Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Matching Problems in Forensics

Presented by: Xiao Hui Tai

Accepted by: Department of Statistics

Readers:

_____
William F. Eddy, Advisor

_____
Nicolas Christin

_____
Brian W. Junker

_____
Joseph B. Kadane

_____
Rebecca Nugent

Approved by the Committee on Graduate Degrees:

_____
Richard Scheines, Dean Date

Carnegie Mellon University

Matching Problems in Forensics

A Dissertation Submitted to the Graduate School

in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics

by

Xiao Hui Tai

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

June 2019

Acknowledgements

I would first like to thank all the people involved in this work. Bill Eddy, thank you for teaching me not only about statistics, but also about life and countless other topics. Thank you to my committee members, Nicolas Christin, Brian Junker, Jay Kadane and Rebecca Nugent. Nicolas, thank you for welcoming a random statistician to your research group. It has been a lot of fun. Thanks also to Kyle Soska and the rest of the cybercrime group, especially Emily Rah for contributing data on criminal complaints. Jay, thank you for all your suggestions on improving this thesis document. I was sad to see the royal "we"s go. Rebecca, thank you for numerous insightful comments. Thank you to the forensics group, especially Robin Mejia for keeping things organized. Thanks also to the Data Science Initiative research group for many interesting discussions. Thank you finally to Steve Fienberg and Joel Greenhouse, for getting me started on this path.

I would also like to thank all the people who not only got me through graduate school, but made it fun. Kayla Frisoli, I could not have done this without you. Ilmun Kim, thank you for all your help, especially with beautifying this thesis and my slides. (One might describe tikz as a game changer!) Matteo Bonvini, thank you for always bringing joy and laughter. Kevin Lin, thank you for random food and for suggestions on my talk. Thanks also to Alden Green, Neil Spencer, and many other students in the department, for support, motivation and for contributing to a positive work environment. A final thanks to Commonplace Coffee where I spent countless productive hours.

To my family and friends back home, especially my sister Xiao Fang, thank you for all your support!

Abstract

Forensic evidence refers to DNA, fingerprints, bullets and cartridge cases, shoeprints and digital evidence left behind when a crime is committed. The underlying assumption is that the perpetrator of the crime, or tools they might have used, leave identifiable characteristics on the evidence that can be traced back to the source. This is the basis of forensic matching, where pairs of evidence are compared, to infer if they came from the same source. For specific pairs of comparisons, such as whether a particular cartridge case comes from a suspect's gun, an inference of a match could have probative value and be used as testimony in courts. With the exception of DNA evidence, this is done manually by trained examiners who make judgments based on their experience and training. Despite the widespread use of forensic evidence in courts, as well as the high stakes involved in criminal investigations, there has been a lack of scientific research to back up this claim of being able to reliably match evidence to source. Examiner error rates are unknown, and it is difficult to attach a quantitative value of the weight of evidence to a subjective opinion. Beginning in the 1990s, exonerations due to DNA evidence revealed problems in many forensic science disciplines, and examiners have been found to have overstated forensic results, leading at least in part to wrongful convictions. As a result, there has been a push in recent years towards automatic methods for making comparisons.

In this thesis, the goal is to develop such methods; in particular, to produce similarity scores for pairwise comparisons of evidence. Apart from addressing the issues raised, automatic methods can be used in the following other ways. They can generate 1) a ranking of similarities of pairs, which could be used to generate investigative leads, or for blind verification; 2) a match or non-match conclusion; 3) a linked or disambiguated data set; 4) random match probabilities or likelihood ratios, as measures of the weight of evidence. Now, record linkage in statistics is the process of inferring which entries in different databases correspond to the same real-world identity, in the absence of a unique identifier. Forensic matching can be thought of as an application of record linkage; simply think of records as evidence and real-world entity as the source of the sample. Steps commonly used in forensic matching can be demonstrated to correspond to steps typically used in record linkage. By thinking about forensics problems in the context of record linkage, one immediately has well-developed frameworks and tools at one's disposal.

I describe a framework that can be used to develop automatic forensic matching methods in a systematic manner. This simplifies the record linkage process and adapts it to a forensic context. I apply this to develop automatic methods for two forensic matching problems. The first is firearms identification, where cartridge cases are compared to infer if they were fired from the same gun. I develop an open source, fully automatic method to compare 2D optical images and 3D topographies, and evaluate performance on over a dozen publicly available data sets. The second problem is matching accounts on anonymous marketplaces. I use marketplace data scraped over eight years, and generate a set of features that is costly for an adversary to mimic. Through these examples, I demonstrate how forensic matching problems can be tackled in general to achieve various objectives, in a more principled manner. I hope that this is a step in the direction of making forensic matching more scientific and rigorous.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

When a crime is committed, the perpetrator almost invariably leaves behind traces of evidence, which could take various forms: DNA, fingerprints, bullets, cartridge cases, shoeprints or digital evidence. Forensic matching involves comparing pairs of samples, to infer if they originated from the same source. The underlying assumption is that pieces of evidence have identifiable characteristics that can be traced back to their source. In current practice, forensic evidence is used in courts to tie suspects to crimes. With the exception of DNA evidence, testimony is provided by trained examiners who make such judgments based on their experience and training. Despite the high stakes involved in criminal investigations, there has been a lack of scientific research to back up this claim of "individualization," or being able to reliably match evidence to source.

Historically, methods have been vetted by the legal system, as opposed to the scientific community. Beginning in the 1990s, exonerations due to DNA evidence revealed problems in many forensic science disciplines (Bell et al., 2018). Examiners have been found to have overstated forensic results, leading (at least in part) to wrongful convictions (Murphy, 2019). A 2009 National Academy of Sciences report (National Research Council, 2009) called for an overhaul of the forensic science system. With respect to forensic matching, it stated that with the exception of nuclear DNA analysis, disciplines had neither scientific support nor a proper quantification of error rates or limitations. In 2016, the President's Council of Advisors on Science and Technology (PCAST) followed up with an independent report (President's Council of Advisors on Science and Technology, 2016), specifically addressing any scientific developments in the various pattern matching disciplines. It found that little progress had been made in the intervening seven years. It recommended that scientific standards be established regarding the validity and reliability of forensic methods. Methods should then be evaluated on an ongoing basis, with respect to these standards. Additionally, methods in three important areas, DNA mixtures, latent fingerprints and firearms, should be converted "from currently subjective methods, with their heavy reliance on human judgment, into objective methods, in which standardized, quantifiable processes require little or no judgment."

With this as a backdrop, the goal of this thesis is to develop automatic methods that produce similarity scores for comparison of evidence in different domains. In Chapter 2 I first provide additional context, and describe the range of possible conclusions that can be drawn using such automatic methods. These are not restricted to addressing the problems raised in the preceding paragraph. I introduce record linkage, which is the process of finding records corresponding to the same real-world identity across different data sets. I draw the link between forensic matching problems and record linkage problems, and explore how the latter can inform the former. In particular, thinking about forensic matching as an application of record linkage gives structure to forensic matching problems, and facilitates the generation of new ideas. I establish a unified framework to develop forensic matching methods in a principled manner. In Chapters 3 and 4 I then apply this framework to two specific evidence types: firearms and digital evidence.

