up the inspection.

An assessment of the matching procedure is also complex because of the lack of performance measures, such as accuracy, precision, or recall. Generating such measures relies on a classification task, which was not possible to construct due to the absence of reliable labelled or ground truth data (i.e., knowing if a pair of listings are about the same product). The lack of ground truth data is why the system relies on similarity scores and a ranking procedure of cryptomarket and eBay listing pairs. The advantage of performance metrics, such as the ones mentioned above, is that they help us understand how well a classifier works and in which situations the system could be worth applying. Assuming that we might have ground truth data in the future, we could train a supervised method and consider how well a system should perform to be practical. Next to accuracy, precision and recall are essential scores that help assess a classifiers' capabilities. In the context of the current system, precision would indicate the fraction of identified identical products that are truly identical from all classified identical products (including false positives). Recall (also called sensitivity) would indicate the fraction of identified identical products from all identical products that could be identified. Thus, both measures are detailed performance measures of how well the system classifies identical products. Precision is crucial when assessing how well the system avoids false positives or how reliable positive predicted cases are truly positive (i.e., identical products). While the precision of the current system is unknown, consider what it would mean if it had a precision of 0.85. Such a precision would indicate that 85% of all predicted identical products are truly identical, while 15% are not (false positives), which could falsely signal good performance to practitioners. To explain, assume that the system would be implemented on large-scale tasks,

such as online shopping platforms, and 10,000 products are identified as identical to counterfeits on cryptomarkets, around 1,500 of these would be falsely identified. Depending on how positively classified cases are handled (e.g., automated removal of the listing or the vendor), many vendors or consumers could be unintentionally harmed financially or physically (e.g., through the removal of genuine medicines). Including a human in the loop would address such an approach but require time and effort. Similar to a ranking system, to improve efficiency, identified identical products could be ranked based on the classifier's confidence and then be further examined manually. For example, further investigations could examine the vendors, what they previously sold and whether they list multiple automatically identified products. Depending on the type of products, the online shopping platform and the affected brand could then be informed to coordinate further steps, such as gathering more information about the possibility that the products are counterfeits. A further discussion about the possible implementation and practical implications of automated systems can be found in Chapter 7 (General Discussion), in section 7.6.

6.7.4 Quantitative vs Qualitative results

Throughout this study, several quantitative results suggested possible problems with the current ranking system. For example, the low agreement scores between annotators on similarity suggested limited utility to building a regression model to predict an aggregate similarity score. With low agreements, the regression model would likely have difficulty in making reliable predictions; this was observed in the low performance of the model as measured using the MAPE scores. Furthermore, we observed some unexpected negative associations in the regression coefficients for text (e.g., q-gram, WMD) and image features (e.g., ORB, SURF), suggesting that some of the automated similarity scores negatively impacted the overall similarity score. However, despite these shortcomings, we observed a big difference qualitatively. That is, our comparison of the top 50 ranked pairings and a random sample of pairings suggest that the current system rankings can generally discriminate between better and worse matches between product pairs. Thus, albeit the quantitative indications of poor performance, the system showed utility qualitatively.

6.7.5 Limitations

Some of the steps in the current approach of collecting eBay data might have a strong impact on the matching procedure and could be improved in the future. Specifically, the eBay search query might act as a bottleneck for finding good product matches. The currently used default setting of the eBay search function is to find “best matches”, an opaque setting but most likely related to the entered keywords. Many of the cryptomarket product names that were used for the eBay search are very short and less descriptive, resulting in less accurate search results. In contrast, the titles of good matches mostly contained some product details, such as the Nike Air Jordan shoe, containing an exact model number. Automated systems for keyword generation, such as “KeyBERT” (Grootendorst, 2021), could be tested to generate more detailed search queries from lengthy product descriptions. Further refined search queries

could also be made through price ranges or product specifics (e.g., shoe size, model number, colour). However, determining a specific price range might also be difficult. Intuitively, a higher price for eBay products seems reasonable since we would assume that sellers want to make a profit. However, prices can fluctuate over time; observed cryptomarket listings might represent wholesales, which would complicate comparisons to listings of single items, or vendors might provide exact prices only after customer inquiries, which is a recurring practice (Soska & Christin, 2015). Therefore, eBay counterfeits might be sold for the same or even a lower price than the listed cryptomarket price in some instances.

The current system finds some highly similar products across platforms, but we do not know how they relate to each other and if there are any interdependencies. For example, products might be highly similar because the same vendor sells them; they might be resold by an individual who purchased the goods on a cryptomarket; or vendors on cryptomarkets may have researched surface web platforms to determine which products should be counterfeited and offered on cryptomarkets. Alternatively, product similarities might also originate from a complex relationship between manufacturers and sellers, or listing information (e.g., images, texts) may be copied from other advertisements without a direct connection between vendors. Thus, further research (discussed in the next section) would have to be conducted to explore those possible interdependences.

An additional problem the currently implemented system faces are the low agreement scores between annotators who rated the similarities between cryptomarket and eBay products. The human-rated similarity scores are essential in informing the regression model and determining how the individual automated scores should be weighted. The low agreement scores not only show how difficult judging similarity can be, but would also seem to jeopardize meaningful training of a regression model. By taking the average score of the rated similarities, we cannot alleviate all these concerns. However, given the results, and contrary to the expectations, the system seemed to be able to utilize human judgments to some extent, as we could find exact product matches. Thus, the model seemed to apply some general tendency of agreed similarity and, in doing so, demonstrated utility.

Nonetheless, more consensus between annotators is desirable. Therefore, future studies should aim to understand how and why annotators do not agree with each other and how this can be addressed. Annotators might need a revised version of instructions or additional text and image similarity definitions. Furthermore, more annotators for each product pair might be required to find a more robust similarity rating, which could be found through an average or majority voting. Similarly, they may be invited to revise their ratings after others have rated the same item(s), applying an iterative process of refining the ratings, or raters may be asked to discuss items collectively rather than independently.

6.7.6 Future work

For this study, we collected product information from only some of the available counterfeits on cryptomarkets. Future studies could automate some of the collection processes to expand the collection of counterfeits. For example, collecting all product information from a pre-selected and manually identified category of each cryptomarket. Instead of collecting all data from a market, which can be very time-consuming, the relevant product categories could be manually determined and provided to a scraper. Including such a manual element will help speed up the automated collection procedure and help to avoid the collection of irrelevant listing data (e.g., drugs, firearms, digital services). Future studies could also conduct more frequent data collection over a prolonged period (e.g., once a month over a year) to examine if and how the offers change over time more accurately. In addition, the search for the same products could be expanded to other platforms (e.g., Amazon, Etsy, Gumtree, and Otto). Similarly, the analyses of counterfeit listings could be expanded to include the vendors and compare them across platforms. For example, whether vendors on cryptomarkets that sell various counterfeit types (e.g., shoes, watches) show similarities to vendors on surface web platforms (e.g., display a similar product portfolio) and if specific product types might be sold together.

Looking at the Support Vector Regression models' coefficients, we can examine the influence the individual features (text and image metrics) have on the final similarity score. Although most features were non-significant, many showed a negative relationship with the dependent variable, while we expected more positive relationships. For example, S-Bert shows a negative association with product names, and the Universal Sentence Encoder shows a positive association, while we would have expected similar behaviour since the models are based on similar principles. Thus, future studies could examine the contributions of text and image similarity metrics in more detail to better understand their importance and to find a more optimal combination for generating a unified similarity score.

While manually inspecting matched cryptomarkets and eBay products, we observed many pairs that showed strong resemblance in shapes and type but differences in their colour scheme. Thus, accurate colour detection of the products seems difficult. A possible reason for such colour matching might be the colour histogram, which contains pixels of the entire image, including the background colours, possibly skewing the distribution to an unfavourable comparison. Future approaches could test whether masking the background – effectively separating the product from the background – could support a better colour analysis of the image. However, finding almost the same products in some cases but with a different colour scheme is also helpful. As for shoe sizes, which are often specified in the description, some cryptomarket products are also available in different colours, specified in the text but not visible in the example images. Although finding the same products is favourable, very similar products should not be discarded straight away to find potential counterfeits.

6.8 Conclusion

With the current work, we devised an automated system that finds eBay products that are similar to openly sold counterfeits on cryptomarkets. Although we do not know if the identified eBay products are counterfeits, some would warrant further inspection, as would the associated sellers. We also found some evidence to suggest that finding similar or the same products seems to have become more difficult over time (at least for the two periods considered) and depends on product type and origins. We identified several possible avenues on how to improve the current approach, such as integrating human judges in a more streamlined manner, not only in evaluating good product matches, but also in finding suitable cryptomarkets, finetuning the applied models to receive more robust similarity ratings, and finding practical cut-off scores, which can make the manual inspections more efficient. Thus, future versions of our approach could be used to investigate further the possible connections between cryptomarket and surface web listings, as well as hold practical value in supporting the detection of counterfeits on the surface web.

Chapter 7: General Discussion

This thesis has explored how data science might help combat online consumer fraud, which encompasses various types of fraud including non-deliveries, fraudulent billings, fake reviews, and counterfeits. The chapters of this thesis vary in terms of the types of fraud they addressed, but the common theme is that the frauds considered take place on online platforms. While online platforms enable fraud to be committed at scale, they also enable us to conduct analyses at scale, and some of these types of analyses have been explored in this thesis. This general discussion begins with a summary of the main findings by revisiting how various stakeholders tackle fraud and continues by examining the challenges and promises of automated methods to combat fraud. The chapter concludes with a discussion of the limitations of the work, suggestions for future avenues of research, and a consideration of the practical implications of the work reported.

7.1 Summary of main findings

This section will summarize the main findings of the chapters in this thesis, beginning with the background and literature review of online consumer fraud and anonymity networks, followed by the remaining chapters.

Current approaches taken by industry and (non-)governmental institutions to tackle fraud were reviewed in Chapter 2 and summarized in Table 2.1. Chapter 2 also introduced and described the various types of online fraud that focus on fraud enabled through online platforms, including fraudulent billings, non-deliveries, or selling lower-value products, including counterfeits. While approaches to combat such frauds vary, depending on the specific problem, companies are interested in maximizing profits and often use brand protection agencies or internal divisions to detect fraud (Ganguly, 2015; Pointer Brand Protection, 2019; Yellow Brand Protection, 2019). Their main goal is to remove and report fraudsters on their platforms through internal monitoring systems or complaints. Government bodies (e.g. the intellectual property office in the UK) and law enforcement agencies also react to complaints and use intelligence (e.g., investigations) on possible frauds to deter fraudsters and to limit their financial impact on society (FBI, 2018; Raine et al., 2015). (Non-)governmental institutions (e.g. Cifas, a fraud prevention organisation in the UK) often take a different approach by informing vulnerable individuals through guidelines or recommendations about how to avoid becoming a victim of fraud (Beals et al., 2015; Deevy & Beals, 2013; M. DeLiema et al., 2019; Peaston, 2019; Stanford Center on Longevity, 2019). The chapter further describes current data science approaches to combating fraud and closes with possible future approaches, such as automating manual tasks, uncovering previously undetected patterns in existing data, and facilitating the understanding of online markets at scale, which was explored in subsequent chapters.

Chapter 2 also introduced anonymity networks, small sub-parts of the deep web, which are only accessible with specialized software and allows for highly anonymized communication

between individuals and servers. Cryptomarkets, platforms on anonymity networks with similar functions as eBay, provide a relatively safe space for trading illicit goods. Although drugs are the predominant goods exchanged (up to 80% of listings), fraud-enabling products (2-5% of listings), such as defrauding guides, fake documents (e.g., passports, driver licenses, food stamps), credit card information, and counterfeits (e.g., watches, clothes, electronics) are also sold there. Therefore, anonymity networks and specifically cryptomarkets present an interesting opportunity to learn about the fraud and counterfeit economy. However, collecting data from anonymity networks on a larger scale can be challenging due to hurdles associated with automating the process.

Chapter 3 and 4 focused on the challenges of automated methods, such as the applicability of supervised machine learning methods in detecting fraudulent activity. More specifically, those chapters aimed to answer two questions: which annotation practices are important for creating a dataset usable for training a supervised model within the fraud context, and how can training data created experimentally or found (e.g., collected from online platforms) affect model performance?

Chapter 3 tackled the former question by exploring the hurdles of creating a labelled data set of suspicious and non-suspicious eBay listings that could be used for training a machine learning classifier to detect suspicious listings. The chapter shows that recruiting a sufficient number of experts to annotate the required amount of data can be challenging. More importantly, experts and non-experts did not agree well on what constitutes “suspicious”, either within their groups or across them, which limited the usability of the labelled data for training a machine learning classifier. Possible reasons for the observed low agreements were examined. These included the labels used and the labelling process employed (e.g., unspecified inspecting time of the listing). According to qualitative feedback from participants, other reasons for the low agreements included missing listing information needed for the annotation (the seller information) and the possible misalignment of the annotators’ expertise with the annotation task (i.e., an expert might only know how to identify a specific fraud). Based on the identified issues, the chapter provided recommendations for future studies and annotation tasks, such as more precise instructions and definitions for the labelling process, increased control during labelling (e.g., giving participants equal time), providing detailed seller information, and tailoring the labelling task to the annotator’s specific fraud-expertise.

Chapter 4 examined the second question by investigating possible confounds that can be introduced when combining data to create training datasets. In that chapter, a machine learning (ML) classifier was trained to detect fake online smartphone reviews. By obtaining ground truth data through an experimental procedure, the need to label the data as in Chapter 3 was circumvented. However, the chapter illustrated how data confounds, such as the origin of the data (i.e., data within the dataset is sourced differently) or whether the review writer owns a product (or not), can impact ML classification performance and lead to false conclusions. More precisely, reviews for which each class (genuine vs fake review) originated from a different source boosted classification accuracy between 20.85-44.27%, depending on

the review sentiment (positive vs negative). Similarly, if reviews originate from the same source but were written by product owners and non-owners, the classification accuracy was boosted between 6.15-9.84% (depending on sentiment). A combination of both confounds led to an increase of 24.89-46.23% accuracy, the largest increase. Such impacts on model performance lead to accuracy overestimations and wrongful conclusions about which features are important in distinguishing between classes of items, highlighting the importance of stronger experimental controls during dataset creation.

Chapters 5 and 6 focused on the potential promises of automated methods and looked at what might be learned about counterfeits from offerings on cryptomarkets. Both chapters sought to answer three questions: how prevalent counterfeits are on cryptomarkets, whether we can expand insights from border seizures and complaint statistics about the counterfeit economy with information from cryptomarkets (Chapter 5); and whether computational methods can be used to search for counterfeits on surface web markets using data extracted from cryptomarkets (Chapter 6).

To answer how prevalent counterfeits are on cryptomarkets and whether we can gather new insights for practitioners, Chapter 5 utilized openly accessible archival data ranging from January 2014 to September 2015 across multiple cryptomarkets. By using the information from product listings, the chapter explored one way of automatically estimating the number of counterfeit types, their origins, and sales volume. The chapter then compares the results to other measures, such as data collected by government border forces during seizures and complaint statistics collected by EU and UK authorities. Cryptomarkets were found to harbour many more watches but fewer clothes, electronics, footwear, tobacco, or other counterfeits than those seized at borders. However, cryptomarket listings and border seizure measures might illuminate the counterfeit economy from slightly different perspectives, such as capturing different moments in the product lifecycle (e.g., offered or already purchased), potentially complicating a direct comparison. In addition, the comparison also bears some uncertainty as seizures are highly dependent on the border authority's activity and mostly do not cover domestically produced and consumed products. The analysis of cryptomarkets also suggested that 80% of all counterfeits originate from China and Hong Kong, a finding that is also reflected in seizure measures. Counterfeits identified on cryptomarkets (within the analysis timeframe) were valued at \$1.8 million, but their estimated value on the surface web would be much higher if they were sold as original items. The results of this chapter suggest that by monitoring cryptomarkets, insights into the fraud and counterfeit economy can be gained in conjunction with other measures (e.g., border seizures) and should be considered by researchers and practitioners in the future.

Chapter 6 investigated how computational methods could be utilized to discover possible connections between crypto and surface web markets. The chapter examined one way of automating a process to see if products offered on cryptomarkets that were clearly labelled as counterfeits could be automatically identified on the surface-web platform eBay. Information concerning 453 counterfeits was collected from four cryptomarkets, all operating

in 2021. Product information from 134,000 eBay listings (collected across two waves of data collection) was collected using the counterfeit product titles as a search query. Through the combination of image and text similarity metrics, product matches were ranked. Based on the inspection of 200 product pairs, we found identical and highly similar products (e.g., shoes, smartphones, watches) on eBay, which were also openly sold on cryptomarkets as counterfeits and would warrant further investigation by law enforcement agencies. By comparing the top-ranked product pairings to random samples, we assessed the utility of the ranking procedure and found that within the top 50 ranked products, three times more identical products were found than for pairs of items sampled at random. We also found some indications that similarities between product matches decreased with time and depended on product categories and product origins. While some quantitative measures indicated poor performance of the applied methods (e.g., low annotator agreements, poor regression model performance), the system showed utility qualitatively and could hold practical value in the future. The chapter showed how possible connections between anonymity networks and the surface web could be investigated and how the search for counterfeits on the surface web could be supported through automation, possibly narrowing down counterfeit-affected product types that might be prioritised.

7.2 How studies relate to each other

The chapters in this thesis differ in which fraud type they are examining, but all show how data science methods could be utilized to combat online consumer fraud. Specifically, chapters 3 and 4 showed the hurdles researchers and practitioners face using supervised machine learning methods to detect online consumer fraud. Both chapters illustrate the difficulties of creating reliable labelled data and how biased datasets impact supervised methods. The different approaches to generating training data originate from the absence of usable ground truth data. Therefore, researchers and practitioners test various ways of creating datasets (described in more detail in section 2.3.2.), trying to circumvent the inaccessibility of ground truth data. Here, data originating from anonymity networks and cryptomarkets might help to provide ground truth data, at least for frauds involving counterfeits. Since counterfeits are openly sold on cryptomarkets as counterfeits, and we assume that vendors have few reasons to deceive possible customers about the products being counterfeits, we think that the product labels (i.e., products being counterfeits) are more reliable than we can obtain through other means (e.g., manual annotations). As a result, the product information from counterfeits on cryptomarkets might be able to support data science approaches that require data with reliable labels (i.e., ground truth data). Chapter 5 explored the counterfeit landscape on cryptomarkets to understand if and how such data could be utilized better. Chapter 6 extended the idea of using information from cryptomarkets and collected new data on counterfeits, and provided a proof of concept of how manual investigations of counterfeits on the surface web could be supported through automation.

Thus, we can regard the collection of information from cryptomarkets as an extension of obtaining reliable data, which is better suited than other means of creating datasets for

utilizing data science methods to better understand and combat online consumer fraud on the surface web. How we can further use the information from anonymity networks and cryptomarkets is discussed in section 7.4.

7.3 Generalization of findings

This thesis examined various hurdles in combating online consumer fraud (chapters 3 and 4) and how data from anonymity networks could help combat the sale of counterfeits (chapters 5 and 6). Similarly, we can discuss how those hurdles are generalizable to other online frauds and whether anonymity network data could also be utilized for combating those frauds. Two hurdles of utilizing data science approaches to combat online consumer fraud are generating reliable labels of fraud cases (e.g., fraudulent vs non-fraudulent advertisement) through (expert) annotations and data confounds introduced when creating datasets that can impact detection performances of machine learning models.

Ground truth data are missing for most online frauds, including the ones not discussed in this thesis, such as identity theft, voice phishing (vishing), romance scams, click frauds (false clicks in pay-per-click advertisements), chargeback frauds (claiming a monetary refund while retaining the purchased goods), or frauds related to fake websites (e.g., for phishing or non-deliveries). Thus, practitioners or researchers who aim to implement detection approaches (e.g., supervised machine learning models) for such frauds will likely rely on creating labelled datasets. However, finding capable annotators for those frauds might be difficult due to some ambiguity as to who should annotate the data and that annotators might not always be readily available (e.g., lack of time or funding). As a result, similar issues described in Chapter 3 might arise, complicating the creation of reliable data labels. While issues around reliable labels for consumer fraud data are likely also an issue for other online frauds, it is unclear whether the data confounds examined in Chapter 4 are also present when investigating other fraud types or platforms (e.g., online shopping platforms, booking websites). Chapter 4 investigated the effects of data confounds for text data (smartphone reviews), which are context-dependent and possible confounds (presence or strength) might differ for other reviews (e.g., hotel reviews). In addition, fraud datasets might also consist of other data types than text, such as images or behavioural data (e.g., user profiles), for which we do not know if the same confounds are present. We would have to conduct further tests to determine if and how confounds are introduced for different datatypes and other contexts or domains. However, researchers should be critical of their datasets and consider what confounds might be present or could be introduced when creating datasets.