Matching firearms evidence has a long history in forensic science. The Association of Firearm and Tool Mark Examiners was formed in 1969 as a professional organization for practitioners of firearm and/or toolmark identification. It has hundreds of members, publishes a journal and administers a certification program. Separately, automatic methods have been used by law enforcement in the United States since 1999. The field has recently come under fire together with the other pattern matching disciplines, as described in the preceding paragraphs. Matching digital evidence, on the other hand, has had much less prominence in forensic science. The specific problem that I tackle is matching seller accounts on anonymous marketplaces; these marketplaces are a very recent phenomenon, with the first starting operations in 2011. When an online account is involved in criminal activity, such as selling drugs or stolen personal information, it is of interest to identify the real-world individual operating the account. There are no investigators dedicated solely to solving these types of cases, and no automatic methods being developed by law enforcement (to my knowledge). Court testimony with respect to linking seller accounts has been used, however, and is not immune to the problems in other forensic disciplines of overstating claims. I explain in more detail in Chapter 4 why matching seller accounts is a worthwhile problem from a forensics point of view, and demonstrate that it is possible to make inroads into solving this problem.

In the remainder of this chapter, Section 1.1 introduces Chapter 2, Section 1.2 introduces Chapter 3, and Section 1.3 introduces Chapter 4. These sections are intended to be a summary and some text is repeated from the main chapters. This work draws from three different papers, (1) Record Linkage and Matching problems in Forensics, published in the IEEE 18th International Conference on Data Mining Workshops (ICDMW) (Tai, 2018), (2) A Fully Automatic Method for Comparing Cartridge Case Images, published in the Journal of Forensic Sciences (Tai and Eddy, 2018), and (3) Adversarial Matching of Dark Net Market Vendor Accounts, accepted to the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Tai et al., 2019). Some text has also been adapted from my contribution to a book chapter in "Open Forensic Science in R" on cartridge cases (to be released at `https://github.com/sctyner/OpenForSciR`), as well as the documentation from the R packages that I have developed, `cartridges`, `cartridges3D`, and `heisenbrgr`.

Since (2) and (3) are joint work, whenever I use "we" in Chapters 3 and 4, I refer to me and my collaborators, Bill Eddy in Chapter 3, and Nicolas Christin and Kyle Soska in Chapter 4.

1.1 Developing Forensic Matching Methods

Forensic matching refers to comparing pairs of samples to infer if they come from the same source. Typically, one of these samples is from the crime of interest, and is called the *questioned sample*. Pairs that are from the same source are *matched pairs* or simply *matches*, and pairs from different sources are *non-matches*.

Comparisons can be thought of from various perspectives. One common distinction in forensics is between *investigation*, and *evaluation*: an investigation is identifying a number of candidates that could potentially come from the same source as the questioned sample, and an evaluation compares the questioned sample to a specific sample, to determine if they can be attributed to the same source. The evaluative step in particular has been dogged by controversy, due to the subjective nature of the current system, in which examiners make judgments based on their experience and training. Automatic methods have been suggested to alleviate these problems, since they produce reproducible similarity scores and can be designed to produce quantitative measures of uncertainty, such as error rates and measures of the weight of evidence.

Specifically, automatic methods can be used to generate the following forensic conclusions, not necessarily restricting oneself to either the investigative or evaluative perspective: 1) a ranking of similarities of pairs; 2) a match or non-match conclusion; 3) a linked or disambiguated data set (i.e., each item is attributed to a unique source); 4) estimating random match probabilities or likelihood ratios, as measures of the weight of evidence. Much more can be said about the reasons one might like to draw these conclusions, and this will be elaborated upon in Section 2.2. An automated method is then able to make these conclusions, through the appropriate selection of cutoffs, for example. Additionally, methods should be open source, for full transparency and proper evaluation. They should not require human input, if they are to address the concerns raised about subjectivity. Other considerations are scalability and resilience to adversarial attacks.

Record linkage or data matching in statistics or computer science is the process of inferring which entries in different databases correspond to the same real-world identity, in the absence of a unique identifier (Christen, 2012). Depending on the field, data matching is known by different names, in particular in statistics, "record linkage" is used, in applications such as linking Census records, death records, bibliographic and other databases. A typical record linkage process includes the following steps. Beginning with records from two databases A and B to be linked, fields such as name and address are pre-processed and standardized. This results in some set of features for each record. Next indexing methods might be used to reduce the comparison space, for example candidate pairs in a demographic database might be restricted to those born in the same month, with all other pairs not being considered for linkage. Now pairwise comparisons are generated, consisting of one or more similarity measures. These pairwise data are then classified into matches and

non-matches (and possibly a third category, potential matches). This classification step may be designed to incorporate restrictions such as one-to-one links, or preserving transitivity. A clerical review may also be conducted for potential matches, where a human examines pairs manually to decide the appropriate classification. Finally an evaluation of the results is conducted.

Forensic matching can be thought of as an application of record linkage; simply think of records as evidence samples, and real-world entity as the source of the sample. Steps traditionally used in forensic matching can be demonstrated to correspond to the steps described in the preceding paragraph. By thinking about forensics problems in the context of record linkage, one immediately has well-developed frameworks and tools at one's disposal.

Later in Chapter 2, I describe a framework that can be used to develop automatic forensic matching methods in a systematic manner. This is a simplification of the record linkage process, adapted to a forensic context. It includes an optional post-processing step to generate transitive closures, using hierarchical agglomerative clustering. The evaluation step is standardized and uses precision and recall, together with the area under the precision-recall curve as a numerical summary. This framework will be used in Chapters 3 and 4.

Chapter 2 concludes with some other specific suggestions on how the record linkage literature can inform forensic matching. To summarize, the contributions of Chapter 2 are:

1. Demonstrating a link between forensic matching problems and record linkage
2. Establishing a framework to approach forensic matching
3. Suggesting ways that record linkage can inform forensic matching.

1.2 Matching Firearms Evidence

Firearms evidence was one of the three important areas listed in the PCAST report that could benefit from more objective analyses. Firearms evidence includes marks left on bullets as well as cartridge cases due to the firing process of the gun. For bullets, striation marks are left due to imperfections in the gun barrel due to the manufacturing process. In the case of cartridge cases, there are at least two types of marks that are of interest. First, the firing pin hits the primer material at the base of the cartridge, leaving a firing pin impression. The subsequent explosion (which launches the bullet) also causes the cartridge case to be pressed against the breech block of the gun, leaving impressed marks known as breechface marks. All of these types of marks are thought to individualize a gun, hence law enforcement officers frequently collect bullets and cartridge cases from crime scenes, in hopes of connecting these to retrieved guns, or connecting crime scenes where the same weapon was used. In this chapter, one type of firearms evidence is considered: breechface marks left on cartridge cases.

In current practice, retrieved cartridge cases are entered into a national database called the National Integrated Ballistics Information Network (NIBIN), through a computer-based platform which was developed and is maintained by Ultra Electronics Forensic Technology Inc. (FTI). This platform captures an image of the retrieved cartridge case and runs a proprietary search algorithm, returning a list of top ranked potential matches from the database. Firearms examiners then examine this list and the associated images, to make a judgment about which potential matches warrant further investigation. The physical cartridge cases associated with these images are then located and examined under a comparison microscope. The firearms examiner decides if there are any matches, based on whether there is "sufficient agreement" between the marks (AFTE Criteria for Identification Committee, 1992), and may bring this evidence to court. As described, this current system has come under scrutiny, and there has been a push in recent years towards automatic comparison methods.

Such methods have been developed by engineers and scientists both in industry and in academia, and a full review is in Section 3.1. As far as one can tell, none of these methods are open source, which is an impediment to transparent evaluation.

I apply the matching framework developed in Chapter 2 to the problem of matching cartridge cases. In particular, since the data involved are images, the steps of generating individual-level features and pairwise similarities are non-trivial. The methodology I develop is fully automatic, and can be used for both 2D photographic images captured using reflectance microscopes, and 3D topographies captured using confocal microscopes. Compared to existing methods, there are several specific improvements. First, the selection of breechface marks is automated using image processing techniques such as edge detection for 2D, and RANdom SAmple Consensus (RANSAC) (Fischler and Bolles, 1981) for 3D. The addition of a pre-processing step to remove circular symmetry reduces the likelihood of spurious correlations. The pairwise similarity score computed is a correlation between pixel values of the two images, maximized over rotations and translations. An unsupervised approach is used where different thresholds on this similarity score are considered. The evaluation step is standardized using precision-recall graphs, and a numerical summary can be attached by reporting the area under the curve. This is an improvement over both visual examinations of the overlap region of match and non-match similarity score distributions and methods which estimate this region by making assumptions about the data. Finally, pairs are evaluated independently and may result in intransitive links; resolving these and generating a linked data set has not been given much thought in the firearms literature. There are situations in which having intransitive links is undesirable, for example when comparing algorithm results with examiner proficiency tests, or when evaluating on reference data sets (more details in Section 3.2). Application of the last step of the matching framework addresses this issue.

I evaluate performance using over 1000 images from an open-access research database of bullet and cartridge case toolmark data maintained by the National Institute of Standards and Technology (NIST). Overall performance for 3D topographies is superior to that for 2D optical images, and excellent performance

is achieved (area under the precision-recall curve of over .9) in 9 out of the 13 data sets in 3D. Performing the clustering step generally improved performance for both 2D and 3D, particularly for data sets with middling to good (but not excellent) results. Performance is generally comparable to other published work, although some evidence suggests that the Congruent Matching Cells method developed by NIST (Song, 2013) is an improvement over the standard measure of correlation that we used, for some subsets of data.

As an additional step I examine some claims and existing anecdotal evidence about the ease of comparing specific brands of guns, or types of ammunition. One conclusion is that it is not possible to make blanket statements such as "Sig Sauers mark poorly," due to a large variance in performance among different models of the gun brands in different studies. Ammunition effects have been suggested, where it is harder to match cartridge cases when different brands of ammunition are fired from the same gun. There is some evidence to suggest this is true, but it is hard to make a definite conclusion since only two data sets involved different ammunition brands, and this is confounded with other factors such as the firearm make and model used. A final hypothesis was that firearms that are consecutively manufactured may be more difficult to differentiate, resulting in false links. This was not found to be the case.

To summarize, some specific contributions of Chapter 3 are:

1. Developing an open source and fully automatic method for comparing cartridge cases, in both 2D and 3D
2. Proposing ways to generate a linked data set and standardize evaluation
3. Comprehensively evaluating on over 1000 images from 13 data sets.

1.3 Matching Seller Accounts on Anonymous Marketplaces

Online anonymous marketplaces are digital marketplaces that provide anonymity protections beyond traditional marketplaces such as Amazon and eBay (Soska and Christin, 2015). They run on the dark web using anonymizing browsers such as Tor, and payment is typically made using cryptocurrencies such as Bitcoin. Because of such anonymity protections, these marketplaces are most commonly used for the sale of illicit products, in particular, drugs. Silk Road, an early such marketplace, began in February 2011, and was shut down by law enforcement in October 2013. Since then, other marketplaces have opened and closed; sellers have done business on multiple marketplaces concurrently, as well as moved between marketplaces (suggesting that they do not necessarily have any particular loyalty to particular marketplaces, and are open to operating multiple accounts). By conducting such business, sellers invariably leave evidence of criminal activity, since these marketplaces are in the public domain. Anyone with some level of technological skills would be able to browse user profiles and products sold through these marketplaces and accounts. To put it simply, these sellers hide in plain sight. The forensic challenge then is to track down the real-world individuals behind these accounts.

Law enforcement has made over 100 anonymous marketplace-related arrests starting in 2012. Marketplace operators, sellers and buyers have all become targets of investigation. For example, in 2017, in a global effort between various agencies, two of the largest marketplaces, Hansa and AlphaBay were taken down. In 2018 the Department of Justice put together a multi-agency team which has subsequently targeted both sellers and buyers (Federal Bureau of Investigation, 2018). Often, linking accounts on the same or on different marketplaces helps investigators identify real-world individuals, in the same way that matching pattern evidence in other forensic disciplines helps to generate leads in investigations. To this end, investigators have been known to try to link various online accounts. Based on descriptions in court records, investigators have used techniques such as manually matching account handles, cryptographic public keys (frequently advertised on accounts), items sold, as well as searched forum discussions for account mentions (United States District Court, Eastern District of New York, 2016; United States District Court, Eastern District of California, 2016). All of these rely on manual investigation, which can be lengthy and time-consuming. No attempt appears to have been made by law enforcement to automate this process. Additionally, even though the misuse of forensic evidence and overstatement of forensic results has so far been most problematic in the pattern matching disciplines, it is not implausible to see this playing out in the form of digital evidence. Imagine a hypothetical situation in which a victim of a fentanyl overdose is known to have purchased the product from a particular online account (call this the questioned account). Now, an investigator may testify with absolute certainty that a different account that is operated by a known individual has the same ownership as the questioned account, implicating this known individual. This statement could be backed up by a seemingly scientific analysis, for example the style of writing or use of particular symbols. This could result in wrongful convictions in the same way as in other pattern matching disciplines. Use of automatic methods in matching digital evidence can similarly remove this potential subjectivity. Needless to say, they should first be comprehensively tested, and error rates should be properly quantified.

In the literature, several attempts to automatically match seller accounts exist, but most of these rely on exact matching schemes that are inflexible and vulnerable to impersonation attacks. In Chapter 4, I again apply the matching framework developed in Chapter 2. Here I make use of data scraped by collaborators from publicly available marketplace sites in the period from 2011 to 2018. An individual record is all the captured historical information relating to an individual account. This includes profile information, items sold and their associated information, such as item descriptions, prices and feedback received. Data are pre-processed and fields are standardized. Pairwise comparisons are then generated, by extracting string and numerical similarity measures. Many of these are common in the record linkage literature, and are adapted to the current context. In anonymous marketplaces, there is an incentive to conceal one's identity or impersonate other sellers with good reputations. Hence features are selected to be costly for such adversaries to mimic. Ground truth for whether pairs of accounts belong to the same seller are unavailable, so labels are generated using different heuristics. A supervised random forest approach is then used (Breiman, 2001).

Depending on the goals of the analysis, there is utility in performing the final hierarchical clustering step to generate a linked data set: for example data involved here are scraped data rather than reference data, and so can be thought of as a census of marketplace vendors. Linking the data can hence give an estimate of the number of sellers involved in such marketplaces. Further, researchers studying marketplace ecosystems might be interested in conducting analysis related to seller behavior, which is more suitably done on a seller-level than an account-level.

The evaluation is then done using precision and recall, using two sets of generated labels. The model is compared to a baseline of a threshold-based approach using edit distances between account IDs. Finally, several case studies are examined as a further evaluation on specific examples of interest, where ground truth is available through criminal complaints, forum discussions and/or news reports.

In terms of the results, using generated labels, the model can achieve more than 75% precision and recall by selecting appropriate cutoffs for the classifier. Performance is superior to using a baseline of only edit distance between IDs. The methodology works particularly well for accounts with significant sales volume (achieving around 90% recall at 75% precision for the top 30% of accounts by sales volume). Performing the clustering step slightly improves model performance. After the clustering step and using a set of chosen parameters, the 22,163 accounts with at least one confirmed sale can be mapped to 15,652 distinct sellers. 12,155 sellers (77%) operate only one account, while the remainder operate up to 11 accounts. Finally in case studies the model discover links documented in court records in 4 out of 7 cases. In some situations it can automatically discover reported impersonation attempts, non-trivial links between accounts, and instances in which the labels are incorrect.