This thesis uses data about counterfeits from anonymity networks and cryptomarkets to better understand the counterfeit landscape and examine how manual work of searching for counterfeits on the surface web could be supported. Utilizing data about counterfeits from anonymity networks, as in this thesis, is possible since those counterfeits are products that could also be sold on the surface web, such as large online shopping platforms. Such an approach is less feasible for many other online (consumer) fraud types (e.g., romance scams,

chargeback frauds). However, for frauds, such as phishing, identity theft, or click fraud, data from anonymity networks could still be helpful since some vendors offer products (e.g., phishing guides) or services (e.g., hacking services) that may contain information related to such fraud. In contrast to counterfeits, for which we can utilize their information through observation, other frauds might require stronger involvements, such as interactions with the vendors to obtain details about their offer or even require sample purchases. For example, obtaining defrauding guides might increase our knowledge about fraud strategies, which could help implement preventative measures. Similarly, hacking services could be purchased against fake targets to assess the hacking strategies and find possible vulnerabilities. Using such an approach, for example, researchers have found that online website security certificates can be bought, revealing security issues in acquiring them (Maimon et al., 2020). Overall, utilizing counterfeit data from anonymity networks and cryptomarkets is not precisely transferable to other frauds, but fraud-related data is still usable, and the approaches used in this thesis valuable.

7.4 Limitations and Outlook

This section will discuss the limitations of this thesis and will follow with possible future work on how to address those issues. First, we will discuss data quality, including the availability of ground truth data and relevant recent data (i.e., data about recent fraud). Next, we will discuss the temporal data coverage related to cryptomarket data, including the difficulties of collecting and sharing data covering multiple markets over longer periods.

7.4.1 Data quality

One recurring issue for utilizing data science methods to identify online fraud is the data collection procedures and usage of datasets. For supervised machine learning models, well-curated and high-quality data are necessary to prevent classification biases and ensure accurate model performance. Ideally, ground truth data would be used for such purposes, but ground truth data are scarce in online fraud research if not fully absent. Researchers mainly employ three strategies to address the lack of ground truth data, all with limitations. First, data are collected from online platforms, for which the labels are given by the platform or inferred using some rule (Barbado et al., 2019; Fazzolari et al., 2021; Mukherjee et al., 2013a; Rahman et al., 2015; Ren & Ji, 2019; D. Zhang et al., 2016). In such cases, data labels must be trusted often without the possibility of validation since most platforms are not transparent about their labelling process. Second, data labels are determined by experts, journalistic activity, or theoretical models, but the validity of obtained labels can, in most cases, not be verified (Flood, 2012; Fornaciari & Poesio, 2014; Hernandez-Castro & Roberts, 2015). This approach was adopted in Chapter 3, illustrating some obstacles faced when experts are asked to annotate data. Third, ground truth data labels are created through experimental work (Gutierrez-Espinoza et al., 2020; Perez-Rosas & Mihalcea, 2014; Salvetti et al., 2016), which was also adopted in Chapter 4, but experiments are often criticized for lack of external validity (Crawford et al., 2015; Mukherjee et al., 2013a; D. Zhang et al., 2016). While all three

approaches have their merits in some contexts, they cannot replace ground truth data of actual fraud cases. Therefore, many fraud (detection) studies rely on data for which the labels are determined without high certainty of their validity, which limits the possible inferred conclusions (Barbado et al., 2019; Rahman et al., 2015).

Furthermore, datasets are often very specific, resulting in trained classifiers that do not generalize well when tested in different circumstances (e.g., data from different online platforms or product types) (Geirhos et al., 2020). Similarly, fraudsters continuously change their strategies leading to different online traces and patterns in the data. Thus, supervised models can become quickly outdated, and their performance can drop if they do not adapt to such shifts through iterative re-training. Although the concept of an arms race between perpetrators and crime prevention measures is not new (Ekblom, 1997), the time needed to collect data and re-train supervised methods might be too great for the measures to be effective. Considering the many hurdles associated with generating high-quality and timely labelled data in the fraud domain, utilizing supervised models to predict fraud might not be practical enough currently, at least with the approaches employed here.

However, some of the discussed quality issues could be addressed by making ground truth data more accessible through better data-sharing practices between researchers, practitioners, companies, and consumers. Since stakeholders pursue different goals with their data, each stakeholder has valuable data that would benefit the research enterprise concerned with better understanding and detecting fraud. Coordinated data-sharing efforts have already been established in other domains, such as the GIFCT's hash-sharing database⁴⁷, which collects extremism content from various stakeholders and shares the hashed values. Hashing is the process of transforming the content (e.g., video, image, text) into a representative (often shorter) value of characters. In short, a hash is a short representation of content that cannot be reverse-engineered, effectively hiding its original information. Therefore, using the same hash algorithm allows various stakeholders to share and cross-reference content more efficiently without revealing potentially sensitive information. Similarly, the National Fraud Database (NFD)⁴⁸, maintained by Cifas, a non-profit association for fraud-prevention in the UK, contains fraudulent or suspicious data that association members can access. The data mostly contains transactional data and is primarily used for verifying transactions and identities, but it could be extended by consumer-fraud-related instances. The advantage of the NFD is that the infrastructure to collect and share data already exists. However, only association members currently have access, making collaboration with researchers more complicated. Other institutes, such as ODISSEI⁴⁹, have started with similar approaches by implementing secure analysis environments (e.g., SANE⁵⁰) that are intended to facilitate data sharing of sensitive data through secure analyses (Meer et al., 2022). Such a centralized database could tackle the problem many organisations face of sharing personal

⁴⁷ <https://gifct.org/hsdb/>

⁴⁸ <https://www.cifas.org.uk/fraud-prevention-community/member-benefits/data/nfd>

⁴⁹ Open Data Infrastructure for Social Science and Economic Innovations: <https://odissei-data.nl/en/en-odissei/>

⁵⁰ Secure Analysis Environment

information by only providing access to aggregated data or the analysis of results accessed through API calls. Therefore, such a database could act as a secure data environment. Alternatively, data could be anonymized through tools, such as ARX (quantitative data) or Textwash (text data) (Haber et al., 2022; Kleinberg et al., 2022), that would facilitate data sharing for sensitive or personal content. Ideally, a new database designed for online (consumer) fraud should be implemented, including data that captures consumers' perspectives on how they were defrauded, and the role of the platforms used (e.g., online shopping); such data would improve researchers' scope for understanding signs of possible fraud. Data capture could be facilitated through the very online platforms where fraud occurs by implementing easily accessible reporting tools. Similar to the NFD and the GIFCT's hash-sharing database, online (shopping) platforms could collaborate on a consumer fraud database, facilitating better fraud detection on each platform.

Another important step to improve data sharing—and thereby improve the reliability of applied ML models—is data documentation practices. Datasets are mostly collected with specific intentions (e.g., detecting fake reviews) related to the investigated research question and the associated research design. While those datasets are in most cases adequate for their respective use cases, making them available to the community will entail re-using the data with different investigatory intentions. Here, data documentation practices are important to make others (e.g., researchers and practitioners) aware of the dataset's limitations and possible included biases (Heger et al., 2022; Olteanu et al., 2019). Some approaches are currently developed but are not as widely used yet; these include model cards for models, which aim to document properties of trained ML models to increase awareness of model biases and other ethical issues (Mitchell et al., 2019), and similarly datasheets for datasets (Gebu et al., 2021) and system cards (Gursoy & Kakadiaris, 2022). Developing such documentation practices further with respect to the fraud domain would support re-using existing datasets. Since online fraud can be multifaceted, well-documented datasets can support decisions in choosing the most suitable data for a particular use case (e.g., research questions, application in practice) to avoid biases. Such developments could be made through an iterative and interactive process. First, researchers could be asked to fill out an initially developed documentation template and survey their opinion on its usability. Second, the template will be revised based on the feedback and re-distributed. Here, informing template users about error frameworks, which are theoretical models of potential biases when collecting data, could be useful to stimulate reflections on data collection processes further (Amaya et al., 2020; Sen et al., 2021).

7.4.2 Temporal data coverage

Other data quality issues, such as the temporal data coverage of multiple cryptomarkets, can also be limited. Automated data collection is inherently difficult due to the anonymized space and precautions taken by website administrators to secure and protect the data from automated crawlers. Platform users often undergo lengthy registration procedures and are regularly prompted to solve advanced CAPTCHAs (i.e., a small task, such as identifying traffic

lights on images to prove you are not a machine). Furthermore, the anonymization procedure of the networks (e.g., Tor) of relaying the internet traffic can slow down the functionality of the websites and the navigation, leading to prolonged scraping procedures. These issues may lead to data gaps during data collection and should be considered when analysing and interpreting data collected using automated approaches (Ball et al., 2019; Van Buskirk et al., 2016). The issues can be further exacerbated when data is collected from multiple platforms over time. Such issues were faced in Chapter 6 when information from counterfeits on cryptomarkets was collected, and the same products were searched on eBay. The data is limited to a few markets and covers only two points in time for the collected eBay data. Therefore, the conclusions drawn from the results in Chapter 6 should be seen as preliminary and should be re-examined in future studies. For example, the temporal coverage of cryptomarket and eBay data should be expanded by collecting data for a year or longer in set intervals (e.g., once a month) to generate a better picture of the counterfeit and fraud landscape and allow for a better examination of possible connections between platforms. Understanding if crypto and surface web markets affect each other could be useful for trend detection and valuable for consumers, online markets, and authorities. For example, previous journalistic work has suggested that producers of counterfeited shoes interact with the Reddit online community⁵¹ to understand which shoes customers would like to see counterfeited next (D. Thomas, 2018). Therefore, we already have some hints of possible connections between online platforms (e.g., forums and social media) and the counterfeit economy, which would be worth examining further by expanding the current analyses of cross-platform connections further, including more shopping and forum platforms.

The issue of limited temporal data coverage relates to data quality, including data accessibility with respect to anonymity networks. Since anonymity network data is difficult to collect and can contain sensitive information, most data are not shared. However, sharing data is important to provide others with research opportunities – particularly those without the capability of collecting such information – and to promote replicability, which is essential in the research process. As the Cambridge Cybercrime Centre⁵² has provided underground forum or other crime-related data to others through data-sharing agreements, cryptomarket data could be collected and shared similarly. The challenges lie in the required technical infrastructure (e.g., customized automated scripts) and its maintenance, such as finding new markets, building new scrapers, and addressing any errors during scraping.

7.5 Theoretical perspectives

Since this thesis takes a strong data-driven approach, the discussion of theoretical perspectives has been relatively limited in the previous chapters. Therefore, this section will examine how broader crime theories, such as Routine Activity Approach (L. E. Cohen & Felson, 1979), rational choice (Clarke & Cornish, 1985; Cornish & Clarke, 1987), and crime scripts (Cornish, 1994a, 1994b) relate to this thesis and how they could be valuable in future works.

⁵¹ www.reddit.com

⁵² <https://www.cambridgecybercrime.uk/datasets.html>

7.5.1 Routine Activity Approach and Controllers

As discussed in Chapter 1, the Routine Activity Approach (RAA) focuses on the situation in which a crime can occur and assumes that when a motivated offender (someone willing to commit a crime) and a suitable target (e.g., a desired item, such as valuables) converge absent a capable guardian (e.g., a bike lock that protects the target), a crime will be more likely to take place (L. E. Cohen & Felson, 1979). Initially, RAA was developed to understand crime in the physical world but has recently also been applied to cybercrimes, such as malware, fraud, hacking or phishing (Bossler & Berenblum, 2019; Hutchings & Hayes, 2009; Kigerl, 2021; Ngo et al., 2020; Reyns & Randa, 2020; Simpson et al., 2014). When the RAA is applied, the focus can be on understanding and explaining cybercrime victimization through the profiling of user activity online (Drew, 2020; Hutchings & Hayes, 2009; Ngo et al., 2020; Reyns & Randa, 2020), the behaviours and motivations of offenders (Harrison et al., 2020), or the concept of guardianship (Williams, 2016). The places where crimes occur are also important, and more recent work concerned with place management extends the initial RAA framework (Eck, 1994; Felson, 1995). The updated “crime triangle” shown in Figure 7.1 is used to illustrate concepts by showing a crime occurring (in the centre) when a motivated offender encounters a suitable target in a particular type of place (Figure 7.1, inner triangle). Associated with each of the RAA (inner triangle) components are controllers (shown in the outer triangle). These can be understood as forms of supervisors, which includes handlers (e.g., parents or friends that are related to the offender), who can directly influence the behaviour of an offender; guardians, whose role it is to protect the target; and, managers, such as the owner of an establishment, who has responsibility for that place, including preventing crime within it (Eck, 1994; Felson, 1995; Sampson et al., 2010).



Figure 7.1. The crime triangle represents how when the various RAA components converge, a crime is more likely to happen. Adapted from (Sampson et al., 2010).

A crime is assumed to be more likely when any of the controllers is absent or ineffective (Felson, 2008). Considering online (consumer) fraud, individuals who interact with an online

platform (the equivalent of a place) to make purchases can be seen as (suitable) targets since they are willing to spend money and can be easily reached by motivated offenders. Some efforts are made to make platform users less suitable to fraud (see Chapter 2; Table 2.1) by determining their susceptibility and providing them with guides and awareness of defrauding schemes (Beals et al., 2015; M. DeLiema et al., 2019; Stanford Center on Longevity, 2019). However, as discussed in section 2.2, regarding online fraud, online platform (place) managers (e.g., website operators) might not be as effective as necessary, and active guardianship (e.g., real-time detection mechanisms and responses) is often absent. Thus, parts of this thesis can be placed within the larger theme of efforts to increase such guardianship, and deterring offenders from committing a crime.

Given the cross-national jurisdictional issues of authorities and the fact that many stakeholders may be involved in addressing online fraud (e.g., an online platform on which the fraud occurs, local authorities responsible for affected individuals), improving guardianship is further complicated. Here, wider theoretical considerations of super controllers can be helpful (Mui & Mailley, 2015; Sampson et al., 2010). To explain, super controllers control the controllers, but often only indirectly interact with them. They can include a broad range of regulatory bodies and financial or political organisations. In the case of online fraud, as discussed in 2.2, incentives for implementing fraud detection approaches are not always aligned and the owners of platforms on which fraud occurs could be further incentivised by regulatory and political bodies (super controllers). Such incentives could include changes in the law to regulate the controllers' actions in dealing with fraud. Thus, changes to incentives for site operators might lead to stronger collaborations with authorities or researchers to improve their automated detection approaches.

7.5.2 Rational Choice

The rational choice perspective takes the view from the (possible) offender and assumes that they make rational decisions by considering the perceived risks and benefits of their actions (Clarke & Cornish, 1985; Cornish & Clarke, 1987). Although the offenders' decisions might seem rational to themselves, others might not perceive them as so. The idea of situational crime prevention is that by changing the situational circumstances around the possible offender, the perceived risks and rewards can be changed, reducing the chances of offending. Such changes can include the increase of perceived risk of detection, the increase of efforts, or by reducing the perceived rewards and excuses for committing a crime. As such, the rational choice perspective is essential to situational crime prevention efforts (Clarke, 1995; Freilich & Newman, 2018). The rational choice perspective and the Routine Activity Approach are closely related since both inform us of similar measures for crime reduction. For example, increased guardianship would also translate into increased perceived risk (of detection) or effort the offender needs to commit a crime.

As previously discussed, existing research suggests that cybercrime offenders perceive the chances of being detected as low (Hutchings, 2013). Thus, low perceived risk (or lack of

capable guardianship) supports the current understanding of the need for automated detection approaches. By examining how data science methods can facilitate our understanding and detection of fraud, the thesis aimed to provide insights that could help change the perceived risks and efforts for possible offenders. Next to automated detection approaches, insight into which products are counterfeited (discussed in Chapter 5) could support online platforms, brands, or authorities to implement preventative measures. Other approaches that were not explored could also change the offender's perception. For example, similar to reassurance mechanisms on online platforms for consumers (e.g., money-back guarantee, certified vendor batches), features or notices that would stress the detection of fraudulent activity could be implemented to change the perceived risks. Some research has suggested that increased seller anonymity increases the risk of fraudulent activity (Harrison et al., 2020). Thus, increasing (perceived) accountability and decreasing the anonymity of vendors could also be explored further.

7.5.3 Crime scripting

Crime scripts are used to inform our understanding of crime by sequentially describing the steps involved in their commission (Cornish, 1994a, 1994b). The idea is that by identifying the necessary actions for a crime, procedures can be developed to disrupt them, thereby hindering the execution of the crime. Recent years have seen a strong increase in the use of crime scripts, often in the cybercrime domain (Dehghanniri & Borrion, 2021). For example, scripts were used to better understand attacks on online banking, carding (unauthorized trafficking of credit card information), phishing, and identity fraud (Dehghanniri & Borrion, 2021; Holt & Lee, 2022; Hutchings & Holt, 2015; C. Lee, 2020). For example, Holt & Lee (2022) examined the procedures involved in obtaining counterfeit documents from 19 surface and crypto market vendors by qualitatively examining the text and images of the websites containing such offers. They identified various initiation and entry steps (e.g., ad creation, customer interaction with ad and seller electronically), vendor actualizations (e.g., placing an order and electronic payment, counterfeit document creation, shipping), and exit steps (e.g., receiving the product, conflict resolution in case of faulty product or shipment). The findings showed that the procurement of counterfeit documents mostly depends on cryptocurrencies and suggests that vendors have access to government equipment and personnel. The results also showed similarities to other findings investigating cryptomarket processes of obtaining stolen data (Hutchings & Holt, 2015), which could be interesting to compare to purchasing counterfeit apparel and whether differences are present. Other studies have used crime scripting to understand online auction fraud in which a fraudster uses stolen credit card information to purchase products on online shopping platforms (Hartel et al., 2010). The authors suggest disruption approaches for various steps in the crime script, such as the stronger policing of stolen credit card information, improved analyses of fraudulent transactions, or disrupting the procurement of stolen data by subverting legitimate with false information (e.g., posting fake advertisements on cryptomarkets). Others have examined crime scripting for organised fraud (Levi, 2008) or the counterfeiting of pharmaceuticals (Kennedy et al., 2018). Thus, future studies would benefit from creating crime scrips of

consumer fraud types on online shopping platforms and selling counterfeit apparel on anonymity networks and the surface web.

Such future work could take (at least) two perspectives in creating crime scripts. First, through observations and theoretical considerations, scripts could be created for online consumer fraud on surface web platforms (e.g., eBay, Amazon). Such scripts could be extended by looking at cryptomarkets, which offer fraud-related products and services, which might provide additional insights into the specifics of committing consumer fraud. Some research has already examined the steps needed when interacting with cryptomarkets (Holt & Lee, 2022; Hutchings & Holt, 2015), which could be used for comparisons. Second, a more interactive approach could be taken by making purchases of counterfeits and fraud guides from cryptomarkets that could further inform crime scripts. Only a few studies have made purchases on cryptomarketplaces before, mainly for drug analyses to determine their quality, shipping procedures, and if they aligned with how they were advertised (Arce, 2019; Jurásek et al., 2021; Rhumorbarbe et al., 2016). Others tested DRDoS attack service capabilities offered on cryptomarkets and compared their advertised and actual service (Hyslip & Holt, 2019). Such approaches are associated with ethical and legal issues but might generate valuable information. Purchasing fraud guides and counterfeited apparel could inform researchers and authorities about differences between advertised and delivered goods and services on cryptomarkets, as well as possible other defrauding methods and the counterfeits themselves. Such information could include the quality of the counterfeits (e.g., materials used) and the shipping process, which could be valuable for identifying counterfeits elsewhere (e.g., surface web platforms, border seizures). Sample purchases could also be made on surface web platforms of items found (through automated searches introduced in Chapter 6) to be the same as counterfeits offered on cryptomarkets. Products could then be compared to determine whether they are indeed both counterfeits. However, conducting test purchases also poses a legal risk to researchers and other individuals involved. Thus, any illegal activity should be coordinated with the appropriate authorities (e.g., national crime agency, justice department), which can be further complicated when shipments cross national borders (Rhumorbarbe et al., 2016). A risk-benefit analysis before test purchases could provide further support in deciding whether purchases should be made by considering what can be learned from the purchased items and how potential risks can be minimized. For example, guidelines on appropriately registering, navigating and conducting purchases on cryptomarkets could be implemented to prevent leakage of personal information to protect the researchers.

7.5.4 Applicability and limitations of theoretical perspectives

Chapters 3 and 4 of this thesis are concerned with the usability of automated methods for detection approaches of online consumer fraud. With the Routine Activity Approach (RAA) and the rational choice perspective, such methods can be placed in a broader framework and be understood as tools to change the situational circumstances of possible offenders by increasing the risks and efforts for offending. Those theoretical perspectives help locate where

interventions could be placed (e.g., on online shopping platforms) and how they could be conceptualised.

However, ambiguities can arise when those theoretical concepts are applied to cyberspace. For example, the RAA conceptualises that suitable targets are required for a crime to be more likely to occur, but how suitability is determined and if those targets are individuals, products, or services can sometimes be unclear. The RAA describes that the suitability of a target is dependent on physical characteristics (e.g., weight, height) and accessibility. However, such characteristics are less relevant as items will not have to be moved to or from individuals unless counterfeits are physically created and shipped (or services provided) (Nikitkov et al., 2014). Furthermore, target suitability is mostly examined through the lens of fraud victimizations, which are mainly about the individuals' demographics (age, gender, etc.) and are correlational (Holtfreter et al., 2008; Kemp & Erades Pérez, 2023; Sarno & Black, 2023). However, suitability likely also depends on how the online environment shapes fraud opportunities through the individuals' decision-making when navigating the online space (Pratt et al., 2010). Thus, the online platforms' affordance to fraud should be included when assessing the suitability of a target.

While the concept of a suitable target, as thought of in the physical space, might not always be transferable to cyberspace, the offenders' perception of possible (suitable) targets might also change. Specifically, fraudsters might not consider what a suitable target is (as individuals) when utilizing deceptive advertisements or webpages because such approaches follow a strategy similar to phishing methods (e.g., spam e-mails) that target everyone without selecting specific individuals. Therefore, the theoretical considerations within the RAA of what a suitable target is and how suitability is determined are more ambiguous when applied to cyberspace and online consumer fraud.

Situational crime prevention (SCP) offers practical and concrete examples of preventative measures, but as they are primarily designed for the physical space, they are not always easily transferable to cyberspace (Clarke, 1980, 1995; Freilich & Newman, 2018). For example, methods that increase the efforts for offenders, such as deflecting offenders (e.g., street closures) or controlling tools/weapons, are challenging to implement for online consumer fraud. In e-mail fraud, some offenders might be deflected through spam filters, but such methods are more difficult to implement on online platforms. Furthermore, defrauding individuals on online shopping platforms does not require additional tools (e.g., software) that would have to be acquired. Similarly, methods that aim to reduce the rewards (e.g., conceal or remove targets, identify property, deny benefits) or reduce provocations (e.g., reduce frustrations and stress, avoid disputes, reduce emotional arousal) are difficult to translate to cyberspace. Many crime prevention techniques require physical locations or objects that are not always present or replicated within the online environment. The relatively passive role of fraudsters in online consumer fraud further complicates the translation of SCP preventative techniques, such as concealing or removing targets, since fraudsters mostly do not actively search for the targets. Concepts that are more easily applicable in cyberspace (e.g., target

hardening, extending guardianship, assisting in surveillance, reducing anonymity) are often challenging to implement due to jurisdictional issues or the misaligned stakeholder incentives (as discussed in 2.2).

However, the place where online consumer fraud occurs (e.g., online platforms) is vital for fraudsters to scale up their schemes, which is why many technological solutions, such as detection approaches, are and should be implemented on these platforms. Such approaches mostly fall within the category of increasing the risk for offenders, but platforms might also be helpful for methods that would increase offenders' efforts. However, current SCP techniques for increasing the effort struggle to incorporate the place of online fraud since the physicality of places changes in cyberspace. Previous studies investigated how platform structures and layouts influence the decision-making process are not new (Di Geronimo et al., 2020; M. Bhoot et al., 2021; Mathur et al., 2019), but similar research on psychological mechanisms in online fraud is scarce (Shang et al., 2023). Thus, more work examining platform designs and affordances might be helpful to find and extend current SCP techniques to online consumer fraud.

Chapters 3 and 4 highlight issues with the implementations of supervised machine learning methods, but both chapters do not consider theoretical aspects for possible preventative measures. However, Chapter 5 discusses how knowledge of counterfeited products could support others (e.g., brands, online platforms, authorities) to better implement preventative measures. In particular, brands and manufacturers could benefit from information about cryptomarket counterfeits since they are directly involved in developing, creating and distributing the product and could most easily intervene at any of the lifecycle steps of the affected product (category). Thus, the applied data science methods to collect and analyse data from cryptomarkets do not act as preventative tools but enable others to act. As such, the perspectives of the Routine Activity Approach and the rational choice theory are helpful in conceptualising who would benefit from information about counterfeits from cryptomarkets and how such information could be used. However, as described in Chapter 5, detailing practical preventative measures depends on the product and the information receiver (e.g., authorities, brands) and would have to be further specified with the implementers.

Chapter 6 examined how some of the manual work authorities face could be automated. While the focus was to enable authorities to work more efficiently, mainly supporting prosecution, some preventative effects might also occur. For example, an increased capability for authorities to surveil the markets would also increase the risks for offenders to be detected, resulting in a possible deterrence effect. Deterrence can be regarded as part of the rational choice perspective, as possible punishments are included when possible offenders consider the costs associated with a crime (Akers, 1990; Piliavin et al., 1986).

The thesis explored various approaches in which data science can help combat online consumer fraud, and the theoretical perspectives can, in most cases, provide a framework in

which those approaches can be situated and be described in terms of how they might work for fraud prevention. However, translating the theory to cyberspace can, in some cases, be difficult due to the change of physical properties of targets and places, as well as the change of proximity between fraudsters and victims in the online space.

7.6 Practical Implications

Fraud reports have shown annual (measured between 2016-2017) prevalence estimations per individual from 14.3% to 17.5%⁵³ in the general US public, and around half (53.7%) were related to or dependent on the internet (K. Anderson, 2019). Such measures illustrate that online platforms play an important role in fraud, but they also show that automated detection is further complicated due to the imbalance of genuine and fraudulent data, also referred to as a low base rate. The problem of low base rate events (infrequent events) is more pronounced in other domains, such as terrorism or threat assessment research (Kleinberg, 2019; van der Vegt, 2021; van der Vegt et al., 2019). However, the issues persist in detecting online fraud or counterfeits, which have a prevalence estimation of around 3.3% of worldwide trade (OECD/EUIPO, 2019). The issue of detecting low base rate events is that a highly accurate detection system would still suffer from a high false positive rate (wrongly detected fraud cases) due to the class imbalance in the data (e.g., fraud vs non-fraud). In such cases, the wrong assumption that a highly accurate system will always lead to reliable and good predictions can also be referred to as the base rate fallacy. We can take the example of predicting (non-)violence from text data, in which violence-actualizers (individuals who act violently after writing threatening text) are a rare occurrence (exhibit a low base rate) (van der Vegt, 2021). In that example, a theoretical detection system with 95% accuracy (correctly identifying violence-actualizer and non-actualizers) that is tasked with classifying 100 million documents from (non-)violent actualizers in which 1% of documents originate from violent-actualizers, would wrongly predict 4,95 million and correctly predict 950,000 documents as originating from violent actualizers. Thus, the theoretical system would only predict 16.10% correctly as violent-actualizers, resulting in a large false positive rate. Although the base rate of violent-actualizer is lower than that of online fraud, the same problem of making predictions with strong class imbalances (counterfeit vs non-counterfeit) within the data persists. We can hypothesize a situation in which a counterfeit detection system with an accuracy of 95% is aimed at detecting counterfeits on an online shopping platform. Put differently, the system would correctly classify counterfeits (i.e., sensitivity or recall) and non-counterfeits (i.e., specificity) 95% of the time. If 5% of all items were counterfeits (a high estimation) and 100 million items were to be classified, only 50% of all items predicted to be counterfeit (4,750,000) would be correctly classified, reflecting a precision of 50. In turn, 250,000 would be wrongly predicted as non-counterfeits (false negative). Therefore, a high-performance (e.g., 95% accuracy) will still result in many false positives, which might be too many to be practically useful. Table 7.1 is a confusion matrix, which shows the distribution of

⁵³ Some individuals were victim of fraud more than once during the measured year.

correctly predicted counterfeits (recall), non-counterfeits (specificity), as well as wrongly predicted counterfeits (false positive), and non-counterfeits (false negative).

		Prediction		
		Fraudulent	Non-fraudulent	Total
Reality	Fraudulent	4,750,000	250,000	5,000,000
	Non-fraudulent	4,750,000	90,250,000	95,000,000
	Total	9,500,000	90,500,000	100,000,000

Table 7.1. Theoretical distribution of predicted counterfeits with 5% occurrence from a hypothetical detection system with a 95% accuracy; adapted from (Kleinberg, 2019; van der Vegt, 2021; van der Vegt et al., 2019).

With the risk of many false positives (i.e., detecting an item of interest as fraudulent, which is, in truth, non-fraudulent), we also must consider their possible impact. Next to financial costs to practitioners or companies (e.g., due to misallocated investigations or lost revenue from suspended vendors), great hardships on many individuals might be imposed (e.g., due to wrongful accusations, loss of income, damaged seller reputation, etc.). The consequences of false positives are further exacerbated due to the scalability inherent to automation. Since fraud also has negative financial and personal impacts, any costs associated with false negative cases (i.e., missed fraudulent cases) should not be disregarded, but their comparatively rare occurrence will, in most cases, be outweighed by the costs associated with false positives. Thus, depending on how a prediction system would be implemented, it could result in great mismatches between allocated financial and human resources. Hence, any system that predicts fraud should minimize the number of false positives.

Anyone interested in understanding and detecting fraud must be aware of the limitations of the different quantitative measures applied to assess how well an automated system works (e.g., detecting fraudulent activity). For example, a supervised machine learning system with high performance (e.g., as indicated by a high classification accuracy) might create a false sense of security among practitioners due to poor labelling, possible data biases, or the base rate fallacy. Thus, quantitative assessment measures, commonly applied, cannot always capture all the processes or limitations of a system, and high performance (as measured using the types of metrics commonly employed in studies such as those reported here) does not always equate to high qualitative performance. To better understand what the quantitative measures can or cannot capture, researchers should aim to convey their limitations more easily to practitioners to support appropriate applications. For example, describing how the training data was created, how the data labels were determined, and how the system was tested can provide a better understanding of what good performance means. Similarly, knowing how the training data was acquired and from where can help understand in which circumstance(s) the detection system (often a supervised model) could be useful. Conveying such information could be achieved through better data documentation practices, as mentioned above, or other approaches, such as the ALGO-CARE⁵⁴ guideline (Oswald et al., 2018), which highlights legal and practical concerns around risk assessment tools, or the

⁵⁴ Advisory, Lawful, Granularity, Ownership, Challengeable, Accuracy, Responsible, Explainable

VISOR-P⁵⁵ checklist (van der Vegt et al., 2022), which aims to sensitize practitioners to the usefulness of computational linguistic threat assessment tools. A similar guideline or checklist could be developed for fraud-related tools to convey possible limitations and best practices.

Assuming a large-scale detection system would be implemented on an online shopping platform (with the costs of false positives) to remove vendors and advertisements from the platform, fraudsters might also be displaced instead of permanently removed. Displacement may occur when an intervention to disrupt a crime is implemented and the crime's target, locations, time, procedure or offender changes (Bowers, 2011; Cornish & Clarke, 1987; Johnson et al., 2014; Tompson et al., 2023). Displacement may also occur for cybercrime, which has been shown when authorities shut down cryptomarkets, and vendors and buyers quickly migrate to other or newly created markets (Décary-Hétu & Giommoni, 2017; Ladegaard, 2019; Zambiasi, 2022). Thus, some fraudsters might adapt to avoid detection, such as migrating to other platforms, targeting different products, or changing their fraud strategy. With those possibilities, the effects of implementing large-scale detection methods should be continuously evaluated to limit the possible displacement of fraud. However, forcing fraudsters to migrate to other platforms might result in temporal disruption of their efforts, which could still be valuable.

Given the issues and limitations around predicting fraud or deceit (also discussed in Chapters 3 and 4), implementing prediction models might not be practical. The uncertainties and low reliability associated with predicting fraud and high false positives rates might not be cost-effective. As a result, data science methods might be more suitable to help human decision making by supporting the organization and understanding of data through scalable analyses. Similarly, implementing supportive tools that take a human-integrated approach and automate some of the manual-intensive work currently faced by authorities or online market platforms might be more fruitful. For example, automation-supported processes may include intelligence gathering (as explored in chapters 5 and 6), advanced web searches, or other investigative steps discussed in Chapter 2. Future research could explore how data originating from investigations could be more easily used for data science. To that end, researchers could observe daily investigative routines of practitioners to identify processes that might be suitable for automation approaches and identify information that is not currently recorded, but that would be beneficial for the sorts of analyses reported here. Information could be recorded more structurally but also include free text. For example, when investigations involve online websites, their content could be captured and annotated once the investigation is over. In that way, a database could be created over time, allowing for large-scale analyses.

⁵⁵ Validity, Indicators, Scientific quality, Openness, Relevance, Performance

Conclusion

This thesis examined how data science methods might help combat online fraud. Cyber-enabled fraud schemes exist in various forms, are easily scalable, and affect many individuals. Scalability is often achieved through internet (shopping) platforms through which fraudulent interactions are facilitated. Those online platforms, which can facilitate the convergence of suitable targets and motivated offenders, can also serve as an entry point for data science approaches to better understand the space (e.g., markets, consumers, vendors) and employ automated methods to support fraud detection. Although this thesis looked at only a few forms of fraud, the methods applied could be extended to other forms of online fraud that occur on online (shopping) platforms.

However, using data science methods to predict fraud has limitations and faces many challenges, making some explored prediction approaches currently unsuitable in practice. Fraud prediction problems are multifaceted but often originate from inadequate (training) data. In turn, conducting large-scale analyses to understand the fraud landscape better and partly automate manually intensive work in a human-integrated approach seems more promising. Future work should advance collaborations between the various stakeholders affected by fraud to address the many fraud types within multiple knowledge domains. Importantly, any future approaches to combat fraud should be transparent about their methodologies and limitations to allow for an appropriate assessment by anyone interested in using them.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *arXiv:1605.08695 [Cs]*. <http://arxiv.org/abs/1605.08695>
- Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68, 90–113. <https://doi.org/10.1016/j.jnca.2016.04.007>
- Abidi, W. U. H., Daoud, M. Sh., Ihnaini, B., Khan, M. A., Alyas, T., Fatima, A., & Ahmad, M. (2021). Real-Time Shill Bidding Fraud Detection Empowered With Fused Machine Learning. *IEEE Access*, 9, 113612–113621. <https://doi.org/10.1109/ACCESS.2021.3098628>
- Ablon, L., Libicki, M., & Ablar, A. (2014). *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar*. RAND Corporation. <https://doi.org/10.7249/RR610>
- Adamsson, H. (2017). *Classification of illegal advertisement*. Uppsala University.
- Ailipoaie, A., & Shortis, P. (2015). *From Dealer to Doorstep – How Drugs Are Sold On the Dark Net*. Global Drugs Policy Observatory.
- Akers, R. L. (1990). Rational Choice, Deterrence, and Social Learning Theory in Criminology: The Path Not Taken. *The Journal of Criminal Law and Criminology (1973-)*, 81(3), 653. <https://doi.org/10.2307/1143850>
- Almendra, V., & Enachescu, D. (2011). A Supervised Learning Process to Elicit Fraud Cases in Online Auction Sites. *2011 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 168–174. <https://doi.org/10.1109/SYNASC.2011.15>
- Almendra, V., & Enachescu, D. (2012). A Fraudster in a Haystack: Crafting a Classifier for Non-delivery Fraud Prediction at Online Auction Sites. *2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 233–239. <https://doi.org/10.1109/SYNASC.2012.21>
- Alzahrani, A., & Sadaoui, S. (2018). Scraping and Preprocessing Commercial Auction Data for Fraud Classification. *arXiv:1806.00656 [Cs, Stat]*. <https://doi.org/10.6084/m9.figshare.6272342>
- Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, 8(1), 89–119. <https://doi.org/10.1093/jssam/smz056>
- Anderson, K. (2019). *Mass-Market Consumer Fraud in the United States: A 2017 Update* (p. 153). Bureau of Economics, Federal Trade Commission.
- Anderson, K. (2022). Mass-Market Consumer Frauds: What the Statistical Data Show. In Y. Hanoch & S. Wood, *A Fresh look at Fraud: Theoretical and Applied Perspectives*. Routledge.
- Anderson, M., & Magruder, J. (2012). Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *The Economic Journal*, 122(563), 957–989. <https://doi.org/10.1111/j.1468-0297.2012.02512.x>
- Anderson, R., Barton, C., Böhme, R., Clayton, R., van Eeten, M. J. G., Levi, M., Moore, T., & Savage, S. (2013). Measuring the Cost of Cybercrime. In R. Böhme (Ed.), *The Economics of Information Security and Privacy* (pp. 265–300). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39498-0_12
- Arce, J. L. T. (2019). *Differences in Cocaine Quality Sourced from Cryptomarkets and Traditional Drug Markets*.

- Arthur, D., & Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. *Stanford*, 11.
- Baccouri, N. (2020). *Deep_translator* (1.6.0) [Python]. https://github.com/prataffel/deep_translator (Original work published 2020)
- Ball, M., Broadhurst, R., Niven, A., & Trivedi, H. (2019). *Data Capture and Analysis of Darknet Markets*. 15.
- Baravalle, A., & Lee, S. W. (2018). Dark Web Markets: Turning the Lights on AlphaBay. In H. Hacid, W. Cellary, H. Wang, H.-Y. Paik, & R. Zhou (Eds.), *Web Information Systems Engineering – WISE 2018* (Vol. 11234, pp. 502–514). Springer International Publishing. http://link.springer.com/10.1007/978-3-030-02925-8_35
- Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4), 1234–1244. <https://doi.org/10.1016/j.ipm.2019.03.002>
- Bauerly, R. J. (2009). ONLINE AUCTION FRAUD AND EBAY. *Marketing Management Journal*, 19(1), 134–144.
- Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3), 346–359. <https://doi.org/10.1016/j.cviu.2007.09.014>
- BBC. (2015, May 18). Gucci sues Alibaba over “counterfeit goods.” *BBC News*. <https://www.bbc.com/news/technology-32781236>
- Beals, M., DeLiema, M., & Deevy, M. (2015). *Framework for a taxonomy of fraud* (p. 40). Financial Fraud Research Center; Stanford Center on Longevity; FINRA Investor Education Foundation. <http://162.144.124.243/~longevl0/wp-content/uploads/2016/03/Full-Taxonomy-report.pdf>
- Belavadi, V., Zhou, Y., Bakdash, J. Z., Kantarcioglu, M., Krawczyk, D. C., Nguyen, L., Rakic, J., & Thuriasingham, B. (2020). MultiModal Deception Detection: Accuracy, Applicability and Generalizability. *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 99–106. <https://doi.org/10.1109/TPS-ISA50397.2020.00023>
- Bergman, M. K. (2001). White Paper: The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1). <http://dx.doi.org/10.3998/3336451.0007.104>
- Biddle, P., England, P., Peinado, M., & Willman, B. (2003). The Darknet and the Future of Content Protection. In J. Feigenbaum (Ed.), *Digital Rights Management* (Vol. 2696, pp. 155–176). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-44993-5_10
- Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. *O’Reilly Media Inc*, 504.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1), 17–35. <https://doi.org/10.1214/07-AOAS114>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4/5), 993–1022.
- Booij, T. M., Verburch, T., Falconieri, F., & Wegberg, R. S. van. (2021). Get Rich or Keep Tryin’ Trajectories in dark net market vendor careers. *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 202–212. <https://doi.org/10.1109/EuroSPW54576.2021.00028>
- Bossler, A. M., & Berenblum, T. (2019). Introduction: New directions in cybercrime research. *Journal of Crime and Justice*, 42(5), 495–499. <https://doi.org/10.1080/0735648X.2019.1692426>

- Bowers, R. T. G., Kate J. (2011). Assessing the Extent of Crime Displacement and Diffusion of Benefits: A Review of Situational Crime Prevention Evaluations *. In *Crime Opportunity Theories*. Routledge.
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. <https://doi.org/10.18653/v1/D15-1075>
- Bracci, A., Nadini, M., Aliapoulios, M., McCoy, D., Gray, I., Teytelboym, A., Gallo, A., & Baronchelli, A. (2021a). Dark Web Marketplaces and COVID-19: After the vaccines. *arXiv:2102.05470 [Physics]*. <http://arxiv.org/abs/2102.05470>
- Bracci, A., Nadini, M., Aliapoulios, M., McCoy, D., Gray, I., Teytelboym, A., Gallo, A., & Baronchelli, A. (2021b). Dark Web Marketplaces and COVID-19: Before the vaccine. *EPJ Data Science*, 10(1), 6. <https://doi.org/10.1140/epjds/s13688-021-00259-w>
- Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Branwen, G., Christin, N., Décary-Héту, D., Andersen, R. M., StExo, El Presidente, Anonymous, Lau, D., Sohlhz, Kratunov, D., Cacic, V., Whom, McKenna, M., & Goode, S. (2015). *Dark Net Market archives, 2011-2015* (2015-07-12). <https://www.gwern.net/DNM-archives>
- Broadhurst, R., & Ball, M. (2020). *Availability of COVID-19 related products on Tor darknet markets*. Australian Institute of Criminology. <https://doi.org/10.52922/sb04534>
- Business Insider. (2021, August 16). *Bezahlt, aber nicht bestellt: Die dubiose Masche mit den Amazon-Paketen*. Business Insider. <https://www.businessinsider.de/wirtschaft/handel/brushing-die-dubiose-masche-mit-amazon-paketen-a/>
- Button, M. (2021). Hiding behind the Veil of Action Fraud: The Police Response to Economic Crime in England and Wales and Evaluating the Case for Regionalization or a National Economic Crime Agency. *Policing: A Journal of Policy and Practice*, 15(3), 1758–1772. <https://doi.org/10.1093/police/paab022>
- Calis, T. (2018). *Multi-homing sellers and loyal buyers on darknet markets*. Erasmus University.
- Çalışkan, E., Minárik, T., & Osula, A.-M. (2015). *Technical and Legal Overview of the Tor Anonymity Network* (p. 32). NATO Cooperative Cyber Defence Centre of Excellence. https://ccdcoe.org/uploads/2018/10/TOR_Anonymity_Network.pdf
- Caneppele, S., & Aebi, M. F. (2019). Crime Drop or Police Recording Flop? On the Relationship between the Decrease of Offline Crime and the Increase of Online and Hybrid Crimes. *Policing: A Journal of Policy and Practice*, 13(1), 66–79. <https://doi.org/10.1093/police/pax055>
- Capuozzo, P., Lauriola, I., Strapparava, C., Aioli, F., & Sartori, G. (2020, August 3). *Automatic Detection of Cross-language Verbal Deception*.
- Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strophe, B., & Kurzweil, R. (2018). Universal Sentence Encoder. *arXiv:1803.11175 [Cs]*. <http://arxiv.org/abs/1803.11175>
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5), 2022–2038. <https://doi.org/10.3758/s13428-019-01273-7>
- Chang, W.-H., & Chang, J.-S. (2012). An effective early fraud detection method for online auctions. *Electronic Commerce Research and Applications*, 11(4), 346–360. <https://doi.org/10.1016/j.elerap.2012.02.005>