To summarize, some specific contributions of Chapter 4 are:

1. Developing an open source and fully automatic method for matching seller accounts on anonymous marketplaces
2. Demonstrating an application of record linkage to an adversarial context
3. Comprehensively evaluating on eight years of marketplace data.

Chapter 2

Developing Forensic Matching Methods

Chapter 1 introduced the idea of forensic evidence and matching problems in forensics. I stated that the goal of forensic matching was to infer if pairs of evidence came from the same source. There are various nuances involved, and in this chapter I start by going over some preliminaries. I then write about the types of forensic conclusions that might be of interest, and restate the goals in terms of being able to make these conclusions. I then introduce record linkage, and argue that forensic matching problems can be thought of as an application of record linkage. Finally, I describe a framework that will be used in Chapter 3 and 4. I conclude the chapter by exploring other specific ways the record linkage literature can inform forensic matching.

2.1 Preliminaries

First, define *forensic matching* as comparing pairs of samples to infer if they come from the same source. One of the samples in question is often a sample from a crime of interest, for example a latent print, retrieved cartridge case, or DNA sample. This is called the *questioned sample*. Comparisons can then be made with the questioned sample; pairs that are from the same source are called *matches* or *mated pairs*, while pairs from different sources are called *non-matches* or *non-mated pairs*. These comparisons can be thought of from various perspectives. One distinction is between *investigation*, and *evaluation*: an investigation is identifying a number of candidates that could potentially come from the same source as the questioned sample, and an evaluation compares the questioned sample to a specific sample, to determine if they can be attributed to the same source.* The investigative problem can be thought of as a one-to-many comparison to generate

*A distinction between "same-source" and "specific-source" is also sometimes made, where the former refers to coming from a common, unknown source, while the latter refers to coming from a specific, known source. This is less important for the current purposes, since statistically they are the same problem.

investigative leads from a database, while the evaluation problem is a one-to-one comparison, for example when evidence is compared with a sample taken from a suspect.

The *evaluation* problem is also sometimes called an *identification* or *individualization*, in the sense of an examiner identifying that a particular questioned sample came from a specific, individual source (see e.g., AFTE Criteria for Identification Committee, 1992). Much debate has surrounded this issue, because of the lack of scientific evidence backing up claims of individualization for many forensic disciplines. Instead of making unsupported statements of absolute certainty, there has been a push in recent years to 1) quantify examiner error rates in terms of false positives and false negatives when making such conclusions, and 2) attach a degree of uncertainty or measure of the weight of evidence to these statements. Automatic approaches have been suggested to deal with these two issues; these can provide the associated quantitative measures with respect to the points raised. Taking the human entirely out of the loop from the decision-making process eliminates associated biases in court testimony due to subjective opinions. The focus of this thesis will be on such automatic approaches. Before going into the details, the second issue of weight of evidence is examined in more detail as follows.

The weight of evidence refers to its probative value. For example, imagine a hypothetical scenario where a bank robber's getaway car was a 1957 Ferrari. Implicating someone associated with a 1957 Ferrari would carry much more weight in courts compared to if the getaway car was a Honda Accord. This example is fairly straightforward to understand, but in cases which require an assessment of the "rarity" of DNA or fingerprint evidence, the probative value is less obvious. To this end, the likelihood ratio has been put forth as a suitable quantitative measure summarizing the weight of evidence. European courts have moved decisively in this direction, with guidelines for forensic laboratories now recommending the reporting of likelihood ratios (Champod et al., 2016). The likelihood ratio is first defined for matching problems, in the context of the following hypothesis test:

H_0: The sample is from the suspected source, i.e., the comparison pair is a match

H_A: The sample is not from the suspected source, i.e., the comparison pair is a non-match

This is known as the *source hypothesis*. This is sometimes confused with but is distinct from the *offense hypothesis*, which is that the suspect is guilty. Here the concern is only with the source hypothesis. The likelihood ratio is then defined in Equation 2.1.

$$LR = \frac{\mathbb{P}[\text{Observe evidence} \mid H_0 \text{ true}]}{\mathbb{P}[\text{Observe evidence} \mid H_0 \text{ false}]} \tag{2.1}$$

There is no consensus on how these likelihood ratios should be estimated, however. For example in DNA, a generative model is used to determine the probabilities of observing particular DNA sequences. More details are in Section 2.5; this is an example of a feature-based likelihood ratio (FLR), defined in Equation 2.2:

$$FLR = \frac{f(x \mid \theta_0)}{f(x \mid \theta_A)}, \tag{2.2}$$

where θ_0 and θ_A are the model parameters under the null and alternative hypotheses, f is a density and x represents *pairwise* features (possibly multivariate). This requires that the distributions of the comparison vector under the null and alternative hypotheses be known. In many forensic fields this is not the case, because of a lack of knowledge of a scientific basis in which say, bullet striations are produced. In such cases, features are often summarized into a single similarity score, and instead of feature-based likelihood ratios, score-based likelihood ratios (SLR) have been suggested. These compare distributions of similarity scores instead of features (Hepler et al., 2012), as defined in Equation 2.3:

$$SLR = \frac{g(s(x) \mid \theta_0)}{g(s(x) \mid \theta_A)}, \tag{2.3}$$

where $s(.)$ is a function that maps pairwise features to a score. A useful similarity score would have high values for matching pairs and low values for non-matching pairs.

Both parametric and non-parametric methods have been suggested for estimating these distributions (Riva and Champod, 2014). A separate issue is what scenarios should be included in the alternative hypotheses (Iyer and Lund, 2017); in other words, if the suspect is not the source, what alternative sources are to be considered? Other issues have also been raised, such as non-monotonicity of likelihood ratios with scores (Park, 2018). Given these concerns, despite Europe's enthusiasm in embracing likelihood ratios, the question of quantifying weight of evidence is far from being resolved.

In the literature a two-step approach has been proposed (see e.g., Park, 2018), where the steps are (1) Develop a statistical method to produce a score or measure of similarity for a comparison pair, and (2) Assess the value of evidence through the score-based likelihood ratio (SLR). In this thesis I focus on the first step. Instead of using scores only to estimate the weight of evidence, in Section 2.2 I list some other ways that scores can be used. I also provide some suggestions for the second step of the two-step process (estimating measures of the weight of evidence), but overall the weight of evidence has not been an emphasis in this thesis. Nevertheless, it is an important issue in forensics and will be the subject of future work.

2.2 Types of Possible Forensic Conclusions

This thesis focuses on developing tools to automatically generate similarity scores for pairwise comparisons. The following is a range of possible conclusions that can be drawn, depending on the goals and data available, without limiting oneself to the investigative or evaluative perspectives. First, given a pair or a set of comparisons, one might be interested in which pair is more or most similar. This situation is straightforward. The set of pairwise similarity scores is sufficient to make a statement such as "Comparison 1 contains images

that are more similar than Comparison 2," or "these are the top 10 pairs with highest similarity scores out of the 100 comparisons being made." Such conclusions could be used to generate investigative leads, where the top 10 (say) pairs are selected for further investigation. A different context in which this type of conclusion could be used is by examiners for blind verification. Blind verification means that an examiner first comes to their own conclusion, and then verifies this using an automatic method, making a conclusion such as "Based on my experience and training, this pair of cartridge cases come from the same gun. The same pair also had a score of .7, the highest similarity score returned by *[[some algorithm]]* among *[[some subset of pairs being considered]]*."