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, X., Cheng, W., Ouellet, M., Li, Y., Maimon, D., & Wu, Y. (2021). Identifying Darknet Vendor Wallets by Matching Feedback Reviews with Bitcoin Transactions. *2021 International Conference on Data Mining Workshops (ICDMW)*, 788–797. <https://doi.org/10.1109/ICDMW53433.2021.00102>
- Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *JOURNAL OF MARKETING RESEARCH*, *10*.
- Christin, N. (2013). Traveling the silk road: A measurement analysis of a large anonymous online marketplace. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 213–224. <https://doi.org/10.1145/2488388.2488408>
- Christin, N., Yanagihara, S. S., & Kamataki, K. (2010). Dissecting one click frauds. *Proceedings of the 17th ACM Conference on Computer and Communications Security - CCS '10*, 15. <https://doi.org/10.1145/1866307.1866310>
- Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*(4), 261–283. <https://doi.org/10.1007/BF00116835>
- Clarke, R. V. (1980). “Situational” Crime Prevention: Theory and Practice. *The British Journal of Criminology*, *20*(2), 136–147.
- Clarke, R. V. (1995). Situational Crime Prevention. *Crime and Justice*, *19*, 91–150. <https://doi.org/10.1086/449230>
- Clarke, R. V., & Cornish, D. B. (1985). Modeling Offenders’ Decisions: A Framework for Research and Policy. *Crime and Justice*, *6*, 147–185. <https://doi.org/10.1086/449106>
- CNET. (2020). *Phone Reviews*. CNET. <https://www.cnet.com/topics/phones/products/>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, L. E., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*(4), 588–608. <https://doi.org/10.2307/2094589>
- Committee of Public Accounts. (2023). *Progress combatting fraud* (43). House of Commons.
- Conlon, S. (2017, June 29). How Chanel Successfully Curbed Counterfeiters. *British Vogue*. <https://www.vogue.co.uk/article/chanel-wins-counterfeit-case-amazon-sellers>
- Cornish, D. B. (1994a). THE PROCEDURAL ANALYSIS OF OFFENDING AND ITS RELEVANCE FOR SITUATIONAL PREVENTION. *Crime Prevention Studies*, *3*(1), 151–196.
- Cornish, D. B. (1994b). Crimes as scripts. *Proceedings of the International Seminar on Environmental Criminology and Crime Analysis*.
- Cornish, D. B., & Clarke, R. V. (1987). Understanding Crime Displacement: An Application of Rational Choice Theory. *Criminology*, *25*(4), 933–948. <https://doi.org/10.1111/j.1745-9125.1987.tb00826.x>
- Cornish, D. B., & Clarke, R. V. (2003). Opportunities, precipitators and criminal decisions: A reply to Wortley’s critique of situational crime prevention. *Crime Prevention Studies*, *16*, 41–96.
- Corsearch. (2023). Content Protection: Investigation Services. *Corsearch*. <https://corsearch.com/investigation-services/>
- Counterpoint Reserach. (2020). *Smartphones*. Counterpoint Research. <https://www.counterpointresearch.com/devices/smartphones/>

- Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. *Quality Engineering*, 14(3), 391–403. <https://doi.org/10.1081/QEN-120001878>
- Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1), 23. <https://doi.org/10.1186/s40537-015-0029-9>
- Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. Free Press. <https://books.google.de/books?id=XIOAAAAIAAJ>
- Daeef, A. Y., Ahmad, R. B., & Yacob, Y. (2016). Websites Phishing Detection using URLs N-Grams as a Discriminating Features. *International Journal of Control Theory and Applications*, 9(42), 10.
- Danilák, M. (2021). *Mimino666/langdetect* [Python]. <https://github.com/Mimino666/langdetect> (Original work published 2014)
- Décary-Hétu, D., & Giommoni, L. (2017). Do police crackdowns disrupt drug cryptomarkets? A longitudinal analysis of the effects of Operation Onymous. *Crime, Law and Social Change*, 67(1), 55–75. <https://doi.org/10.1007/s10611-016-9644-4>
- Deevy, M., & Beals, M. (2013). *The scope of the problem* (p. 46). Financial Fraud Research Center; Stanford Center on Longevity; FINRA Investor Education Foundation.
- Dehghanniri, H., & Borrión, H. (2021). Crime scripting: A systematic review. *European Journal of Criminology*, 18(4), 504–525. <https://doi.org/10.1177/1477370819850943>
- DeLiema, M., Fletcher, E., Kieffer, C. N., Mottola, G. R., & Pessanha, R. (2019). *Exposed to scams. What separates victims from non-victims?* (p. 24). Stanford Center on Longevity, Federal Trade Commission, FINRA Foundation, International Association of Better Business Bureaus, BBB Institute for Marketplace Trust.
- DeLiema, M. I., Mottola, G. R., & Deevy, M. (2017). *Findings from a Pilot Study to Measure Financial Fraud in the United States*. Stanford Center on Longevity; FINRA Investor Education Foundation. <https://www.ssrn.com/abstract=2914560>
- Dellarocas, C. (2006). Reputation Mechanisms. In T. Hendershott (Ed.), *Handbook on Economics and Information Systems* (pp. 629–660). Elsevier.
- Demant, J., Munksgaard, R., & Houborg, E. (2018). Personal use, social supply or redistribution? Cryptomarket demand on Silk Road 2 and Agora. *Trends in Organized Crime*, 21(1), 42–61. <https://doi.org/10.1007/s12117-016-9281-4>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1), 74–118. <https://doi.org/10.1037/0033-2909.129.1.74>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [Cs]*. <http://arxiv.org/abs/1810.04805>
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73. <https://doi.org/10.1145/2500499>
- Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., & Bacchelli, A. (2020). UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376600>

- Dijk, J. van, Tseloni, A., & Farrell, G. (2012). *The International Crime Drop: New Directions in Research*. Springer.
- Dong, F., Shatz, S. M., & Xu, H. (2009). Combating online in-auction fraud: Clues, techniques and challenges. *Computer Science Review*, 3(4), 245–258.
<https://doi.org/10.1016/j.cosrev.2009.09.001>
- Douglas, M. B. (2015). *Pharmaceutical Crime on the Darknet*. 24.
- Drew, J. M. (2020). A study of cybercrime victimisation and prevention: Exploring the use of online crime prevention behaviours and strategies. *Journal of Criminological Research, Policy and Practice*, 6(1), 17–33. <https://doi.org/10.1108/JCRPP-12-2019-0070>
- Du, P.-Y., Zhang, N., Ebrahimi, M., Samtani, S., Lazarine, B., Arnold, N., Dunn, R., Suntwal, S., Angeles, G., Schweitzer, R., & Chen, H. (2018). Identifying, Collecting, and Presenting Hacker Community Data: Forums, IRC, Carding Shops, and DNMs. *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 70–75.
<https://doi.org/10.1109/ISI.2018.8587327>
- Duhigg, C. (2019, October 10). Is Amazon Unstoppable? *The New Yorker*.
<https://www.newyorker.com/magazine/2019/10/21/is-amazon-unstoppable>
- Dwoskin, E., & Timberg, C. (2018). How merchants use Facebook to flood Amazon with fake reviews. *Washington Post*. https://www.washingtonpost.com/business/economy/how-merchants-secretly-use-facebook-to-flood-amazon-with-fake-reviews/2018/04/23/5dad1e30-4392-11e8-8569-26fda6b404c7_story.html
- eBay. (2020a, November 1). *Seller ratings*. eBay. <https://www.ebay.co.uk/help/buying/resolving-issues-sellers/seller-ratings?id=4023>
- eBay. (2020b, November 1). *Writing item descriptions | eBay Seller Centre | Business*.
<https://sellercentre.ebay.co.uk/business/item-description>
- Eck, J. E. (1994). *Drug markets and drug places: A case-control study of the spatial structure of illicit drug dealing* [Ph.D.]. University of Maryland.
- Eisenbrand, R. (2018, February 26). Händler versenden en masse nicht bestellte Ware – wegen Amazon SEO? *OMR*. <https://omr.com/de/amazon-paket-nicht-bestellt-fake-reviews/>
- Ekblom, P. (1997). Gearing up against crime: A dynamic framework to help designers keep up with the adaptive criminal in a changing world. *International Journal of Risk Security and Crime Prevention*, 249–266.
- ElBahrawy, A., Alessandretti, L., Rusnac, L., Goldsmith, D., Teytelboym, A., & Baronchelli, A. (2020). Collective dynamics of dark web marketplaces. *Scientific Reports*, 10(1), 18827.
<https://doi.org/10.1038/s41598-020-74416-y>
- Elshaar, S., & Sadaoui, S. (2020). Semi-supervised Classification of Fraud Data in Commercial Auctions. *Applied Artificial Intelligence*, 34(1), 47–63.
<https://doi.org/10.1080/08839514.2019.1691341>
- EMCDDA-Europol. (2017). *Drugs and the darknet: Perspectives for enforcement, research and policy*. Publications Office of the European Union.
- Ester, M., Kriegel, H.-P., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press*, 226–231.
- EUIPO. (2019). *2019 Status Report on IPR infringement*. European Union, Intellectual Property Office.
<https://euipo.europa.eu/tunnel->

- web/secure/webdav/guest/document_library/observatory/documents/reports/2019_Status_Report_on_IPR_infringement/2019_Status_Report_on_IPR_infringement_en.pdf
- Europol. (2017). *INTELLECTUAL PROPERTY CRIME ON THE DARKNET*. Enforcement Cooperation. <https://www.europol.europa.eu/publications-documents/intellectual-property-crime-darknet>
- Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1–20. <https://doi.org/10.3758/s13428-021-01694-3>
- Fazzolari, M., Buccafurri, F., Lax, G., & Petrocchi, M. (2021). Experience: Improving Opinion Spam Detection by Cumulative Relative Frequency Distribution. *Journal of Data and Information Quality*, 13(1), 1–16. <https://doi.org/10.1145/3439307>
- FBI. (2017). *2017 Internet Crime Report* (p. 29). Federal Bureau of Investigation.
- FBI. (2018). *Internet Crime Report*. Federal Bureau of Investigation. https://pdf.ic3.gov/2018_IC3Report.pdf
- Federal Bureau of Investigation. (2014). *2014 Internet Crime Report*. U.S Department of Justice. https://pdf.ic3.gov/2014_IC3Report.pdf
- Federal Bureau of Investigation. (2015). *2015 Internet Crime Report* (p. 236). U.S Department of Justice.
- Federal Bureau of Investigation. (2016). *2016 Internet Crime Report*. U.S Department of Justice. https://pdf.ic3.gov/2016_IC3Report.pdf
- Felson, M. (1995). Those who discourage crime. *Crime and Place*, 4.
- Felson, M. (2008). Routine activity approach. In *Environmental Criminology and Crime Analysis*. Willan.
- Feng, G. C. (2015). Mistakes and How to Avoid Mistakes in Using Intercoder Reliability Indices. *Methodology*, 11(1), 13–22. <https://doi.org/10.1027/1614-2241/a000086>
- Flood, A. (2012, September 4). Sock puppetry and fake reviews: Publish and be damned. *The Guardian*. <http://www.theguardian.com/books/2012/sep/04/sock-puppetry-publish-be-damned>
- Ford, B. J., Xu, H., & Valova, I. (2013). A Real-Time Self-Adaptive Classifier for Identifying Suspicious Bidders in Online Auctions. *The Computer Journal*, 56(5), 646–663. <https://doi.org/10.1093/comjnl/bxs025>
- Fornaciari, T., Cagnina, L., Rosso, P., & Poesio, M. (2020). Fake opinion detection: How similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, 54(4), 1019–1058. <https://doi.org/10.1007/s10579-020-09486-5>
- Fornaciari, T., & Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 279–287.
- Freilich, J. D., & Newman, G. R. (2018). Situational Crime Prevention. *Oxford Research Encyclopedia of Criminology*, 29. <https://doi.org/10.1093/acrefore/9780190264079.013.3>
- Ganguly, P. (2015, January 8). How e-retailers such as Flipkart, Amazon are keeping the fake products at bay. *The Economic Times; New Delhi*, 2.
- Garg, V., Afroz, S., Overdorf, R., & Greenstadt, R. (2015). Computer-Supported Cooperative Crime. In R. Böhme & T. Okamoto (Eds.), *Financial Cryptography and Data Security* (Vol. 8975, pp. 32–43). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-47854-7_3

- Gayialis, S. P., Kechagias, E. P., Papadopoulos, G. A., & Masouras, D. (2022). A Review and Classification Framework of Traceability Approaches for Identifying Product Supply Chain Counterfeiting. *Sustainability*, 14(11), 6666. <https://doi.org/10.3390/su14116666>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Gehl, R. W. (2018). Archives for the Dark Web: A Field Guide for Study. In Lewis Levenberg, T. Neilson, & D. Rheams (Eds.), *Research Methods for the Digital Humanities* (pp. 31–51). Springer International Publishing. https://doi.org/10.1007/978-3-319-96713-4_3
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673. <https://doi.org/10.1038/s42256-020-00257-z>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Ghosh, S., Das, A., Porras, P., Yegneswaran, V., & Gehani, A. (2017). Automated Categorization of Onion Sites for Analyzing the Darkweb Ecosystem. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 1793–1802. <https://doi.org/10.1145/3097983.3098193>
- Ghosh, S., Porras, P., Yegneswaran, V., Nitz, K., & Das, A. (2017). ATOL: A Framework for Automated Analysis and Categorization of the Darkweb Ecosystem. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.
- Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.(JAIR)*, 57, 345–420.
- Goldberg, Y., & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722 [Cs, Stat]*. <http://arxiv.org/abs/1402.3722>
- Gray, H. (2019). *Dark Web Map*. <https://www.hyperiongray.com/dark-web-map/#zoom=0.8521016982969332&x=0.5064520330563047&y=0.572866049039204>
- Great Britain, & National Audit Office. (2016). *Protecting consumers from scams, unfair trading and unsafe goods* (HC 851). National Audit Office; Department for Business, Energy & Industrial Strategy.
- Grootendorst, M. (2021). *MaartenGr/KeyBERT: BibTeX* (v0.1.3) [Computer software]. Zenodo. <https://doi.org/10.5281/ZENODO.4461265>
- Gursoy, F., & Kakadiaris, I. A. (2022). *System Cards for AI-Based Decision-Making for Public Policy* (arXiv:2203.04754). arXiv. <https://doi.org/10.48550/arXiv.2203.04754>
- Gutierrez-Espinoza, L., Abri, F., Siami Namin, A., Jones, K. S., & Sears, D. R. W. (2020). Ensemble Learning for Detecting Fake Reviews. *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1320–1325. <https://doi.org/10.1109/COMPSAC48688.2020.00-73>
- Haber, A. C., Sax, U., Prasser, F., & the NFDI4Health Consortium. (2022). Open tools for quantitative anonymization of tabular phenotype data: Literature review. *Briefings in Bioinformatics*, 23(6), bbac440. <https://doi.org/10.1093/bib/bbac440>
- Harper, J. (2021, March 10). Trader gets painted stones instead of \$36m of copper. *BBC News*. <https://www.bbc.com/news/business-56330378>

- Harrison, A. J., Dilla, W. N., & Mennecke, B. E. (2020). Relationships within the Fraud Diamond: The Decision Processes That Influence Fraudulent Intentions in Online Consumer Fraud. *Journal of Information Systems*, 34(1), 61–80. <https://doi.org/10.2308/isys-52627>
- Hartel, P. H., Junger, M., & Wieringa, R. J. (2010). *Cyber-crime Science = Crime Science + Information Security*. <https://research.utwente.nl/en/publications/cyber-crime-science-crime-science-information-security>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- Hayes, D., Cappa, F., & Cardon, J. (2018). A Framework for More Effective Dark Web Marketplace Investigations. *Information*, 9(8), 186. <https://doi.org/10.3390/info9080186>
- He, B., Patel, M., Zhang, Z., & Chang, K. C.-C. (2007). Accessing the deep web. *Communications of the ACM*, 50(5), 94–101. <https://doi.org/10.1145/1230819.1241670>
- Heger, A., Marquis, E. B., Vorvoreanu, M., Wallach, H., & Vaughan, J. W. (2022). *Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata* (arXiv:2206.02923). arXiv. <https://doi.org/10.48550/arXiv.2206.02923>
- Herley, C. (2010). The Plight of the Targeted Attacker in a World of Scale. *The Ninth Workshop on the Economics of Information Security (WEIS) 2010, Harvard University, USA.*, 12.
- Hernandez-Castro, J., & Roberts, D. L. (2015). Automatic detection of potentially illegal online sales of elephant ivory via data mining. *PeerJ Computer Science*, 1, e10. <https://doi.org/10.7717/peerj-cs.10>
- Ho, T. N., & Ng, W. K. (2016). Application of Stylometry to DarkWeb Forum User Identification. In K.-Y. Lam, C.-H. Chi, & S. Qing (Eds.), *Information and Communications Security* (Vol. 9977, pp. 173–183). Springer International Publishing. https://doi.org/10.1007/978-3-319-50011-9_14
- Hollis, M. E., & Wilson, J. (2014). Who are the guardians in product counterfeiting? A theoretical application of routine activities theory. *Crime Prevention and Community Safety*, 16(3), 169–188. <https://doi.org/10.1057/cpcs.2014.6>
- Holt, T. J., & Lee, J. R. (2022). A Crime Script Analysis of Counterfeit Identity Document Procurement Online. *Deviant Behavior*, 43(3), 285–302. <https://doi.org/10.1080/01639625.2020.1825915>
- Holtfreter, K., Reising, M. D., & Pratt, T. C. (2008). Low Self-Control, Routine Activities, and Fraud Victimization*. *Criminology*, 46(1), 189–220. <https://doi.org/10.1111/j.1745-9125.2008.00101.x>
- Homer, E. M. (2020). Testing the fraud triangle: A systematic review. *Journal of Financial Crime*, 27(1), 172–187. <https://doi.org/10.1108/JFC-12-2018-0136>
- Hong, J. (2020, November 25). *Apple's iPhone XR dominates smartphone model shipment ranking in 2019—Omdia*. <https://technology.informa.com/621286/apples-iphone-xr-dominates-smartphone-model-shipment-ranking-in-2019>
- Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* [Computer software].
- Houser, D., & Wooders, J. (2006). Reputation in Auctions: Theory, and Evidence from eBay. *Journal of Economics & Management Strategy*, 15(2), 353–369. <https://doi.org/10.1111/j.1530-9134.2006.00103.x>
- Hughes, J. (2022). *Measuring Agreement Using Krippendorff's Alpha Coefficient (2.0)* [R]. <https://cran.r-project.org/web/packages/krippendorffsalpha/krippendorffsalpha.pdf>

- Hutchings, A. (2013). *Theory and Crime: Does it Compute?* [PhD Thesis, Griffith University].
<https://doi.org/10.25904/1912/2800>
- Hutchings, A. (2018). Leaving on a jet plane: The trade in fraudulently obtained airline tickets. *Crime, Law and Social Change*, 70(4), 461–487. <https://doi.org/10.1007/s10611-018-9777-8>
- Hutchings, A., & Hayes, H. (2009). Routine Activity Theory and Phishing Victimization: Who Gets Caught in the ‘Net’? *Current Issues in Criminal Justice*, 20(3), 433–452.
<https://doi.org/10.1080/10345329.2009.12035821>
- Hutchings, A., & Holt, T. J. (2015). A Crime Script Analysis of the Online Stolen Data Market. *British Journal of Criminology*, 55(3), 596–614. <https://doi.org/10.1093/bjc/azu106>
- Hyslip, T. S., & Holt, T. J. (2019). Assessing the Capacity of DRDoS-For-Hire Services in Cybercrime Markets. *Deviant Behavior*, 40(12), 1609–1625.
<https://doi.org/10.1080/01639625.2019.1616489>
- Ibrahim, M. (2012). *THEMATIC ANALYSIS: A CRITICAL REVIEW OF ITS PROCESS AND EVALUATION*. 1(1).
- Ihaza, J. (2017, August 11). The Instagram watchdog that calls out fashion counterfeits. *The Outline*.
<https://theoutline.com/post/2089/yezybusta-fake-fashion-instagram>
- Intellectual Property Office. (2019). *Annotation task* [Letter to Felix Soldner].
- IP Crime Group. (2015). *IP Crime Report 2014/15* (p. 52). Intellectual Property Office UK.
- Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist*, 11(2), 37–50. <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>
- Jain, A. K., & Gupta, B. B. (2018). Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, 68(4), 687–700.
<https://doi.org/10.1007/s11235-017-0414-0>
- Joffe, H. (2011). Thematic Analysis. In *Qualitative Research Methods in Mental Health and Psychotherapy* (pp. 209–223). John Wiley & Sons, Ltd.
<https://doi.org/10.1002/9781119973249.ch15>
- Johnson, S. D., Guerette, R. T., & Bowers, K. (2014). Crime displacement: What we know, what we don’t know, and what it means for crime reduction. *Journal of Experimental Criminology*, 10(4), 549–571. <https://doi.org/10.1007/s11292-014-9209-4>
- Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition—Third Edition draft*.
<https://web.stanford.edu/~jurafsky/slp3/>
- Jurásek, B., Čmelo, I., Svoboda, J., Čejka, J., Svozil, D., & Kuchař, M. (2021). New psychoactive substances on dark web markets: From deal solicitation to forensic analysis of purchased substances. *Drug Testing and Analysis*, 13(1), 156–168. <https://doi.org/10.1002/dta.2901>
- Kamps, J., Trozze, A., & Kleinberg, B. (2022). Cryptocurrencies: Boons and curses for fraud prevention. In *A Fresh Look at Fraud*. Routledge.
- Karami, E., Prasad, S., & Shehata, M. (2017). *Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images*.
- Keegan, J. (2020, July 21). Is This Amazon Review Bullshit? – The Markup. *The Markup*.
<https://themarkup.org/ask-the-markup/2020/07/21/how-to-spot-fake-amazon-product-reviews>
- Kemp, S., & Erades Pérez, N. (2023). Consumer Fraud against Older Adults in Digital Society: Examining Victimization and Its Impact. *International Journal of Environmental Research and Public Health*, 20(7), 5404. <https://doi.org/10.3390/ijerph20075404>

- Kennedy, J. P., Haberman, C. P., & Wilson, J. M. (2018). Occupational Pharmaceutical Counterfeiting Schemes: A Crime Scripts Analysis. *Victims & Offenders, 13*(2), 196–214. <https://doi.org/10.1080/15564886.2016.1217961>
- Kigerl, A. (2021). Routine activity theory and malware, fraud, and spam at the national level. *Crime, Law and Social Change, 76*(2), 109–130. <https://doi.org/10.1007/s10611-021-09957-y>
- Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology, 72*(6), 558–569. <https://doi.org/10.4097/kja.19087>
- Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. 1–15.
- Kleinberg, B. (2019). *Towards detecting deceptive intentions on a large scale*. [Ph.D.]. University of Amsterdam.
- Kleinberg, B., Davies, T., & Mozes, M. (2022). *Textwash—Automated open-source text anonymisation* (arXiv:2208.13081). arXiv. <https://doi.org/10.48550/arXiv.2208.13081>
- Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta Psychologica, 213*, 103250. <https://doi.org/10.1016/j.actpsy.2020.103250>
- Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement, 30*(1), 61–70. <https://doi.org/10.1177/001316447003000105>
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573–603. <https://doi.org/10.1037/a0029146>
- Kumar, D. G. R., & Gunasekaran, D. S. (2019). URL PHISHING DATA ANALYSIS AND DETECTING PHISHING ATTACKS USING MACHINE LEARNING IN NLP. *3*(10), 7.
- Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 957–966). PMLR. <http://proceedings.mlr.press/v37/kusnerb15.html>
- Ladegaard, I. (2019). Crime displacement in digital drug markets. *International Journal of Drug Policy, 63*, 113–121. <https://doi.org/10.1016/j.drugpo.2018.09.013>
- Lai, S., Wu, J., Ma, Z., & Ye, C. (2023). BTextCAN: Consumer fraud detection via group perception. *Information Processing & Management, 60*(3), 103307. <https://doi.org/10.1016/j.ipm.2023.103307>
- Lee, C. (2020). A crime script analysis of transnational identity fraud: Migrant offenders' use of technology in South Korea. *Crime, Law and Social Change, 74*(2), 201–218. <https://doi.org/10.1007/s10611-020-09885-3>
- Lee, D. (2020, September 4). Amazon deletes 20,000 reviews after evidence of profits for posts. *Financial Times*. <https://www.ft.com/content/bb03ba1c-add3-4440-9bf2-2a65566aef4a>
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research, 18*(17), 1–5.
- Lenaerts-Bergmans, B. (2023, April 27). *What is Dark Web Monitoring? [Beginner's Guide]* - CrowdStrike. CrowdStrike.Com. <https://www.crowdstrike.com/cybersecurity-101/dark-web-monitoring/>
- Levenshtein, V. I. & others. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, Article 8.

- Levi, M. (2008). Organized fraud and organizing frauds Unpacking research on networks and organization. *Criminology & Criminal Justice - CRIMINOL CRIM JUSTICE*, 8, 389–419. <https://doi.org/10.1177/1748895808096470>
- Levitan, S. I., An, G., Ma, M., Levitan, R., Rosenberg, A., & Hirschberg, J. (2016). Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection. *Interspeech 2016*, 2006–2010. <https://doi.org/10.21437/Interspeech.2016-1519>
- Lewis, S., Jamie. (2016, July 3). *OnionScan Report June 2016—Snapshots of the Dark Web*. Mascherari Press. <https://mascherari.press/onionscan-report-june-2016/>
- Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1566–1576. <https://doi.org/10.3115/v1/P14-1147>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, 6(1), 1–39. <https://doi.org/10.1145/2133360.2133363>
- Loria, S. (2021). *Sloria/TextBlob* [Python]. <https://github.com/sloria/TextBlob> (Original work published 2013)
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, 2, 1150–1157. <https://doi.org/10.1109/iccv.1999.790410>
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- M. Bhoot, A., A. Shinde, M., & P. Mishra, W. (2021). Towards the Identification of Dark Patterns: An Analysis Based on End-User Reactions. *Proceedings of the 11th Indian Conference on Human-Computer Interaction*, 24–33. <https://doi.org/10.1145/3429290.3429293>
- Maimon, D., Wu, Y., Stubler, N., & Singirikonda, P. (2020). Extended Validation in the Dark Web: Evidence from Investigation of the Certification Services and Products Sold on Darknet Markets. *EBCS Reports*. https://scholarworks.gsu.edu/ebsc_reports/2
- Männistö, T., Morini, C., & Hintsa, J. (2021). *Customs Innovations for Fighting Fraud and Trafficking in Cross-border Parcel Flows*.
- Mansfield-Devine, S. (2009). Darknets. *Computer Fraud & Security*, 2009(12), 4–6. [https://doi.org/10.1016/S1361-3723\(09\)70150-2](https://doi.org/10.1016/S1361-3723(09)70150-2)
- Marin, E., Diab, A., & Shakarian, P. (2016). Product Offerings in Malicious Hacker Markets. *arXiv:1607.07903 [Cs]*. <http://arxiv.org/abs/1607.07903>
- Marucheck, A., Greis, N., Mena, C., & Cai, L. (2011). Product safety and security in the global supply chain: Issues, challenges and research opportunities. *Journal of Operations Management*, 29(7–8), 707–720. <https://doi.org/10.1016/j.jom.2011.06.007>
- Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 81:1-81:32. <https://doi.org/10.1145/3359183>
- Meer, L. van der, Claeysens, S., & Brandt, M. (2022, June 23). *SANE - Secure ANalysis Environment @SURF Cloud*. <https://doi.org/10.5281/zenodo.7074192>
- Melnik, M. I., & Alm, J. (2003). Does a Seller's eCommerce Reputation Matter? Evidence from eBay Auctions. *The Journal of Industrial Economics*, 50(3), 337–349. <https://doi.org/10.1111/1467-6451.00180>

- Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 309–312.
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272. <https://aclanthology.org/D11-1024>
- Miramirakhani, N., Starov, O., & Nikiforakis, N. (2017). Dial One for Scam: A Large-Scale Analysis of Technical Support Scams. *Proceedings 2017 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA. <https://doi.org/10.14722/ndss.2017.23163>
- Mishra, V. (2020, February 27). iPhone 11: Second Best Selling Phone of 2019 Globally After Less Than Four Months. *Counterpoint Research*. <https://www.counterpointresearch.com/iphone-11-second-best-selling-phone-2019-less-four-months/>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Mohawesh, R., Xu, S., Springer, M., Al-Hawawreh, M., & Maqsood, S. (2021). Fake or Genuine? Contextualised Text Representation for Fake Review Detection. *Natural Language Processing*, 137–148. <https://doi.org/10.5121/csit.2021.112311>
- Mooij, D. (2018, March 17). 70% of counterfeit products are sold online. *Seal Network*. <https://medium.com/sealnetwork/70-of-counterfeit-products-are-sold-online-c6eafe07083>
- Mui, G., & Mailley, J. (2015). A tale of two triangles: Comparing the Fraud Triangle with criminology's Crime Triangle. *Accounting Research Journal*, 28(1), 45–58. <https://doi.org/10.1108/ARJ-10-2014-0092>
- Muja, M., & Lowe, D. (2011). FLANN - Fast Library for Approximate Nearest Neighbors User Manual. *Visapp*, 331–340.
- Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013a). Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. *Technical Report, Department of Computer Science, University of Illinois at Chicago*.
- Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013b). What Yelp Fake Review Filter Might Be Doing? *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), Article 1.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
- Nadini, M., Bracci, A., ElBahrawy, A., Gradwell, P., Teytelboym, A., & Baronchelli, A. (2021). Emergence and structure of decentralised trade networks around dark web marketplaces. *arXiv:2111.01774 [Physics]*. <http://arxiv.org/abs/2111.01774>
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5), 544–551. <https://doi.org/10.1136/amiajnl-2011-000464>
- Nagi Alsubari, S., N. Deshmukh, S., Abdullah Alqarni, A., Alsharif, N., H. H. Aldhyani, T., Waselallah Alsaade, F., & I. Khalaf, O. (2022). Data Analytics for the Identification of Fake Reviews Using Supervised Learning. *Computers, Materials & Continua*, 70(2), 3189–3204. <https://doi.org/10.32604/cmc.2022.019625>
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5), 665–675.

- Ngo, F. T., Piquero, A. R., LaPrade, J., & Duong, B. (2020). Victimization in Cyberspace: Is It How Long We Spend Online, What We Do Online, or What We Post Online? *Criminal Justice Review*, 45(4), 430–451. <https://doi.org/10.1177/0734016820934175>
- Nguyen, N. (2018). Inside The Ecosystem That Fuels Amazon’s Fake Review Problem. *BuzzFeed News*. <https://www.buzzfeednews.com/article/nicolenguyen/amazon-fake-review-problem>
- Nikitkov, A. N., Stone, D. N., & Miller, T. C. (2014). Internal Controls, Routine Activity Theory (RAT), and Sustained Online Auction Deception: A Longitudinal Analysis. *Journal of Information Systems*, 28(1), 311–337. <https://doi.org/10.2308/isys-50708>
- OECD (Ed.). (2018). *Trade in counterfeit goods and free trade zones: Evidence for recent trends*. OECD Publishing.
- OECD/EUIPO. (2019). *Trends in Trade in Counterfeit and Pirated Goods*. OECD Publishing. <http://www.library.yorku.ca/e/resolver/id/287746375>
- Office for National Statistics. (2020). *Nature of fraud and computer misuse in England and Wales: Year ending March 2019* (p. 29). Office for National Statistics.
- Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2. <https://www.frontiersin.org/articles/10.3389/fdata.2019.00013>
- Ortega, A., & Navarrete, G. (2017). Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing (NHST) in Psychology and Social Sciences. In J. P. Tejedor (Ed.), *Bayesian Inference*. InTech. <https://doi.org/10.5772/intechopen.70230>
- Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and ‘Experimental’ proportionality. *Information & Communications Technology Law*, 27(2), 223–250. <https://doi.org/10.1080/13600834.2018.1458455>
- Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 497–501.
- Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 309–319.
- Palmer, A. (2020, July 30). Amazon sales soar as pandemic fuels online shopping. *CNBC*. <https://www.cnn.com/2020/07/30/amazon-amzn-earnings-q2-2020.html>
- Pandit, S., Chau, D. H., Wang, S., & Faloutsos, C. (2007). Netprobe: A fast and scalable system for fraud detection in online auction networks. *Proceedings of the 16th International Conference on World Wide Web - WWW '07*, 201. <https://doi.org/10.1145/1242572.1242600>
- Paterva. (2019). *Paterva Home*. PATERVA A New Train of Thought. <https://www.paterva.com/index.php>
- Paul, K. (2018). Ancient Artifacts vs. Digital Artifacts: New Tools for Unmasking the Sale of Illicit Antiquities on the Dark Web. *Arts*, 7(2), 12. <https://doi.org/10.3390/arts7020012>
- Peaston, S. (2019). *The Fraudscape* (p. 19). Cifas. <https://www.cifas.org.uk/secure/contentPORT/uploads/documents/Cifas%20Fraudscape%202019%20Full%20Digital%20Report%20.pdf>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 12, 2825–2830.

- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Pele, O., & Werman, M. (2008). A Linear Time Histogram Metric for Improved SIFT Matching. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer Vision – ECCV 2008* (pp. 495–508). Springer Berlin Heidelberg.
- Pele, O., & Werman, M. (2009). Fast and robust Earth Mover’s Distances. *2009 IEEE 12th International Conference on Computer Vision, 460–467*. <https://doi.org/10.1109/ICCV.2009.5459199>
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015* (Version 2015) [Computer software]. Pennebaker Conglomerates. www.LIWC.net
- Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. *Proceedings of the 27th International Conference on Computational Linguistics, 3391–3401*. <https://www.aclweb.org/anthology/C18-1287>
- Perez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural Deception Detection. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), 6*.
- Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1120–1125*. <https://doi.org/10.18653/v1/D15-1133>
- Peteranderl, S. (2019, December 2). Amazon: Was hinter den mysteriösen Überraschungspaketen steckt - DER SPIEGEL - Netzwelt. *Spiegel*. <https://www.spiegel.de/netzwelt/web/amazon-was-hinter-den-mysterioesen-ueberraschungspaketen-steckt-a-1252822.html>
- Piliavin, I., Gartner, R., Thornton, C., & Matsueda, R. L. (1986). Crime, Deterrence, and Rational Choice. *American Sociological Review, 51*(1), 101–119. <https://doi.org/10.2307/2095480>
- Pointer Brand Protection. (2019). *Online Brand Protection With An Impact*. Pointer Brand Protection. <https://pointerbrandprotection.com/>
- Pratt, T. C., Holtfreter, K., & Reisig, M. D. (2010). Routine Online Activity and Internet Fraud Targeting: Extending the Generality of Routine Activity Theory. *Journal of Research in Crime and Delinquency, 47*(3), 267–296. <https://doi.org/10.1177/0022427810365903>
- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data, 1*(1), 51–59. <https://doi.org/10.1089/big.2013.1508>
- Rahman, M., Carbanar, B., Ballesteros, J., & Chau, D. H. (Polo). (2015). To catch a fake: Curbing deceptive Yelp ratings and venues. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 8*(3), 147–161. <https://doi.org/10.1002/sam.11264>
- Raine, J., Mangan, C., & Watt, P. (2015). *THE IMPACT OF LOCAL AUTHORITY TRADING STANDARDS IN CHALLENGING TIMES* (p. 148). The Department for Business, Innovation and Skills and The Trading Standards Institute.
- Rajeev, A. (2019). *SelectorLib* [Computer software]. /
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [Cs]*. <http://arxiv.org/abs/1908.10084>
- Ren, Y., & Ji, D. (2019). Learning to Detect Deceptive Opinion Spam: A Survey. *IEEE Access, 7*, 42934–42945. <https://doi.org/10.1109/ACCESS.2019.2908495>
- Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45–48. <https://doi.org/10.1145/355112.355122>

- Reyns, B. W., & Randa, R. (2020). No honor among thieves: Personal and peer deviance as explanations of online identity fraud victimization. *Security Journal*, 33(2), 228–243. <https://doi.org/10.1057/s41284-019-00182-w>
- Rhumorbarbe, D., Staehli, L., Broséus, J., Rossy, Q., & Esseiva, P. (2016). Buying drugs on a Darknet market: A better deal? Studying the online illicit drug market through the analysis of digital, physical and chemical data. *Forensic Science International*, 267, 173–182. <https://doi.org/10.1016/j.forsciint.2016.08.032>
- Richardson, L. (2019). *Beautiful Soup* (4.7.1) [Python]. <https://www.crummy.com/software/BeautifulSoup/>
- Ríssola, E. A., Aliannejadi, M., & Crestani, F. (2020). Beyond Modelling: Understanding Mental Disorders in Online Social Media. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, & F. Martins (Eds.), *Advances in Information Retrieval* (Vol. 12035, pp. 296–310). Springer International Publishing. https://doi.org/10.1007/978-3-030-45439-5_20
- Roberts, D. L., & Hernandez-Castro, J. (2017). Bycatch and illegal wildlife trade on the dark web. *Oryx*, 51(3), 393–394. <https://doi.org/10.1017/S0030605317000679>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). **stm**: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2). <https://doi.org/10.18637/jss.v091.i02>
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4), 1064–1082. <https://doi.org/10.1111/ajps.12103>
- Rodrigues, V. F., Policarpo, L. M., da Silveira, D. E., da Rosa Righi, R., da Costa, C. A., Barbosa, J. L. V., Antunes, R. S., Scorsatto, R., & Arcot, T. (2022). Fraud detection and prevention in e-commerce: A systematic literature review. *Electronic Commerce Research and Applications*, 56, 101207. <https://doi.org/10.1016/j.eleap.2022.101207>
- Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass*. <https://doi.org/10.1111/spc3.12579>
- Ruble, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571. <https://doi.org/10.1109/ICCV.2011.6126544>
- Rusch, J. J. (1991). *The “Social Engineering” of Internet Fraud*. 12. http://www.isoc.org/isoc/conferences/inet/99/proceedings/3g/3g_2.htm
- Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357. <https://doi.org/10.1016/j.eswa.2018.09.029>
- Salvetti, F., Lowe, J. B., & Martin, J. H. (2016). A tangled web: The faint signals of deception in text—Boulder lies and truth corpus (BLT-C). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 3510–3517. <https://www.aclweb.org/anthology/L16-1558>
- Sampson, R., Eck, J. E., & Dunham, J. (2010). Super controllers and crime prevention: A routine activity explanation of crime prevention success and failure. *Security Journal*, 23(1), 37–51. <https://doi.org/10.1057/sj.2009.17>
- Santos, A. S. dos, Camargo, L. F. R., & Lacerda, D. P. (2020). Evaluation of classification techniques for identifying fake reviews about products and services on the internet. *Gestão & Produção*, 27. <https://doi.org/10.1590/0104-530X4672-20>

- Sarno, D. M., & Black, J. (2023). Who Gets Caught in the Web of Lies?: Understanding Susceptibility to Phishing Emails, Fake News Headlines, and Scam Text Messages. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 001872082311732. <https://doi.org/10.1177/00187208231173263>
- Schafer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., & Lenders, V. (2019). BlackWidow: Monitoring the Dark Web for Cyber Security Information. *2019 11th International Conference on Cyber Conflict (CyCon)*, 1–21. <https://doi.org/10.23919/CYCON.2019.8756845>
- Scheck, A. B., Shane Shifflett and Justin. (2019, August 23). Amazon Has Ceded Control of Its Site. The Result: Thousands of Banned, Unsafe or Mislabeled Products. *Wall Street Journal*. <https://www.wsj.com/articles/amazon-has-ceded-control-of-its-site-the-result-thousands-of-banned-unsafe-or-mislabeled-products-11566564990>
- Schiffer, Z. (2020, September 4). Amazon takes down a five-star fraud in the UK. *The Verge*. <https://www.theverge.com/2020/9/4/21423429/amazon-top-reviewers-uk-fraud>
- Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*, 85(S1), 399–422. <https://doi.org/10.1093/poq/nfab018>
- Sergi, A. (2022). Playing Pac-Man in Portville: Policing the dilution and fragmentation of drug importations through major seaports. *European Journal of Criminology*, 19(4), 674–691. <https://doi.org/10.1177/1477370820913465>
- Shah, H. S., Joshi, N. R., Sureka, A., & Wurman, P. R. (2003). Mining eBay: Bidding Strategies and Skill Detection. In O. R. Zaiane, J. Srivastava, M. Spiliopoulou, & B. Masand (Eds.), *WEBKDD 2002—Mining Web Data for Discovering Usage Patterns and Profiles* (Vol. 2703, pp. 17–34). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-39663-5_2
- Shang, Y., Wang, K., Tian, Y., Zhou, Y., Ma, B., & Liu, S. (2023). Theoretical basis and occurrence of internet fraud victimisation: Based on two systems in decision-making and reasoning. *Frontiers in Psychology*, 14. <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1087463>
- Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., & Nadali, S. (2013). Detecting deceptive reviews using lexical and syntactic features. *2013 13th International Conference on Intelligent Systems Design and Applications*, 53–58. <https://doi.org/10.1109/ISDA.2013.6920707>
- Simpson, S. S., Rorie, M., Alper, M., Schell-Busey, N., Laufer, W. S., & Smith, N. C. (2014). Corporate Crime Deterrence: A Systematic Review. *Campbell Systematic Reviews*, 10(1), 1–105. <https://doi.org/10.4073/csr.2014.4>
- Singh, A., & Chatterjee, K. (2022). A Comparative Approach for Opinion Spam Detection Using Sentiment Analysis. In S. Rawat, A. Kumar, P. Kumar, & J. Anguera (Eds.), *Proceedings of First International Conference on Computational Electronics for Wireless Communications* (Vol. 329, pp. 511–522). Springer Singapore. https://doi.org/10.1007/978-981-16-6246-1_43
- Soper, S. (2021, January 11). eBay Growth Ebbs, Sparking Concern Pandemic Boost Is Over. *Bloomberg.Com*. <https://www.bloomberg.com/news/articles/2020-10-28/ebay-gives-disappointing-sales-forecast-for-holiday-quarter>
- Soska, K., & Christin, N. (2015). Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. *Proceedings of the 24th USENIX Security Symposium*, 33–48.
- Spink, J., Moyer, D. C., Park, H., & Heinonen, J. A. (2013). Defining the types of counterfeiters, counterfeiting, and offender organizations. *Crime Science*, 2(1), 8. <https://doi.org/10.1186/2193-7680-2-8>