In other situations, it might be of interest to designate a similarity cutoff above which some action is taken. The selection of such a cutoff depends on the goal. For example, similar to the above situation, one might be interested in selecting pairs above a cutoff for further manual investigation, instead of simply picking the top 10 pairs. Alternatively, a cutoff could be used to decide if pairs are matches or non-matches, in other words if they come from the same source or not. Such a conclusion could be of interest in criminal cases, where a conclusion of match or non-match is required to decide if a person should be implicated in a crime. In the first investigative case a lower cutoff might be set to ensure high recall, while in the second case a much higher cutoff might be necessary, since the costs associated with falsely implicating a suspect are disproportionately high.

Next, within some data set, one might be interested in performing multiple (or all) pairwise comparisons, to obtain a linked or disambiguated data set, meaning that each item in the data set has been attributed to some source. The said source may be known or unknown; the point is that items are grouped into clusters sharing a common source. An example of when generating a linked data set might be useful is if one would like to compare the performance of an automated method with that of human examiners. A common setup of proficiency tests is to have multiple reference samples from a single source, and a few questioned samples of different origins. Examiners are then asked to compare each of the questioned samples to the reference samples, to make a conclusion of "identification" (match), "elimination" (non-match), or "inconclusive," for each of the questioned samples independently. For example, in the Collaborative Testing Services (CTS) firearms examination 526 test, examiners are given three reference samples from one gun, and four other questioned samples (Collaborative Testing Services, 2015). One of the questioned samples is from the same gun as the reference, and two others are from a common, different gun, while the last is from a third different gun. To be clear, call the three guns guns R, A and B. The source of the reference samples is gun R. Then, the origin of the four questioned samples are gun R, gun A, gun A, and gun B. Examiners are not required to state the results of all pairwise comparisons within this set (only the comparisons of each questioned sample to the set of reference samples), but in order for them to come to their conclusions, they are likely making these multiple comparisons, since three reference samples are provided. To accurately compare the performance of an automated method to that of examiners, it is appropriate to perform all pairwise comparisons to get

a disambiguated data set, and then compare the resulting error rates. There are other situations in which the goal might be to produce a linked data set. These are described in Section 3.2 for cartridge cases, and Section 4.1.3 for seller accounts.

Now, closer to the evaluative context, one might be interested in estimating a probability of getting a higher similarity score by chance, sometimes known as a "random match probability." Such a probability roughly corresponds to the denominator of the likelihood ratio in Equation 2.3. Generating this probability would require appropriate data on the distribution of similarity scores for non-matching pairs in some population of interest. For example, if a similarity of .7 for the pair of interest is obtained, .7 is compared to the corresponding non-match score distribution, and the probability of interest is the probability that a random draw from that distribution is larger than .7, say p_0. The conclusion then, is that if the pair was a non-match, the probability of getting a score higher than .7 is p_0. If the value of p_0 is small, the conclusion provides evidence against the hypothesis that the pair of interest is a non-matching pair. Hypothetically, an examiner could use such a conclusion as court testimony: "The pair had a score of .7, the highest similarity score returned by *[[some method]]* among *[[some subset of interest]]*. If the two cartridge cases are not a match, the probability of observing a higher similarity score among *[[some relevant non-matching population]]* is *[[p_0]]*."

Finally, one might be interested in estimating a score-based likelihood ratio, as defined in Equation 2.3. Estimating such a likelihood ratio would require both a distribution of scores for matching and non-matching pairs, in some population of interest. With a likelihood ratio of l_0, the interpretation in context is that the probability of observing a particular similarity score if the pair is a match is l_0 times that if the pair is a non-match. Values of l_0 larger than 1 provide evidence in favor of the null hypothesis that the sample is from the suspected source. While such a likelihood ratio provides a quantitative measure, one must be careful not to make statements such as an estimated likelihood ratio being *the* likelihood ratio for a particular comparison; any estimate is simply one of many possible estimates that can be derived given the associated assumptions.

In summary, various conclusions can be made, given an automated method producing pairwise similarity scores. Possible outcomes include 1) a ranking of similarities of pairs; 2) a match or non-match conclusion; 3) a disambiguated data set; 4) random match probabilities or likelihood ratios. These conclusions require assumptions to be made, both in relation to the model being used to generate scores, as well as the relevant populations of interest.

2.3 Goals of Automatic Methods

Here the goals of the thesis are restated explicitly. My aim is to develop automatic methods to compare pairs of evidence or samples. Automatic methods should produce a similarity score, that can be used in a variety

of ways to draw different conclusions, as described in the preceding section. Some types of conclusions require the selection of a cutoff or method to differentiate matching scores from non-matching scores, while others require reference distributions for scores of matching or non-matching pairs. Given appropriate data, some guidance is provided on how to accomplish these different tasks.

In light of criticisms of subjectivity in examiners' conclusions, methods should as much as possible operate without human input. There is the possibility of algorithmic bias, but a fully automatic algorithm still has the advantage of producing reproducible results. Additionally, methods should also be open source. Being open source allows methods to be properly tested, and gives any interested party the opportunity to verify the methodology, understand situations in which it succeeds or fails, experiment with various parameters or pre-processing steps, and make tweaks to suit their purposes, if desired.

Some additional properties that might be desirable are scalability and resilience to adversarial situations. In Section 2.1 the distinction between investigation and evaluation was made; for investigation it might be of interest to perform a large number of comparisons, in which case scalability of the methodology is important. In some situations there could be adversaries that are interested in seeing comparison algorithms fail, for example, criminals who alter their fingerprints to evade detection. It would be ideal if automatic comparison methods are resilient to such attacks. Some examples of such adversarial behavior are discussed in Chapter 4.

2.4 Record Linkage

In the current section I introduce the field of record linkage, and do a brief overview of the literature, with a heavier focus on aspects that are pertinent to the work described later in this thesis. A full review is out of the scope of the thesis, but the interested reader can refer to e.g., Christen (2012). The connection between record linkage and forensic matching problems is discussed in Section 2.5.

Record linkage or data matching is the process of inferring which entries in different databases correspond to the same real-world identity, in the absence of a unique identifier. When dealing with duplicate entries in a single database, it is more commonly known as deduplication or duplicate detection. Depending on the field, it is known by different names, in particular in statistics, "record linkage" is used, with applications such as linking Census records, death records, bibliographic databases, and so forth.

A standard framework for record linkage is in Figure 2.2, adapted from Christen (2012). Begin with records from database A and B to be linked. Each of these records might consist of information such as name, address, date of birth, and so forth. These need to be pre-processed and standardized across databases, for example the date of birth field may not always be recorded in the same format. This results in some set of features for each record. Next indexing methods might be used to reduce the comparison space. For example, one might only compare records where the first and last names begin with the same letter. The other records are simply predicted to be non-matches. The next step is to generate pairwise comparisons from these records

to be compared. Each pairwise comparison may consist of one or more similarity measures. These pairwise comparison data are then classified into matches and non-matches (and possibly a third category, potential matches). This classification step may be designed to incorporate certain restrictions such as one-to-one links, or preserving transitivity. A clerical review may also be conducted for potential matches, where a human examines pairs manually to decide the appropriate classification. Finally an evaluation of the results can be conducted. Each of these steps is described in greater detail in the following sub-sections.

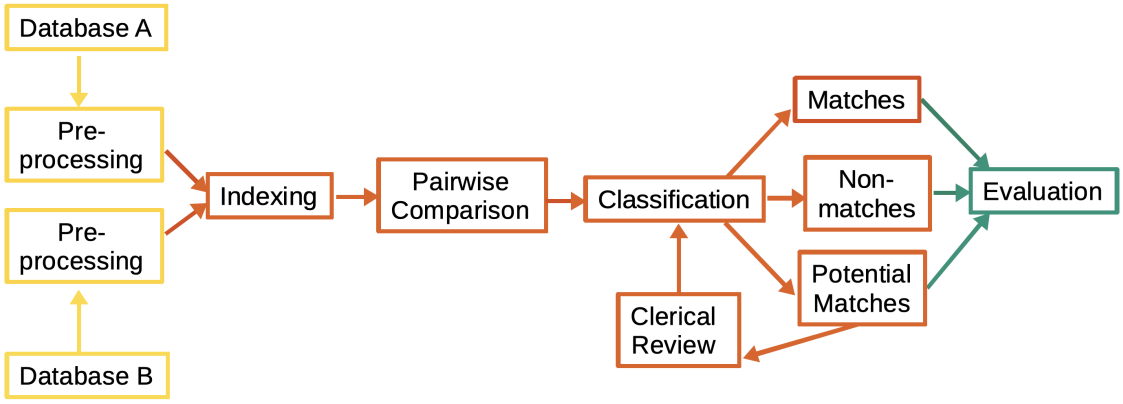


Figure 2.1: Standard framework for statistical record linkage problems, adapted from Christen (2012).

2.4.1 Pre-processing

Pre-processing involves selecting and cleaning up the data fields that are available at an individual record level. Typical steps include handling missing values, standardizing fields such as date of birth, and segmenting and extracting various features. For example, an address may be segmented into street number, street name and apartment number. Other common examples include converting cases and removing whitespace. The outcome of this step is individual records with standardized features that can be compared across data sets.

2.4.2 Indexing

The number of pairs increases quadratically with the number of individual records. Indexing aims to reduce this quadratic complexity of the data matching process through the use of data structures to efficiently generate candidate record pairs that likely correspond to matches. A straightforward and common approach to indexing is blocking, where records are grouped into blocks based on some similarity criteria, for example the first letter of the first name. Pairwise comparisons are generated only for records in the same block, and the other pairs are simply predicted to be non-matches.

2.4.3 Pairwise Comparison

In this step, pairwise similarity measures are generated from the individual-level fields or features. These individual features may include strings, numbers, dates, ages, times, geographical distances, or complex data, e.g., images or households. A similarity function between two attribute values a_i and a_j is defined as $s = \text{sim}(a_i, a_j)$. Typically, $0 \leq s \leq 1$, and a larger value indicates higher similarity. Some common string comparison measures include edit distance, also known as Levenshtein distance (Levenshtein, 1966), defined as the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. The Jaro-Winkler distance is a common name comparison method, which takes into account name-specific features. For example, the similarity between two strings is higher if their beginning is the same and differences only occur toward the middle and end. Another common similarity measure is the Jaccard similarity, defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ for sets A and B (Jaccard, 1908). This can be used with q-grams, which are substrings of length q. For example, let a_i be 'john' and a_j be 'joan'. 2-grams of 'john' are 'jo', 'oh', 'hn', which is treated as set A. 2-grams of 'joan' are 'jo', 'oa' and 'an', which is Set B. Then, $\text{sim}(a_i, a_j) = J(A, B) = \frac{1}{5}$.

2.4.4 Classification

After pairwise comparisons have been generated, there are numerous methods to classify the pairs into matches and non-matches. Several of them are described here. The most straightforward is simply a threshold-based approach. The similarity scores can be summed, and if they exceed a user-specified threshold, the pair is classified to be a match. On the other hand, rules-based approaches rely on heuristics on each of the similarity scores. For example, if similarity measure A exceeds .2 or similarity measure B exceeds .5, label the pair to be a match.

Fellegi and Sunter (1969) proposed a probabilistic framework for assigning matches, and this subsequently gained widespread popularity and is often described as the "traditional" approach to record linkage. Briefly, the framework is as follows. Let A and B be two databases to be linked, and let $a \in A$ and $b \in B$ be generic records in A and B. Let $M = \{(a, b); a = b, a \in A, b \in B\}$ be the matched set, and $U = \{(a, b); a \neq b, a \in A, b \in B\}$ be the unmatched set. Let $\gamma_{ab} = (\gamma_{ab}(1), ..., \gamma_{ab}(k))$ be the comparison vector between a and b, having k components. Then the Fellegi-Sunter method makes use of cutoffs on the following likelihood ratio in favor of $(a, b) \in M$:

$$\frac{\mathbb{P}[\gamma_{ab} = g | (a, b) \in M]}{\mathbb{P}[\gamma_{ab} = g | (a, b) \in U]}, \tag{2.4}$$

where g is the observed k-dimensional comparison vector. If the likelihood ratio exceeds some cutoff, the pair is classified as a match, and if it is below some other cutoff, a non-match. These cutoffs are determined by pre-specified limits on false positive and false negative rates. For estimation of the likelihood ratio, conditional

independence is often assumed, where each component of the comparison vector is assumed to be independent of the others, given the match status M or U. If the data-generating process is known, and population frequencies for each of the k individual-level features used to generate γ_{ab} are known, the likelihood ratio can be computed in a straightforward manner using the conditional independence assumption. Alternatively, when such information is not available, the EM algorithm has been frequently used to estimate parameters involved in the conditional distribution of γ_{ab}, and $\mathbb{P}[(a, b) \in M]$ (Winkler, 2000).

The above are unsupervised approaches; supervised approaches are also commonly used. In a supervised method there needs to be some mechanism for generating training labels for whether a pair is a match or not, for example by manual labeling. Any standard supervised classifier can then be used, for example logistic regression, decision trees, support vector machines, etc.

Additional restrictions sometimes need to be imposed in the matching process, and these are typically done in the classification step. For example, a one-to-one constraint may be added, meaning that a record in database A can be matched to only one record in database B. When multiple pairwise comparisons are done independently, there could be intransitive links, which might be undesirable in the matching process. For example, if A matches B and B matches C, but A does not match C, there is a conflict. This can be resolved either by adding or removing links. One approach to solving this problem is to use hierarchical clustering on pairwise comparison results (as a post-processing step), as is described in Section 2.5.

2.4.5 Evaluation

The classification step generates predictions for whether pairs are matches or non-matches. The performance of this matching process can be assessed in multiple ways. Assuming that there is some set of ground truth labels for whether pairs are matched or not, one can use the same performance measures used in classification problems, at a pairwise level. These measures are based on the number of true positives, false positives, true negatives and false negatives, defined as follows.

True positives Actual matched examples that are predicted to be matched.

False positives Actual non-matched examples that are predicted to be matches.

True negatives Actual non-matched examples that are predicted to be non-matches.

False negatives Actual matched examples that are predicted to be non-matches.

These correspond to the following cells in a typical confusion matrix:

	Predicted	
Truth	0	1
0	True Negative	False Positive
1	False Negative	True Positive

Based on the four counts above, various performance metrics can be generated, such as accuracy, false positive rate, false negative rate, precision and recall. When the cutoff for predicting positives is not fixed, a popular performance metric is the receiver operating characteristic (ROC) curve, which plots True Positive Rate $= \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$ against False Positive Rate $= \frac{\text{False Positives}}{\text{False Positives+True Negatives}}$.

In record linkage problems, there is a large class imbalance since most pairs do not match. Typically, pairs are overwhelmingly true negatives. Measures that have the number of true negatives in the denominator, such as false positive rate, will often be close to zero. Hence popular alternatives are Precision $= \frac{\text{True Positive}}{\text{True Positive + False Positive}}$ and recall, which is the same as true positive rate. Both measures range from 0 to 1, with 1 being perfect classifier performance. The area under the precision-recall graph can also be computed; this also ranges from 0 to 1, and maximum performance is achieved when the graph passes through the top-right corner, giving an area under the curve of 1.