- Spink, J., Moyer, D. C., Park, H., & Heinonen, J. A. (2014). Development of a product-counterfeiting incident cluster tool. *Crime Science*, 3(1). <https://doi.org/10.1186/s40163-014-0003-4>
- Stanford Center on Longevity. (2019). *Financial Fraud – Stanford Center on Longevity*. <http://longevity.stanford.edu/2017/03/29/safeguarding-clients-from-financial-fraud-and-exploitation/>
- Streitfeld, D. (2011, August 19). In a Race to Out-Rave, 5-Star Web Reviews Go for \$5 (Published 2011). *The New York Times*. <https://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html>
- Sullivan, B. A., Chan, F., Fenoff, R., & Wilson, J. M. (2017). Assessing the developing knowledge-base of product counterfeiting: A content analysis of four decades of research. *Trends in Organized Crime*, 20(3), 338–369. <https://doi.org/10.1007/s12117-016-9300-5>
- Suthivarakom, G. (2020, February 11). Welcome to the Era of Fake Products. *Wirecutter: Reviews for the Real World*. <https://www.nytimes.com/wirecutter/blog/amazon-counterfeit-fake-products/>
- Swearingen, J. (2019, August 26). Hijacked Reviews on Amazon Can Trick Shoppers. *Consumer Reports*. <https://www.consumerreports.org/customer-reviews-ratings/hijacked-reviews-on-amazon-can-trick-shoppers/>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. <https://doi.org/10.1109/CVPR.2015.7298594>
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem*, 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- Tang, C. S. (2006). Perspectives in supply chain risk management. *International Journal of Production Economics*, 103(2), 451–488. <https://doi.org/10.1016/j.ijpe.2005.12.006>
- Tareen, S. A. K., & Saleem, Z. (2018). A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK. *2018 International Conference on Computing, Mathematics and Engineering Technologies: Invent, Innovate and Integrate for Socioeconomic Development, iCoMET 2018 - Proceedings, 2018-Janua*, 1–10. <https://doi.org/10.1109/ICOMET.2018.8346440>
- Tcherni, M., Davies, A., Lopes, G., & Lizotte, A. (2016). The Dark Figure of Online Property Crime: Is Cyberspace Hiding a Crime Wave? *Justice Quarterly*, 33(5), 890–911. <https://doi.org/10.1080/07418825.2014.994658>
- The Invisible Internet Project. (2020). *I2P Anonymous Network*. <https://geti2p.net/en/>
- The Tor Project, Inc. (2020). *The Tor Project | Privacy & Freedom Online*. <https://torproject.org>
- Thomas, D. (2018, August 23). We took a trip to the fake sneaker capital of China. *Vice*. <https://www.vice.com/en/article/d3e9mw/we-took-a-trip-to-the-fake-sneaker-capital-of-china>
- Thomas, K., Huang, D. Y., Wang, D., Bursztein, E., Grier, C., Holt, T. J., Kruegel, C., McCoy, D., Savage, S., & Vigna, G. (2015). Framing Dependencies Introduced by Underground Commoditization. *14th Workshop on the Economics of Information Security*. WEIS 2015, Delft, Netherlands.
- Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. *Proceedings of the IEEE International Conference on Computer Vision*, 839–846. <https://doi.org/10.1109/iccv.1998.710815>

- Tompson, L., Steinbach, R., Johnson, S. D., Teh, C. S., Perkins, C., Edwards, P., & Armstrong, B. (2023). Absence of Street Lighting May Prevent Vehicle Crime, but Spatial and Temporal Displacement Remains a Concern. *Journal of Quantitative Criminology*, 39(3), 603–623. <https://doi.org/10.1007/s10940-022-09539-8>
- Trading Standards UK. (2019, February 26). *Interview about possible data science solutions* [Personal communication].
- Trading Standards UK. (2019, May 17). *Interview about Trading Standards areas of operations, and current issues concerning online crime* [Personal communication].
- Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1), 191–211. [https://doi.org/10.1016/0304-3975\(92\)90143-4](https://doi.org/10.1016/0304-3975(92)90143-4)
- UNICRI, & ICC BASCAP. (2013). *Confiscation of the Proceeds of Crime: A Modern Tool for Deterring Counterfeiting and Piracy and Executive Summary*. United Nations Interregional Crime and Justice Research Institute, International Chamber of Commerce ‘Business Action to Stop Counterfeiting and Piracy.’
- UNODC. (2014). *The Illicit Trafficking of Counterfeit Goods and Transnational Organized Crime*. United Nations Office on Drugs and Crime. https://www.unodc.org/documents/counterfeit/FocusSheet/Counterfeit_focussheet_EN_HIRES.pdf
- Van Buskirk, J., Naicker, S., Bruno, R. B., Breen, C., & Roxburgh, A. (2016). *Drugs and the Internet*. https://www.drugsandalcohol.ie/20369/1/NDARC_Drugs&TheInternet_Bulletin1.pdf
- Van Buskirk, J., Roxburgh, A., Farrell, M., & Burns, L. (2014). The closure of the Silk Road: What has this meant for online drug trading?: Editorial. *Addiction*, 109(4), 517–518. <https://doi.org/10.1111/add.12422>
- Van Buskirk, J., Roxburgh, A., Naicker, S., & Burns, L. (2015). A response to Dolliver’s “Evaluating drug trafficking on the Tor network.” *International Journal of Drug Policy*, 26(11), 1126–1127. <https://doi.org/10.1016/j.drugpo.2015.07.001>
- van der Vegt, I. (2021). *Linguistic Threat Assessment: Understanding Targeted Violence through Computational Linguistics* [PhD Thesis, UCL (University College London)]. https://discovery.ucl.ac.uk/id/eprint/10124525/1/thesis_iwjvandervegt_final.pdf
- van der Vegt, I., Gil, P., Macdonald, S., & Kleinberg, B. (2019). Shedding Light on Terrorist and Extremist Content Removal. *Global Research Network on Terrorism and Technology*, 3, 11.
- van der Vegt, I., & Kleinberg, B. (2020). Women Worry About Family, Men About the Economy: Gender Differences in Emotional Responses to COVID-19. In S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, & D. Pedreschi (Eds.), *Social Informatics* (Vol. 12467, pp. 397–409). Springer International Publishing. https://doi.org/10.1007/978-3-030-60975-7_29
- van der Vegt, I., Kleinberg, B., & Gill, P. (2022). *Proceed with Caution: On the Use of Computational Linguistics in Threat Assessment*. PsyArXiv. <https://doi.org/10.31234/osf.io/ncq9d>
- van der Vegt, I., Mozes, M., Gill, P., & Kleinberg, B. (2021). Online influence, offline violence: Language use on YouTube surrounding the ‘Unite the Right’ rally. *Journal of Computational Social Science*, 4(1), 333–354. <https://doi.org/10.1007/s42001-020-00080-x>
- van Wegberg, R., Tajalizadehkhoob, S., Soska, K., Akyazi, U., Ganan, C., Klievink, B., Christin, N., & van Eeten, M. (2018). Plug and Prey? Measuring the Commoditization of Cybercrime via Online Anonymous Markets. *Proceedings of the 27th USENIX Security Symposium*, 1009–1026.

- Vistalworks. (2019). *Vistalworks—Keeps online shoppers safe from harm*. Vistalworks.
<https://vistalworks.com>
- Walters, J. H., & Langton, L. (2013). *Household Burglary, 1994-2011*.
- Wang, X. (2018). *Photo-based Vendor Re-identification on Darknet Marketplaces using Deep Neural Networks* [Master Thesis]. Faculty of the Virginia Polytechnic Institute and State University.
- Wang, X., Peng, P., Wang, C., & Wang, G. (2018). You Are Your Photographs: Detecting Multiple Identities of Vendors in the Darknet Marketplaces. *Proceedings of the 2018 on Asia Conference on Computer and Communications Security - ASIACCS '18*, 431–442.
<https://doi.org/10.1145/3196494.3196529>
- Watson, J. (2018). *Aspects of Online Reviews and Their Effects in Consumer Decisions* [ProQuest Dissertations Publishing]. <http://search.proquest.com/docview/2078904637/?pq-origsite=primo>
- Weise, K. (2020, November 27). Pushed by Pandemic, Amazon Goes on a Hiring Spree Without Equal. *The New York Times*. <https://www.nytimes.com/2020/11/27/technology/pushed-by-pandemic-amazon-goes-on-a-hiring-spree-without-equal.html>
- Weng, H., Ji, S., Duan, F., Li, Z., Chen, J., He, Q., & Wang, T. (2019). CATS: Cross-Platform E-Commerce Fraud Detection. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 1874–1885. <https://doi.org/10.1109/ICDE.2019.00203>
- Weng, H., Li, Z., Ji, S., Chu, C., Lu, H., Du, T., & He, Q. (2018). Online E-Commerce Fraud: A Large-Scale Detection and Analysis. *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, 1435–1440. <https://doi.org/10.1109/ICDE.2018.00162>
- Williams, M. L. (2016). Guardians Upon High: An Application of Routine Activities Theory to Online Identity Theft in Europe at the Country and Individual Level. *British Journal of Criminology*, 56(1), 21–48. <https://doi.org/10.1093/bjc/azv011>
- Woolf, M. (2014, June 17). *A Statistical Analysis of 1.2 Million Amazon Reviews*. Max Woolf's Blog. <https://minimaxir.com/2014/06/reviewing-reviews/>
- WTO. (1994). *TRIPS: Agreement on Trade-Related Aspects of Intellectual Property Rights*. WTO.
https://www.wto.org/english/tratop_e/trips_e/ta_docs_e/1_tripsandconventions_e.pdf
- Xu, J.-C., Shin, K., & Liu, Y.-L. (2016). Detecting Fake Sites based on HTML Structure Analysis. *Proceedings of the 6th International Conference on Communication and Network Security - ICCNS '16*, 86–90. <https://doi.org/10.1145/3017971.3017980>
- Yellow Brand Protection. (2019). *Anti-Counterfeiting | Yellow Brand Protection*.
<https://www.yellowbrandprotection.com/services/anti-counterfeiting>
- Zambiasi, D. (2022). Drugs on the Web, Crime in the Streets. The Impact of Shutdowns of Dark Net Marketplaces on Street Crime. *Journal of Economic Behavior & Organization*, 202, 274–306.
<https://doi.org/10.1016/j.jebo.2022.08.008>
- Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews. *Journal of Management Information Systems*, 33(2), 456–481.
<https://doi.org/10.1080/07421222.2016.1205907>
- Zhang, G., Li, Z., Huang, J., Wu, J., Zhou, C., Yang, J., & Gao, J. (2022). eFraudCom: An E-commerce Fraud Detection System via Competitive Graph Neural Networks. *ACM Transactions on Information Systems*, 40(3), 1–29. <https://doi.org/10.1145/3474379>
- Zimmerman, B. (2020, January 22). Council Post: Why Nike Cut Ties With Amazon And What It Means For Other Retailers. *Forbes*.

<https://www.forbes.com/sites/forbesbusinesscouncil/2020/01/22/why-nike-cut-ties-with-amazon-and-what-it-means-for-other-retailers/>

Appendix A: Background and literature review

A1 Table. 25 techniques of situational crime prevention from Cornish & Clarke (2003)

Increase the Effort	Increase the Risks	Reduce the Rewards	Reduce Provocations	Remove Excuses
1. Target harden	6. Extend guardianship	11. Conceal targets	16. Reduce frustrations and stress	21. Set rules
2. Control access to facilities	7. Assist natural surveillance	12. Remove targets	17. Avoid disputes	22. Post instructions
3. Screen exits	8. Reduce anonymity	13. Identify property	18. Reduce emotional arousal	23. Alert conscience
4. Deflect offenders	9. Utilize place managers	14. Disrupt markets	19. Neutralize peer pressure	24. Assist compliance
5. Control tools/ weapons	10. Strengthen formal surveillance	15. Deny benefits	20. Discourage imitation	25. Control drugs and alcohol

https://www.ebay.co.uk/b/Multipurpose-Batteries-Power/48446/bn_1630451
https://www.ebay.co.uk/b/Other-Sound-Vision/175837/bn_1842037
https://www.ebay.co.uk/b/Performance-DJ-Equipment/48458/bn_1676803
https://www.ebay.co.uk/b/Portable-Disc-Players-Radios/175747/bn_1632017
https://www.ebay.co.uk/b/Smart-Glasses/178894/bn_1842027
https://www.ebay.co.uk/b/Sound-Vision-Manuals-Resources/48644/bn_1840124
https://www.ebay.co.uk/b/TV-Home-Audio-Accessories/14961/bn_1839784
https://www.ebay.co.uk/b/TV-Reception-Set-Top-Boxes/15069/bn_1838822
https://www.ebay.co.uk/b/TV-Home-Audio-Accessories/14961/bn_1839784
https://www.ebay.co.uk/b/Televisions/11071/bn_1839641
https://www.ebay.co.uk/b/Vintage-Sound-Vision/175740/bn_1634826
https://www.ebay.co.uk/b/iPod-and-MP3-Player-Accessories/56169/bn_877451
https://www.ebay.co.uk/b/MP3-Players/73839/bn_875530
https://www.ebay.co.uk/b/Coin-Operated-Gaming/3944/bn_1842041
https://www.ebay.co.uk/b/Video-Game-Controllers-Attachments/117042/bn_552150
https://www.ebay.co.uk/b/Video-Game-Headsets/171821/bn_198270
https://www.ebay.co.uk/b/Video-Game-Manuals-Inserts-Box-Art/182174/bn_16571093
https://www.ebay.co.uk/b/Video-Game-Memory-Cards-Expansion-Packs/117045/bn_550249
https://www.ebay.co.uk/b/Original-Video-Game-Cases-Boxes/182175/bn_16569357
https://www.ebay.co.uk/b/Other-Video-Games-Consoles/187/bn_1841452
https://www.ebay.co.uk/b/Prepaid-Gaming-Cards/156597/bn_1838659
https://www.ebay.co.uk/b/Video-Game-Replacement-Parts-Tools/171833/bn_450773
https://www.ebay.co.uk/b/Video-Game-Strategy-Guides-Cheats/156595/bn_552311
https://www.ebay.co.uk/b/Video-Game-Accessories/54968/bn_1634843
https://www.ebay.co.uk/b/Video-Game-Consoles/139971/bn_450873
https://www.ebay.co.uk/b/Video-Game-Merchandise/38583/bn_450739
https://www.ebay.co.uk/b/Video-Games/139973/bn_450842
https://www.ebay.co.uk/b/Clothing-Shoes-Accessories-Wholesale-Lots/41964/bn_1841091
https://www.ebay.co.uk/b/Computing-Wholesale-Job-Lots/45090/bn_1838745
https://www.ebay.co.uk/b/Consumer-Electronics-Wholesale-Job-Lots/51004/bn_1838810
https://www.ebay.co.uk/b/DVDs-Films-TV-Wholesale-Job-Lots/31606/bn_1841068
https://www.ebay.co.uk/b/Jewelery-Watches-Wholesale-Job-Lots/40131/bn_1839526
https://www.ebay.co.uk/b/Mobile-Home-Phones-Wholesale-Job-Lots/45065/bn_1841752
https://www.ebay.co.uk/b/PC-Video-Gaming-Wholesale-Job-Lots/31583/bn_1839382
https://www.ebay.co.uk/b/Hair-Dryers/11858/bn_2314554
https://www.ebay.co.uk/b/Electric-Shavers/180512/bn_1676366

Appendix C: Confounds and Overestimations in Fake Review Detection: Controlling for Product-Ownership and Data- Origin (Chapter 4)

C1 Table. Amazon review replacements

Matching	Replacements		
	Brand	Random	Total
Non-owners	123	35	158
Owners (truthful)	516	17	533
Owners (deceptive)	421	75	496
Total	1060	127	1187

C2 Table. All features used in the classification experiments

Part of speech							
POS_CC	POS_FW	POS_JJR	POS_NN	POS_PRP\$	POS_RP	POS_VBN	POS_WDT
POS_CD	POS_IN	POS_JJS	POS_NNP	POS_RB	POS_VB	POS_VBP	
POS_DT	POS_JJ	POS_MD	POS_NNS	POS_RBR	POS_VBD	POS_VBZ	
LIWC							
LIWC_Analytic	LIWC_adverb	LIWC_male	LIWC_achieve	LIWC_swear			
LIWC_Clout	LIWC_conj	LIWC_cogproc	LIWC_power	LIWC_netspeak			
LIWC_Authentic	LIWC_negate	LIWC_insight	LIWC_reward	LIWC_assent			
LIWC_Tone	LIWC_verb	LIWC_cause	LIWC_risk	LIWC_nonflu			
LIWC_WPS	LIWC_adj	LIWC_discrep	LIWC_focuspast	LIWC_AllPunc			
LIWC_Sixltr	LIWC_compare	LIWC_tentat	LIWC_focuspresent	LIWC_Period			
LIWC_Dic	LIWC_interrog	LIWC_certain	LIWC_focusfuture	LIWC_Comma			
LIWC_function	LIWC_number	LIWC_differ	LIWC_relativ	LIWC_Colon			
LIWC_pronoun	LIWC_quant	LIWC_percept	LIWC_motion	LIWC_SemiC			
LIWC_ppron	LIWC_affect	LIWC_see	LIWC_space	LIWC_QMark			
LIWC_i	LIWC_posemo	LIWC_hear	LIWC_time	LIWC_Exclam			
LIWC_you	LIWC_negemo	LIWC_feel	LIWC_work	LIWC_Dash			
LIWC_shehe	LIWC_anx	LIWC_bio	LIWC_leisure	LIWC_Quote			
LIWC_they	LIWC_anger	LIWC_body	LIWC_home	LIWC_Apostro			
LIWC_ipron	LIWC_sad	LIWC_health	LIWC_money	LIWC_Parenth			
LIWC_article	LIWC_social	LIWC_ingest	LIWC_relig	LIWC_OtherP			
LIWC_prep	LIWC_family	LIWC_drives	LIWC_death				
LIWC_auxverb	LIWC_friend	LIWC_affiliation	LIWC_informal				
Bigrams							
BI_also_batteri	BI_everi_day	BI_like_phone	BI_phone_much	BI_samsung_galaxi			
BI_android_phone	BI_front_camera	BI_look_good	BI_phone_price	BI_sd_card			
BI_batteri_life	BI_good_phone	BI_much_better	BI_phone_realli	BI_smart_phone			
BI_bought_phone	BI_good_price	BI_new_phone	BI_phone_screen	BI_sound_qualiti			
BI_camera_good	BI_great_phone	BI_oper_system	BI_phone_work	BI_take_photo			
BI_camera_phone	BI_intern_storag	BI_phone_batteri	BI_phone_would	BI_use_phone			
BI_camera_qualiti	BI_iphon_7	BI_phone_camera	BI_pretti_good	BI_want_phone			
BI_cheap_phone	BI_ive_ever	BI_phone_ever	BI_qualiti_phone	BI_work_well			
BI_dont_know	BI_ive_phone	BI_phone_everyth	BI_realli_good	BI_would_recommend			

BI_dont_want BI_even_though	BI_last_day BI_last_long	BI_phone_good BI_phone_ive	BI_recommend_phone	
--------------------------------	-----------------------------	-------------------------------	--------------------	--