2.5 Forensic Matching and Record Linkage

At first glance, record linkage and forensic matching might seem like separate problems, but there is a correspondence between the two – approaches traditionally used in forensic matching fit into the framework of statistical record linkage problems described in Section 2.4. This correspondence has been noted by statisticians in the past, but to my knowledge this has not been formalized or exploited. Copas and Hilton (1990) mention comparing an unidentified fingerprint with fingerprints of known individuals as a possible application of computer matching. Skinner (2007) establishes a correspondence between statistical disclosure control and forensic statistics based on their common use of the concept of "probability of identification," focusing on how disclosure control can learn from the literature on forensic identification.

Automated or semi-automated approaches have been developed for forensic matching by people working in different fields, in an ad-hoc manner. Using the examples of DNA and bullet matching, I demonstrate how these methods can be treated as record linkage problems, noting that this extends to automated forensic matching methods for other evidence types. Thinking about forensics problems in the context of record linkage lends structure to forensic problems and allows one to easily generate new ideas. Tools from the record linkage literature can be leveraged for forensic matching, in ways that are explored in this section.

From a record linkage perspective, forensic matching problems are unique in the sense that the data are often complex, unstructured data, such as multivariate data, images and text. In some applications there might also be an adversarial element, where individuals might stand to gain from the failure of the matching process, either by avoiding being matched, or being falsely matched. Examples of this are discussed in Chapter 4.

DNA

DNA matching is often described as the gold standard of forensic matching methods, because there is a scientific basis for performing comparisons and computing likelihood ratios. The United States government maintains the Combined DNA Index System (CODIS), which is a database containing DNA profiles from crime scenes as well as from known offenders.

Each DNA profile consists of information from 13 (or more) locations on the DNA. In particular, these are locations with short tandem repeats (STRs), sections of DNA with repeating patterns. The number of repeats at each location occurs with different frequencies for each individual, and the number of repeats at each of the 13 locations form a DNA profile. By biological theory, all of the STRs are independent, and the distribution of the number of repeats at each location differs by race (White, African American, Asian, etc.). If information at all of the STR locations match exactly, a comparison is reported as a match. A likelihood ratio can then be estimated as

$$LR = \frac{\mathbb{P}[\text{13 STRs identical} \mid \text{Samples are from the same source}]}{\mathbb{P}[\text{13 STRs identical} \mid \text{Samples are from different sources}]}. \tag{2.5}$$

The numerator is usually taken to be 1, and the denominator is estimated using a generative model, based on known frequencies of repeats at each location in the relevant reference population (the race of the suspect). The likelihood ratio can be interpreted as the number of times more likely one is to observe the evidence if the profiles come from the same source than if they do not, and hence quantifies the weight of evidence. This number is often reported in courts.

To be explicit in using the steps introduced in Section 2.4, each DNA sample is treated as a record, and is summarized using 13 features which correspond to the 13 STRs. The similarity function for a pair of samples inputs two vectors of length 13, with each entry being a number of repeats at the STR location. Define $d_H(x, y)$ to be the Hamming distance, $d_H(x, y) = \frac{1}{13}\sum_{i=1}^{13} I(x_i \neq y_i)$. The similarity function is $1 - d_H(x, y)$, where $0 \leq d_H \leq 1$. A threshold-based classification method is then used with a cutoff of 1 to be classified as a match. In words, all the 13 STRs need to have the same number of repeats to be considered a match. Since this is an exact matching scheme, transitivity is automatically preserved when multiple comparisons are conducted independently.

The estimation of the likelihood ratio is an additional step, and the methodology used is essentially the same as in the Fellegi-Sunter model, in the case where population frequencies of the features are known (Fellegi and Sunter, 1969). In this case, the distribution of the number of repeats at each location is estimated for each race from standard population databases, and is treated as known.[†] The Fellegi-Sunter model and its connection to likelihood ratios in forensic matching is discussed in greater detail in Section 2.5.2.

[†]Using the language from Fellegi and Sunter (1969), in the numerator of Equation 2.5 are m-probabilities and in the denominator u-probabilities.

Bullet Matching

A gun is thought to leave identifiable marks on bullets, and if these are retrieved from crime scenes, they can be compared to other samples, to infer if they were fired from the same gun. Rifling, manufacturing defects, and impurities in the barrel create striation marks on the bullet.

Bullet matching has a long history, but unlike DNA matching, there is no well-understood scientific basis upon which marks are created, and hence for comparisons to be made. Automatic comparison methods have been developed by engineers and scientists both in industry and in academia (e.g., Hare et al., 2016; Roberge and Beauchamp, 2006). These generally extract a profile or signature from the bullet lands (the surface between two bullet grooves). This profile serves as the features for each record. For pairs of profiles, various similarity metrics have been used, such as the correlation between aligned profiles, maximum number of consecutive matching striae and average Euclidean vertical distance between surface measurements of aligned profiles. For classification, both unsupervised threshold-based (e.g., Roberge and Beauchamp, 2006) and supervised methods (e.g., Hare et al., 2016) have been used. For example, Hare et al. (2016) used a total of seven similarity measures, and a random forest classifier, reporting no classification errors on a test data set.

2.5.1 Framework for Developing Forensic Matching Methods

In Figure 2.2 I introduce a framework that can be used to develop automatic forensic matching methods in a systematic manner. The framework in Figure 2.2 is a simplification of that in Figure 2.1. I also explicitly include a step of generating transitive closures as an optional post-processing step, implemented using hierarchical clustering. Evaluation is to be done using precision and recall. I will apply the framework in Figure 2.2 in Chapters 3 and 4. Comparing Figure 2.2 to Figure 2.1, Figure 2.2 skips the indexing step and the clerical review. The indexing step is not implemented because it is not straightforward in the forensic problems studied how this should be done; for example when the data are images it is unclear how to generate blocks in a blocking approach. Additionally, the current size of the data in the problems studied is not large enough to necessitate the indexing step. Indexing will be the subject of future work. Clerical review was not pursued due to the large manual effort required.

Generating features for each record and similarities for pairs of records is very data-specific, but at a high level the concepts are identical to that described in Section 2.4. The specifics for the two applications in this thesis will be described in detail in Chapters 3 and 4. The classification step is also as previously described. Chapter 3 uses an unsupervised threshold-based approach and Chapter 4 uses a supervised random forest approach.

If one would like to impose the constraint of transitive closures, an optional post-processing step can be performed after all pairs have been classified. The approach in Ventura et al. (2015) is adopted, where

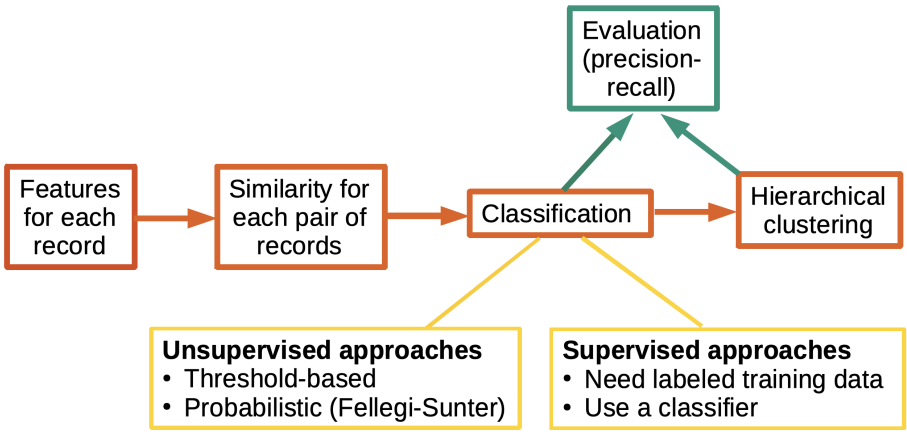


Figure 2.2: Simplified record linkage framework used to develop forensic matching methods.

given pairwise similarity predictions s, one uses $d = 1 - s$ as a distance measure, and performs hierarchical agglomerative clustering using various linkage methods.