Unigrams

UNI_1	UNI_buy	UNI_especi	UNI_high	UNI_mention	UNI_product	UNI_spec
UNI_10	UNI_call	UNI_etc	UNI_hit	UNI_might	UNI_purchas	UNI_specif
UNI_2	UNI_came	UNI_even	UNI_home	UNI_mine	UNI_put	UNI_speed
UNI_2020	UNI_camera	UNI_ever	UNI_honestli	UNI_mobil	UNI_qualiti	UNI_spend
UNI_3	UNI_cant	UNI_everi	UNI_hour	UNI_model	UNI_quickli	UNI_star
UNI_4	UNI_card	UNI_everyday	UNI_howev	UNI_money	UNI_quit	UNI_start
UNI_5	UNI_case	UNI_everyth	UNI_huawei	UNI_month	UNI_ram	UNI_still
UNI_6	UNI_caus	UNI_excel	UNI_huge	UNI_much	UNI_rang	UNI_storag
UNI_7	UNI_cellphon	UNI_except	UNI_id	UNI_multipl	UNI_rate	UNI_store
UNI_8	UNI_chang	UNI_expect	UNI_im	UNI_music	UNI_rather	UNI_super
UNI_abl	UNI_charg	UNI_expens	UNI_imag	UNI_must	UNI_read	UNI_support
UNI_absolut	UNI_cheap	UNI_experi	UNI_import	UNI_need	UNI_real	UNI_suppos
UNI_access	UNI_cheaper	UNI_extra	UNI_impress	UNI_never	UNI_realli	UNI_sure
UNI_actual	UNI_choic	UNI_extrem	UNI_includ	UNI_new	UNI_reason	UNI_system
UNI_addit	UNI_clear	UNI_fact	UNI_instal	UNI_next	UNI_recent	UNI_take
UNI_afford	UNI_color	UNI_fail	UNI_intern	UNI_nice	UNI_recommend	UNI_thank
UNI_ago	UNI_come	UNI_fall	UNI_internet	UNI_normal	UNI_record	UNI_that
UNI_allow	UNI_compani	UNI_far	UNI_iphon	UNI_noth	UNI_releas	UNI_thing
UNI_almost	UNI_compar	UNI_fast	UNI_jsnt	UNI_notic	UNI_reliabl	UNI_think
UNI_alreadi	UNI_complet	UNI_faster	UNI_issu	UNI_offer	UNI_resist	UNI_though
UNI_also	UNI_condit	UNI_featur	UNI_ive	UNI_ofen	UNI_resolut	UNI_time
UNI_although	UNI_connect	UNI_feel	UNI_job	UNI_ok	UNI_respons	UNI_took
UNI_alway	UNI_consid	UNI_find	UNI_keep	UNI_okay	UNI_review	UNI_top
UNI_amaz	UNI_cool	UNI_fine	UNI_know	UNI_old	UNI_right	UNI_total
UNI_amount	UNI_cost	UNI_fingerprint	UNI_jack	UNI_one	UNI_run	UNI_touch
UNI_android	UNI_could	UNI_first	UNI_lag	UNI_open	UNI_samsung	UNI_tri
UNI_annoy	UNI_couldnt	UNI_fit	UNI_larg	UNI_oper	UNI_satisfi	UNI_turn
UNI_anoth	UNI_coupl	UNI_flagship	UNI_last	UNI_option	UNI_say	UNI_two
UNI_anyon	UNI_cover	UNI_found	UNI_least	UNI_os	UNI_scratch	UNI_updat
UNI_anyth	UNI_crash	UNI_freez	UNI_less	UNI_other	UNI_screen	UNI_upgrad
UNI_app	UNI_current	UNI_friend	UNI_let	UNI_overal	UNI_sd	UNI_usag
UNI_appl	UNI_custom	UNI_front	UNI_lg	UNI_own	UNI_second	UNI_use
UNI_applc	UNI_daili	UNI_full	UNI_life	UNI_pay	UNI_see	UNI_user
UNI_arent	UNI_damag	UNI_function	UNI_light	UNI_peopl	UNI_seem	UNI_valu
UNI_around	UNI_data	UNI_galaxi	UNI_like	UNI_perform	UNI_servic	UNI_version
UNI_ask	UNI_day	UNI_game	UNI_line	UNI_person	UNI_set	UNI_video
UNI_avail	UNI_deal	UNI_gave	UNI_littl	UNI_phone	UNI_sever	UNI_want
UNI_away	UNI_decent	UNI_gb	UNI_live	UNI_photo	UNI_short	UNI_wasnt
UNI_back	UNI_decid	UNI_gener	UNI_load	UNI_pick	UNI_show	UNI_watch
UNI_bad	UNI_definit	UNI_get	UNI_long	UNI_pictur	UNI_simpl	UNI_way
UNI_basic	UNI_design	UNI_give	UNI_longer	UNI_piec	UNI_simpli	UNI_week
UNI_bateri	UNI_devic	UNI_given	UNI_look	UNI_pixel	UNI_sinc	UNI_well
UNI_battery	UNI_didnt	UNI_go	UNI_lose	UNI_play	UNI_singl	UNI_whole
UNI_beauti	UNI_differ	UNI_goe	UNI_lot	UNI_plu	UNI_size	UNI_within
UNI_believ	UNI_disappoint	UNI_good	UNI_love	UNI_pocket	UNI_slightli	UNI_without
UNI_besid	UNI_display	UNI_googl	UNI_low	UNI_point	UNI_slow	UNI_wont
UNI_best	UNI_doesnt	UNI_got	UNI_lower	UNI_poor	UNI_small	UNI_work
UNI_better	UNI_dont	UNI_great	UNI_made	UNI_power	UNI_smart	UNI_worth
UNI_big	UNI_download	UNI_half	UNI_main	UNI_present	UNI_smartphon	UNI_would
UNI_bigger	UNI_drop	UNI_hand	UNI_make	UNI_pretti	UNI_softwar	UNI_wouldnt
UNI_bit	UNI_due	UNI_handl	UNI_mani	UNI_previou	UNI_someon	UNI_wrong
UNI_bought	UNI_easi	UNI_happi	UNI_market	UNI_price	UNI_someth	UNI_x
UNI_brand	UNI_easili	UNI_hard	UNI_may	UNI_pro	UNI_sometim	UNI_xiaomi
UNI_bright	UNI_els	UNI_hardwar	UNI_mayb	UNI_probabl	UNI_soon	UNI_year
UNI_bring	UNI_end	UNI_heavi	UNI_mean	UNI_problem	UNI_sound	UNI_your
UNI_broke	UNI_enjoy	UNI_help	UNI_memori	UNI_processor	UNI_space	
UNI_budget	UNI_enough					

C3 Table. Extra Trees classifier settings

n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, in_impurity_decrease=0.0,	bootstrap=False, oob_score=False, n_jobs=None, random_state=319, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None
--	--

C4 Table. Other tested classifiers

Classifier
Random Forest
Decision Tree
MultinomialNB
GaussianNB
GradientBoosting
Logistic
Regression
LinearSVC

C5 Table. All classification performance metrics across all experiments

			Sentiment							
			Positive				Negative			
	Analysis	Testing	Acc.	Precision	Recall	F1	Acc.	Precision	Recall	F1
Pure	1	Veracity	60.26	60.54	60.23	60.04	69.87	70.24	69.87	69.76
	2	Ownership	63.41	63.87	63.41	63.07	58.61	58.65	58.62	58.54
	3	Data-origin	88.33	88.79	88.35	88.23	85.23	85.79	85.26	85.18
Confounded	4	Veracity, Ownership	66.19	66.43	66.17	66.06	74.17	74.41	74.17	74.09
	5	Veracity, Data-origin	86.94	87.35	86.94	86.91	84.44	84.62	84.47	84.43
	6	Veracity, Ownership, Data-origin	88.12	88.34	88.12	88.10	87.26	87.78	87.24	87.19

Appendix D: Counterfeits on Cryptomarkets: A measurement between Jan-2014 and Sep-2015 (Chapter 5)

D1 Table. List of considered markets

Cell properties: **Yellow**: Not in the archive; **Orange**: timeframe too short or data gaps too big; **Blue**: Products are not or only partly categorized; **Grey**: Data gaps too big or no counterfeits were found within the categories; **Green**: market included in analyses

Name	Archived	Data-start	Data-end	Data Categorized?	Notes
Abraxas	Yes	16.12.2014	05.07.2015	Yes	
Acropolis	No	-	-	-	
Agora	Yes	01.01.2014	07.07.2015	Yes	
Alpaca	Yes	24.04.2014	07.11.2014	Partly	
AlphaBay	Yes	21.12.2014	28.01.2017	Yes	
Anarchia	Yes	07.05.2015	05.07.2015	Partly	
Andromeda	Yes	12.04.2014	18.11.2014	Yes	Strong variation on captured listings per scrape
Apple Market	No	-	-	-	
Area51	Yes	22.06.2014	23.01.2015	No	
Black Market	No	-	-	-	
BlackBank Market	Yes	06.02.2014	17.05.2015	Yes	
Blue sky	yes	06.01.2014	28.09.2014	Partly	Several months missing between scrapes
Cloud 9	Yes	11.02.2014	01.11.2014	Yes	
Crypto Market	Yes	19.02.2015	06.07.2015	Yes	Strong variation on captured listings per scrape
Darknet Heroes League	Yes	30.05.2015	04.07.2015	Partly	
Diabolus/SR3	Yes	17.10.2014	05.07.2015	Yes	
Dream Market	Yes	09.01.2014	05.07.2015	Partly	
East India Company	Yes	28.04.2015	05.07.2015	Yes	
Evolution	Yes	21.01.2014	17.03.2015	Yes	
Hansa	No	-	-	-	
House of Lions Market	No	-	-	-	
Hydra	Yes	03.04.2014	27.10.2014	Partly	Counterfeits are not captured in categories
Middle Earth Marketplace	Yes	23.06.2014	05.07.2015	Yes	
Mr Nice Guy 2	Yes	21.02.2015	04.07.2015	-	Cannot inspect data (corrupted)
Nucleus Marketplace	Yes	24.10.2014	07.07.2015	Partly	Most of the scrapes were blocked (no content)
Outlaw Market	Yes	09.01.2014	05.07.2015	Partly	Almost exclusively drugs; newer listings without categorization
Pandora	Yes	25.12.2013	05.11.2014	No	
Pirate market	Yes	25.01.2014	21.09.2014	Yes	Big gaps in the scraped data
RoadSilk	Yes	26.12.2013	15.01.2014	Yes	
Silk Road 2.0	Yes	13.12.2013	06.11.2014	Partly	
Silk Road 3.0	No (see Diabolus)	-	-	-	

Silk Road Reloaded	Yes	18.01.2015	05.07.2015	Yes	No counterfeits found within categories
The Marketplace	Yes	03.01.2014	09.11.2014	Yes	
TheRealDeal	Yes	16.04.2015	05.07.2015	Partly	
Tochka	Yes	05.02.2015	04.07.2015	Partly	
Tom	Yes	05.05.2014	17.12.2014	Partly	2 months of data missing
Topix	No	-	-	-	
Tor Bazaar	Yes	02.02.2014	06.11.2014	Partly	Almost exclusively drugs; Counterfeits not captured in categories
Valhalla	No	-	-	-	

D2 List. Categories included for keyword searches:

“Other”, “Fraud”, “Electronics”, “CustomListings”, “Miscellaneous”, “Accessories”, “Fraud Related”, “Jewelry”, “Weight loss”, “Forgeries”, “Jewelry”, “Listings”, “Tobacco”, “Market”, “Hidden”, “Precious metals”, “Jewels & Gold”, “Other Listings”, “Abraxas”, “Electronics”, “Watches”, “Accessories”, “Clothing”, “Sunglasses”, “Cigarettes”, “Jewelry”, “Collectables”, “Tobacco”, “Metals Stones”.

D3 List. Synonyms used for keyword search:

"copies", "copy", "counterfeit", "replica", "fake", "clone", "deceit", "deception", "bum", "dummy", "facsimile", "gyp", "hoax", "humbug", "imitation", "imposture", "phony", "pseudo", "put-on", "reproduction", "sham", "simulacrum", "bogus", "junque", "likeness", "miniature", "lookalike", "xerox", "forge", "ditto", "dupe", "mimeo", "reduplication", "replication", "repro", "stat", "forgery", "forged", "spurious", "mock", "false", "unreal", "ungenuine", "falsified".

D4 List. Synonyms of authentic:

"genuine", "authentic", "real", "valid", "original", "actual", "official".

D5 List. Keywords used to exclude listings:

"tutorial", "template", "licence", "ID", "refund", "scan", "scans", "COUPONS", "Licence", "License".

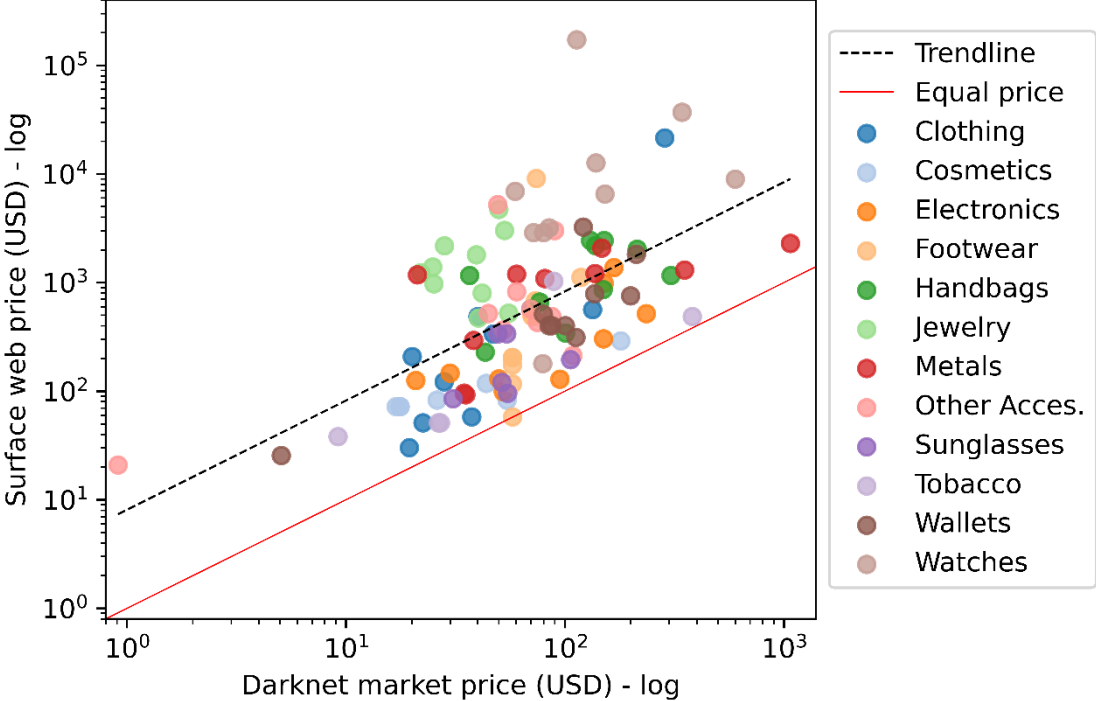
D6 Table. Full lists of percentage counterfeits by OECD/EUIPO

OECD/EUIPO		
Category	Share of custom seizures (%)	Share of seized value (%)
Footwear	22.6	10.45
Clothing, knitted or crocheted	17.00	8.20
Articles of leather	13.55	11.60
Electrical machinery and equipment	12.25	10.75
Watches	5.70	22.75
Optical, photographic and medical instruments	5.15	4.10
Perfumery and cosmetics	3.50	4.95
Toys	2.75	4.65
Jewellery	1.85	5.85
Machinery and mechanical appliances	1.55	0.95
Pharmaceutical products	1.55	1.50
Headgear	1.45	0.30
Other made up textile articles	1.15	1.10
Vehicles	1.10	1.50
Clothing, not knitted or crocheted	1.00	0.95
Plastics and articles thereof	0.65	0.60
Furniture, bedding, cushions, lamps etc.	0.49	0.75
Miscellaneous manufactured articles	0.40	0.70
Foodstuff	0.40	0.56

D7 Table. Full list of percentage counterfeits by IPO

IPO	
Category	%
Optical Media	39.85
Tobacco	28.15
Clothing	7.94
Alcohol	4.42
Footwear	2.77
Circumvention	2.01
Cosmetics	1.10
Handbags	1.44
Software	0.68
E-Games	0.66
File-sharing	0.78
Other	7.41
Watches	1.19
Headphones	0.44
Electrical	0.36
Jewellery	0.51
Pharmaceuticals	0.30

D8 Figure. Product price differences for 10 products in each category between Cryptomarkets and the surface web.



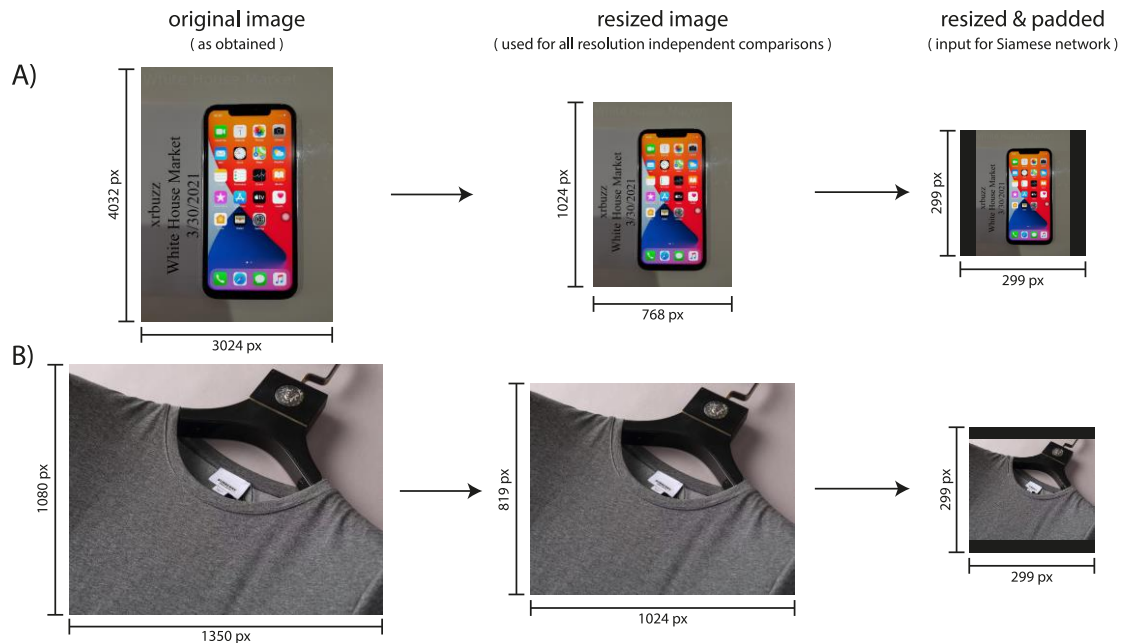
Appendix E: From the anonymity network to the surface web: Scouting eBay for Counterfeits (Chapter 6)

E1 Table. Descriptive statistics of word occurrences

Listing texts	min		max		median		mean		SD	
	CM	eBay	CM	eBay	CM	eBay	CM	eBay	CM	eBay
Title	2	1	25	22	5	12	5.6	11.19	2.92	2.62
Item specifics	-	0	-	690	-	97	-	108.23	-	52.97
Description 1	4	0	891	182	70	0	95.77	12.67	111.95	22.29
Description 2	-	0	-	65060	-	191	-	533.97	-	2762.94

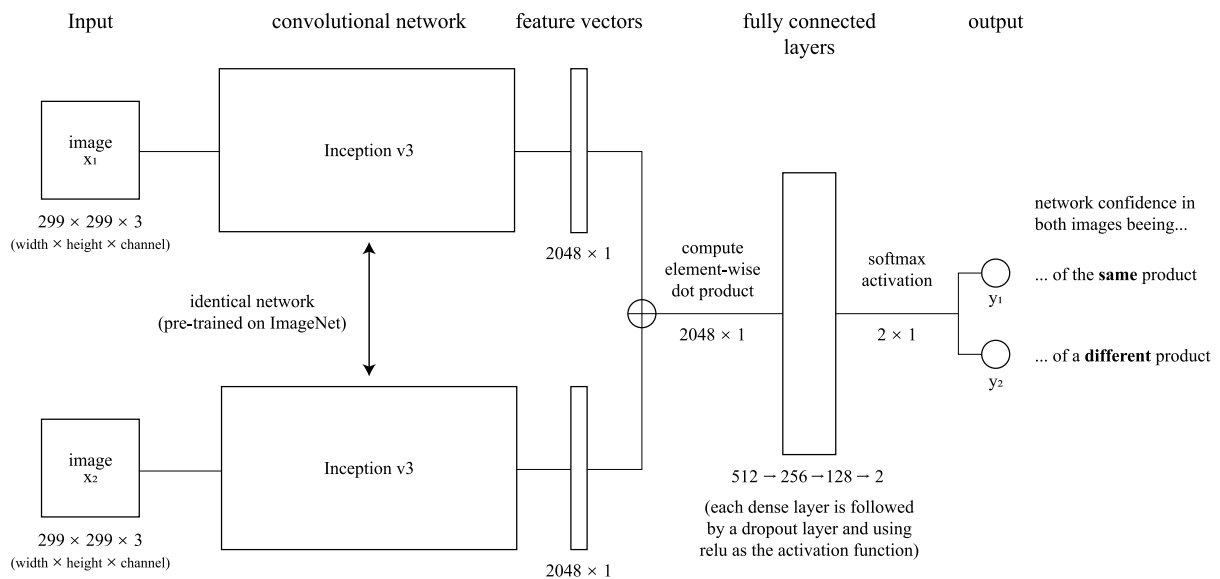
Word distributions of cryptomarket (CM) and eBay listings texts, which are used in our analyses to generate text similarity scores between CM and eBay listings. Item specifics contain factual properties of the product (e.g., product height, weight, brand). eBay listings have two descriptions corresponding to a short (1) and long (2) text in most cases. Due to a change in the eBay webpage, most item specifics were not collected in the second scrape period. Thus, the statistical descriptors for item specifics only refer to the first eBay scrape period. 2,108 eBay listings were fully or partially not in English and were automatically translated with the Google translator in Python, using the package “deep translator” (Baccouri, 2020/2020).

E2 Figure. Example of rescaled cryptomarket images



Example of rescaling procedure of cryptomarket (CM) images, restricting the largest dimension (width, height) to a maximum of 1024 px while preserving the original aspect ratio. Smaller images are not rescaled. All images (from both eBay and CM listings) are rescaled to 299 px x 299 px and padded, if necessary, to fit the input dimensions of the Inception V3 feature extractor backbone of the Siamese network. **A)** rescaled image from listing “SILVER iPhone 12 Pro Max 512GB - Sealed In Box”. **B)** rescaled image from listing “Burberry mercerized cotton t-shirt 77005”.

E3 Figure. Siamese network architecture



Depiction of the flow of information within the Siamese network architecture. The values below each element indicate its dimensions. The inputs are made up of two images to be compared: x_1 , an image from a cryptomarket listing to x_2 , an image from a surface web listing. We use an Inception V3 network, pre-trained on the ImageNet dataset, to produce high-level feature vectors from both input images. The element-wise dot product of the two produced feature vectors is computed, and the resulting vector is passed to a set of fully connected layers, which produce the final output. The network forms two predictions, y_1 and y_2 , the similarity and dissimilarity, respectively, of the two images and can therefore be trained as a binary classifier.

E4 Detailed description of sampling annotation data

To determine a single similarity score for each product pair, we first created a single score describing the similarity of the title, description, and images by merging the different similarity metrics. We aimed to ensure equal weightings for the three aspects by obtaining a single score for each listing attribute (i.e., title, description, images). For each listing attribute, we determined if the score for a metric (e.g., colour histogram) given to a comparison pair was unusually high (or low, depending on the meaning of the score) by calculating if the score value was above (or below) two standard deviations from the mean of all scores for that metric. To avoid possible product-specific scoring biases, we calculated the means and standard deviations for each product group (watches, shoes, etc.) separately and determined whether each product's similarity value deviated unusually from its product group. The merged similarity score for any given product pair attribute consists of the count of unusual deviations ($2 \cdot \text{STD} \pm \text{Mean}$) of each metric. We then normalized each merged score, resulting in three scores for the title, description, and images, each ranging from 0 (low similarity) to 1 (high similarity). By standardizing the scores, we counteract the unequal number of metrics between text and image scores. eBay listings contain several subsections of detailed

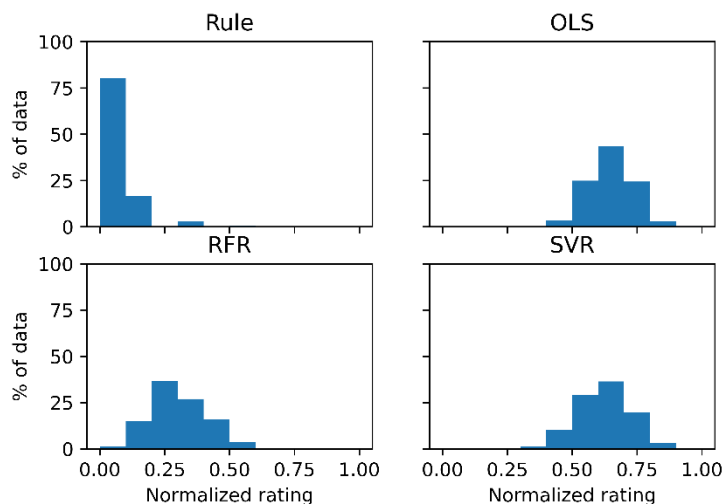
information (e.g., product specifics), which were all merged into one text labelled as product description.

We scored all cryptomarket and eBay (first scrape only) product pairs and ranked them from the highest to lowest similarity scores. To determine if the ranking could capture a preliminary order of similarity, we inspected the top and last 50 ranks of product pairs. Within the top 50 ranks, we found that products were correctly matched on product types, including four products that seemed to depict and describe the same product (2 shoes, a shirt, and a bag chain). Within the last 50 ranks, products seem to match very poorly, including mismatches of product types (e.g., a shoe with a shirt). Since the ranked products appeared to follow a somewhat sensible order, we sampled the top and last 250 products and a random sample of 500 from the remaining products, resulting in a sample of 1000 product pairs.

E5 Table. Regression models settings

Model	Settings
OLS	<i>fit_intercept=True, normalize='deprecated', copy_X=True, n_jobs=None, positive=False</i>
SVR	<i>kernel='linear', degree=3, gamma='scale', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=-1</i>
RFR	<i>n_estimators=100, criterion='squared_error', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=319, verbose=0, warm_start=False, ccp_alpha=0.0, max_samples=None</i>

E6 Figure. Detailed model comparison










Distribution of normalized similarity scores for each scoring method. We re-trained each model on the full set of training data (without folds) and used them to make similarity rating predictions on the unannotated data to compare their predictions to the rule-based scoring system that was used to sample our annotations. To compare the rule-based approach and the other scoring methods, we normalized the scores from 0 (not at all similar) to 1 (completely similar) (Fig. F). The rule-based system differs greatly from the regression models by generating mostly low similarity scores, while the RFR model generated lower to medium scores, and the OLS and SVR models behaved similarly by generating medium to higher scores.

E7 Table. Full product descriptions (eBay scrape period 1)

Fig	Description
4: CM-1	<p>Nike Air Jordan 1 Retro High OG 555088-134 Size: 36 36.5 37 38 38.5 39 40 40.5 41 42 42.5 43 44 44.5 45 46</p> <p>The popularity of Air Jordan 1 this year has remained constant, but Jordan Brand has never stopped developing its new color scheme. The Jordan 1 High OG (University Blue) will be one of the first versions of the Jordan brand in early 2021. This Air Jordan 1 is matched with white, college blue and black colors. Although the images have not yet been leaked, they are expected to have white leather on their uppers, while college blue is on the overlay. Other details will include black trim, white midsole and blue rubber outsole.</p>
4: eBay-1	<p>Item specifics Condition: New with box: A brand-new, unused, and unworn item (including handmade items) in the original packaging (such as the original box or bag) and/or with the original tags attached. See all condition definitions - opens in a new window or tab ... Read more about the condition Pattern: Colorblock Model: Air Jordan 1 Style Code: 555088-134 Style: Sneaker Character: Michael Theme: University US Shoe Size: 9.5 Features: Comfort, Cushioned Color: Blue Country/Region of Manufacture: China Silhouette: Jordan 1 Retro High OG University Blue Upper Material: Leather Brand: Jordan Idset_Mpn: 555088-134 Department: Men Performance/Activity: Basketball Type: Athletic Product Line: Jordan</p> <p>Nike Air Jordan 1 Retro OG High White University Blue 555088-134 Men's Size 9.5. Condition is "New with box". Shipped with FedEx Ground or FedEx Home Delivery.</p>

<p>4: CM-2</p>	<p>APPLE SETUP WHEN YOU FIRST RUN IT: APP STORE IS FUNCTIONAL: ITUNES IS FUNCTIONAL: FLASHED TO MAKE ALL DETAIL LOOK LIKE APPLE: SECRET MENU: ABILITY TO CHAMGE YOUR IMEI: ABILITY TO CHANGE YOUR MEID: ABILITY TO CHABGE YOUR SERIAL NUMBER: PRE-ROOTED: ABILITY TO CLONE AND RECIEVE MESSAGES AS ANYONE IN THE WORLD: ABILITY TO BURN PHONE ON A MOMENTS NOTICE: ABILITY TO MAKE YOU RICH: GREAT FOR CARDING: GREAT FOR ANONYMITY: IN YOUR CART CAUSE YOU ARE BRILLIANT: GREAT CUSTOMER SERVICE WITH AN ACE OF SPADES UP HER SLEEVE JUST IN CASE YOU CAN'T PULL OFF A SALE IN 30 DAYS: DEDICATED SUPPORT TEAM:</p> <p>Put this in your cart, let's make some money</p> <p>Just so you know from jump I have a video of an unboxing that we took in house. Taken by a Professional!!!.....pothead but this dude ain't bad looks....like it was done by someone in at least...3rd grade? I have absolutely no problem taking one for you as well if needed. Ask me to perform certain tasks on it so you can see if it is right for you.</p> <p>Unboxing: You would never in your life know from opening this thing that it is a clone. Ive opened them side by side. No difference.</p> <p>Turn it on: Same startup screen , same layout, same apps. These are the upgraded versions too. So no lag whatsoever, its smooth as ever.</p> <p>Take a picture: Fantastic quality</p> <p>Security: Face Id Fingerprint unlock</p> <p>Insert SIM: IT WORKS! same imei numbers as actual iPhone. Most are valid. You will not get a dud from us with these 12's. They are fresh and we will send a picture of the imei for each buyer to test before shipment to see with their own eyes</p> <p>It's all the same....but the rice....waaaaaay less</p> <p>This shit IS an iPhone..just one that was dropped on its head as a baby and is hooked on droids.... Thats right it is run with android. but ill tell you this - quick handoffs ---hahaha that's what we used to say, FUCK THAT, grab a beer with the dude, go to grandmas for your special holiday and let em pass it around to the whole family to test before purchase..</p> <p>You will receive a nice little package and ill give you a little instruction booklet on how to be really good at pushing these out. honestly the last thing you will ever see of these people will be smiling faces. I'm not exaggerating in the slightest when I say the last guy I hooked up SKIPPED away from me he was so happy. Hah. shit. I just dont understand what people see in these apple products. It looks like a bigger more fucked up version of the 5. IT NEVER CHANGES. But whatever these things are FLYING off the handle at \$12-\$1800 even broken they are worth 8 or 9.</p>
--------------------	--

	<p>Shit BUY 10 break ALL of em and give it for six and youll be rolling in it. Thats THREE THOUSAND DOLLARS JUST FOR BROKEN IPHONES. you know people woll be looking for parts, a screen, a camera WHATEVER, that could be your hustle and nobody would be the wiser.</p> <p>Personally I dont have the time nor the poker face for it so this is what I do. BUT my lovely staff here have each had to get rid of our base models. Part of their interview was to sell 5 of the ones worse than these so if you need ANY help trust me they know what the fuck, where the fuck, how the fuck to pull it off.</p> <p>I'll be putting up Note20s and Note9s soon. Those are beautiful little creatures as well.</p> <p>I refuse to sell the Note10's cause someone fucked up down the line didnt make the screen edge to edge and decided it was acceptable. We only want to give you guys the best.</p> <p>If you need more than one, the price breaks down nicely. Just shoot us a message and we will throw up a personalized listing.</p> <p>Oh and the package includes:</p> <ul style="list-style-type: none"> 1x 512GB iPhone 12 Pro Max 1x iPhone case 1x iPhone screen Protector 1x Lightening to USB-C Authentic Apple Charging Cord 1x Apple welcome packet 1x Sim Card Pin 1x iPhone OEM Box <p>Enjoy</p>
<p>4: eBay-2</p>	<p>Item specifics Condition: Open box : An item in excellent, new condition with no wear. The item may be missing the original packaging or protective wrapping, or may be in the original packaging but not sealed. The item includes original accessories. The item may be a factory second. See the seller's listing for full details and description. See all condition definitions - opens in a new window or tab Seller Notes: "Open Box Condition: Brand new, but box has been opened." Brand: Apple Connectivity: 5G, Lightning, Bluetooth, NFC, Wi-Fi Model: Apple iPhone 12 Pro Max Processor: Hexa Core Style: Bar Operating System: iOS Storage Capacity: 512 GB Manufacturer Color: Graphite Features: Proximity Sensor, Barometer, LiDAR Scanner, Accelerometer, Fingerprint Sensor, E-compass, Ambient Light Sensor, Gyro Sensor, eSIM Camera Resolution: 12.0 MP Color: Gray MPN: MG9K3LL/A Network: Verizon Screen Size: 6.7 in UPC: 0194252019887 EAN: 0194252019887</p> <p>Search our eBay store  iPhone 12 Pro Max - Verizon - 512GB - Graphite - Open Box Device Details  Network: Verizon - Can only be used with Verizon. Is not unlocked for use with other networks. Brand: Apple Model: iPhone 12 Pro Max Storage: 512GB Color: Graphite What's In The Box  iPhone 12 Pro Max Box & All Manufacturers Sealed Accessories Condition    Certified - Open Box: Brand new, but box has been opened. What is a certified device? Before we clear a device for resale it must pass a series of functional, database and cosmetic inspections. These inspections are performed by our expertly trained team of in-house technicians using state of the art software to ensure there are no testing errors. 60 Day - Free Returns If you aren't completely satisfied with your purchase you can return it for a refund or exchange. We'll even cover the return shipping. 1 Year warranty  This device comes with a 12 Month (365 day) warranty starting at the date of purchase. If your device stops working properly from normal use during this timeframe you can ship it back to us and we'll either repair your device, send an exchange, or send a refund. Additional Details This warranty does not cover accidental damage of any kind. This warranty is void if any repairs or modifications have been performed or attempted to device hardware or software. The remedy for your warranty claim will be decided at our sole discretion. Generally speaking, we'll attempt to repair your device or send an exchange before considering a refund. This warranty only extends to the original purchaser; it cannot be transferred if the device is sold / given to a 3rd party. Why Buy From Us? Based in Sunny Florida, [Anonymized name] has been helping people all over the USA buy and sell gadgets since 2016. If you do a quick search for our company you'll see that we have great reviews not only on eBay, but all over the web. Our customers love us for one simple reason - we deliver on promises. Our Promises Quality Devices - We don't sell junk. Our certified devices have been tested extensively to make sure they'll work perfectly right out of the box. Fast Shipping - All orders are shipped the same or next business day after purchase. Orders placed before 2PM EST Tuesday - Friday will be shipped the same day. Amazing Customer Service - Our philosophy on customer service isn't a new one, we treat people how we like to be treated. That means fast & accurate communication, no-hassle 60 day free returns and a 1 year warranty.</p>

4: CM-3	Brand---Rolex quality level---UltimateAAA+ Manufacturer:N Window material:Sapphire glass Bezel material:ceramics Case material:stainless steel Strap material:stainless steel Case Diameter:44 MM waterproof:60 M Movement:ETA 2836 Automatic mechanical movement. Vibration frequency:28800 function:Date.hour.minute.second. Dial luminous:Yes ---more images--- [Anonymized link] ---box images--- [Anonymized link] If you want to buy a box. the certificate please choose 100USD in the shipping option. Because the box is large.the shipping is expensive. When you buy a watch. please fill in the address information in this format: name: address: City/State: Postcode: country: 1. When you order a watch, I expect to send the order within 4-8 days. I need time to order the watch and check the quality. When I send an order, I will provide tracking number information. 2. The United States, Canada, and Australia usually deliver in 10-15 days. EU, UK usually deliver in 10-20 days. 3. I guarantee that the goods can be delivered successfully. If the goods are lost or detained by the customs, I will bear the loss.
4: eBay-3	<p>Item specifics Condition: Pre-owned: An item that has been used previously. The item may have some signs of cosmetic wear, but is fully operational and functions as intended. This item may be a floor model or store return that has been used. See the seller’s listing for full details and description of any imperfections. See all condition definitions - opens in a new window or tab ... Read more about the condition Water Resistance: 3900 m (390 ATM) Water Resistance Rating: Diver \’s 300 m (ISO 6425) Model: Sea-Dweller Band Material: Stainless Steel Country/Region of Manufacture: Switzerland Type: Wristwatch Watch Shape: Round Features: Date Department: Men Customized: No Style: Casual, Sport, Diver, Luxury With Papers: No Case Color: Silver Year Manufactured: 2009 Caseback: Solid Indices: Non-Numeric Hour Marks, Round Indexes MPN: 116660 With Original Box/Packaging: No Hour Markers: Dot, index Dial Color: Black Case Material: Stainless Steel Band Color: Silver Gender: Men Reference Number: 116660 Buckle Type: Folding buckle Display: Analog Box/Papers: Box and Papers Brand: Rolex Movement: Mechanical(Automatic) Case Size: 44 mm Warranty: 2-Year Watchbox UPC: Does not apply</p> <p>[Anonymous name] Jewelry and Loan [Anonymous name] [anonymous rating] Sign up for newsletter Search within store Visit Store: [Anonymous name] Jewelry and Loan Items On Sale Categories Other</p> <p>Rolex Deepsea Reference 116660 Stainless steel 44mm Beast. This is a great watch in excellent condition with no box or card. Ready for your wrist with speedy delivery. We Guarantee Authenticity on all Merchandise. We Are A Long Standing Member of the [Anonymous name] [Anonymous name] in operation since [Anonymous date]. Two locations [Anonymous location] and [Anonymous location] WE ONLY ACCEPT PAY-PAL. AND CREDIT CARDS THROUGH PAYPAL. PLEASE SERIOUS inquiries ONLY FOR ANY FURTHER ASSISTANCE WITH PURCHASED ITEMS PLEASE CONTACT Attn. Zef WITH PROPER RETURN INSTRUCTIONS AND ADDRESS (Buyer is responsible for return shipping) We try to ship within 1 to 3 business days from the payment. Buyer pays a fixed rate for shipping and handling for the United States and International sales Payments must be made in 2 days of purchase. REMINDER : WE DO NOT SHIP TO “FPO/APO, OR” PO BOX “LOCAL PICK UP IS ACCEPTED AND REQUIRES THE CUSTOMER TO PAY 6% SALES TAX OF TOTAL AMOUNT AT THE TIME OF PICK UP” ONLINE ITEMS ARE NOT ELIGIBLE FOR [Anonymous name] UPGRADE AND TRADE IN POLICY Thank You</p>

E8 Table. Full product descriptions (eBay scrape period 2)

Figure	Description
5: CM-1	<p>Nike Air Jordan 1 Retro High OG 555088-134 Size: 36 36.5 37 38 38.5 39 40 40.5 41 42 42.5 43 44 44.5 45 46</p> <p>The popularity of Air Jordan 1 this year has remained constant, but Jordan Brand has never stopped developing its new color scheme. The Jordan 1 High OG (University Blue) will be one of the first versions of the Jordan brand in early 2021. This Air Jordan 1 is matched with white, college blue and black colors. Although the images have not yet been leaked, they are expected to have white leather on their uppers, while college blue is on the overlay. Other details will include black trim, white midsole and blue rubber outsole.</p>
5: eBay-1	Nike Air Jordan 1 Retro High OG GS University Blue 575441-134 Size 7Y.
5: CM-2	<p>Louis Vuitton Bag charm Chain Fleur de Monogram M65111</p> <p>CONDITION IS NEW</p> <p>COMES WITH:</p>

	<p>1x LV CHARM 1x LV DUST BAG 1x CERTIFICATE OF AUTHENTICITY 1x CARE BOOKLET 1x LV BOX 1x LV SHOPPING BAG</p>
5: eBay-2	<p>Louis Vuitton Bag Charm/Key Chain Fleur de Monogram Gold Plated. This is preowned in excellent condition with no signs of use or wear. Comes with box. Buy in US to avoid duty and taxes that I have already paid!! Please message for more info or photos. TY!</p>
5: CM-3	<p>Brand---Audemars Piguet quality level---UltimateAAA+ Manufacturer:JF Window material:Sapphire glass Bezel material:stainless steel Case material:stainless steel Strap material:stainless steel Case Diameter:41MM waterproof:30 M Movement:Clone Automatic mechanical movement. Vibration frequency:28800 function:calendar.Date.hour.minute.second. Dial luminous:Yes ---more images--- [anonymous link] images--- [anonymous link] If you want to buy a box. the certificate please choose 160USD in the shipping option. Because the box is large.the shipping is expensive. When you buy a watch. please fill in the address information in this format: name: address: City/State: Postcode: country: 1. When you order a watch, I expect to send the order within 4-8 days. I need time to order the watch and check the quality. When I send an order, I will provide tracking number information. 2. The United States, Canada, and Australia usually deliver in 10-15 days. EU, UK usually deliver in 10-20 days. 3. I guarantee that the goods can be delivered successfully. If the goods are lost or detained by the customs, I will bear the loss.</p>
5: eBay-3	<p>Audemars Piguet Audemars Piguet Royal Oak Perpetual Calendar Watch 26574OR.OO.1220OR.02 Details Department Unisex Adult Dial Color Blue Dial Pattern Grande Tapisserie Case Size 41 mm Customized No Case Material Pink Gold Seller Warranty Yes Warranty 5 Year Warranty Reference Number [anonymous number] Water Resistance 20m (2 ATM) With Papers Yes Features Perpetual Calendar, Sapphire Crystal Case Color Pink Gold Item description 41 mm 18K pink gold case, 9.5 mm thick, glareproofed sapphire crystal back, screw-locked crown, glareproofed sapphire crystal, blue dial with Grande Tapisserie” pattern, pink gold applied hourmarkers and Royal Oak hands with luminescent coating, blue inner bezel, Manufacture 5134 selfwinding movement with perpetual calendar with week indication, day, date, astronomical moon, month, leap year, hours and minutes, approximately 40 hours of power reserve. Water resistant to 20 meters. Condition: Used• FREE Overnight Express & Insured FedEx Domestic & International overnight shipping. • International import duties, taxes, and charges are not included in the item price or shipping cost. These charges are the buyer’s responsibility. • Full 5-YEAR warranty If the item fails because of a manufacturer’s defect. [anonymous name] will repair or replace the item at absolutely no cost to you. Warranty coverage except: Abuse of the timepiece, accidental water intrusion caused by user, outside modifications and third-party repair attempts of any kind will void the warranty. External damage to the product not covered under warranty: damage resulting from abusive wear, crystal/glass, watch bracelet, watch bezel, straps, screws. • Please feel free to contact us directly if you have any questions. We are happy to assist with any inquiry.</p>