Hierarchical agglomerative clustering is described as follows. Given items $1, \ldots, n$, dissimilarities d_{ij} between each pair i and j, and dissimilarities $d(G, H)$ between groups $G = \{i_1, i_2, \ldots i_r\}$ and $H = \{j_1, j_2, \ldots, j_s\}$, the algorithm starts with each node in a single group, and repeatedly merges groups such that $d(G, H)$ is within a threshold D. $d(G, H)$ is determined by the linkage method, and in this thesis I use single, complete, average (Everitt et al., 2001) and minimax linkage (Bien and Tibshirani, 2011). These linkage methods define $d(G, H)$, as follows.

Single linkage $d_{\text{single}}(G, H) = \min_{i \in G, j \in H} d_{ij}$. Informally, given some cutoff, each item needs to be matched with only one other item in the cluster, and missing links between pairs are filled in.

Complete linkage $d_{\text{complete}}(G, H) = \max_{i \in G, j \in H} d_{ij}$. That is, every item in the cluster needs to match with all other items in the cluster, and links between pairs are cut if this property is not satisfied.

Average linkage $d_{\text{average}}(G, H) = \frac{1}{|G||H|} \sum_{i \in G, j \in H} d_{ij}$. This is the average of all pairwise comparisons involving one item in group G and one in group H. It has no intuitive interpretation.

Minimax linkage $d_{\text{minimax}}(G, H) = \min_{i \in G \cup H} r(i, G \cup H)$, where $r(i, G) = \max_{j \in G} d_{ij}$, the radius of a group of nodes G around i. Informally, each item i belongs to a cluster whose center c satisfies $d_{ic} \leq D$; that is, all items in the cluster need to match to the cluster center.

Other approaches exist to ensure transitivity; the reason this is done as a post-processing step, is to allow the flexibility to do so or not, depending on the goals of the comparison. As described in Section 2.2 one reason one may wish to do so is to compare the results with examiner proficiency tests. This is elaborated on in Section 3.2 for cartridge cases, and other situations will be discussed in Section 4.1.3 for marketplace

accounts. Finally, for the evaluation step, I propose using precision and recall to evaluate the results both after classification and after clustering. Additionally, the area under the precision-recall curve can be reported as a numerical summary.

In terms of the forensic objectives outlined in Section 2.3, the framework in Figure 2.2 allows achievement of at least three of the four objectives outlined. To be specific, one can immediately get a ranking of similarity of pairs, a match or non-match conclusion and a disambiguated data set. Depending on the methodology used and data available, one can estimate random match probabilities or likelihood ratios. More details are in Chapters 3 and 4.

2.5.2 Other Specific Ways Record Linkage can Inform Forensic Matching

The main benefits of drawing the link between forensic matching problems and record linkage are in providing structure to the problem, and enabling the easy generation of new ideas. Thinking about forensic matching as an application of record linkage allows one to take advantage of tools that have been developed, without having to reinvent the wheel. When developing methods for new forensic matching problems, one can simply apply associated techniques. In Chapter 4 I describe how this can be done. Additionally, in domains such as DNA, fingerprints and firearms identification where there are existing well-developed bodies of literature, one can still benefit from thinking about these problems as record linkage problems, and some specific ideas are listed below.

1. **Indexing.** For forensic matching problems, as the sizes of databases grow, reducing computational complexity becomes increasingly important, particularly since in many applications data are often images which are high-dimensional objects. This is also related to the issue of scalability mentioned in Section 2.3. As explained in Section 2.5.1, it is not immediately apparent how to implement indexing methods in forensic problems, which has not been explored fully in this thesis. Indexing will be the subject of future work.

2. **Cutoffs for classification methods.** In existing work in the forensics literature, threshold-based methods have been proposed where cutoffs are selected arbitrarily. For example, in cartridge case comparisons, Song (2013) propose the Congruent Matching Cells method, in which smaller regions on the cartridge cases are compared. If there are six or more matching regions, the pair is determined to be a match. The original rationale was that in casework for bullets, examiners use six consecutive matching striae as a cutoff, so this form of "conventional wisdom" might transfer to cartridge cases.

   The literature in classification and record linkage offer much more principled ways to select and assess cutoffs. For example, in the Fellegi-Sunter framework for unsupervised record linkage (Fellegi and Sunter, 1969), cutoffs are selected to control error rates (false positives and false negatives). In supervised

classification problems receiver-operator curves or precision-recall curves are widely used to evaluate performance. I demonstrate how this can be done, in Chapters 3 and 4.

3. **Resolving intransitive links.** The issue of intransitive links is largely ignored in the forensics literature, even in situations where all pairwise comparisons are conducted on some test data set. Cases where it might be of interest to produce a linked data set are described in Section 3.2 and Section 4.1.3. The issue of intransitive links can easily be addressed by applying techniques from record linkage, as will be clear in Chapters 3 and 4.

4. **Generating additional investigative leads.** Related to the previous point, if one takes the investigative perspective to forensic matching, with a focus on producing leads through database searches, generating clusters within a database could be an easy way to generate additional leads.

5. **Deduplicating existing databases.** The previous suggestion hints at the benefits of managing existing databases. There might be some desire to deduplicate and/or summarize existing national forensic databases, such as fingerprint or cartridge cases. It is unclear if this is currently being done. Managing existing databases is beyond the scope of this thesis, but the record linkage literature can definitely lend itself towards this effort.

6. **Fellegi-Sunter and weight of evidence.** As described in Section 2.4, Fellegi and Sunter (1969)'s probabilistic record linkage framework relies on the likelihood ratio in favor of $(a, b) \in M$:

$$\frac{\mathbb{P}[\gamma_{ab} = g | (a,b) \in M]}{\mathbb{P}[\gamma_{ab} = g | (a,b) \in U]}, \tag{2.6}$$

   where g is the observed k-dimensional comparison vector.

   Notice that Equation 2.6 is essentially a specific formulation of the likelihood ratio in forensic problems, in Equation 2.1 (or Equation 2.5 for the DNA case, specifically). In forensics there has been no consensus as to how these likelihood ratios are to be estimated. In record linkage, methods to estimate likelihood ratios exist in the literature. It is unclear if these could also apply to feature or score-based likelihood ratios in Equations 2.2 and 2.3. Weight of evidence is not a focus of this thesis and these questions have not been explored in any detail, but could be the subject of future work.

The above is not an exhaustive list; record linkage is an active research area and there are many potential ways in which the forensic field might take advantage of recent advances.

In the next two chapters I develop methodology for automatically generating pairwise similarity scores in two specific forensic domains, and explore ideas for achieving the goals outlined in Section 2.3. I demonstrate the use of the framework introduced in Figure 2.2, and implement some of the suggestions noted in Section 2.5.2.

Chapter 3

Matching Firearms Evidence

In this chapter I apply the matching framework that was developed in Chapter 2, to the problem of matching firearms evidence. In particular, this chapter deals with image data, and as a result considerable effort was put into the steps of generating individual-level features and pairwise similarities. Section 3.1 introduces the problem and gives an overview of the literature. Section 3.2 describes the application of the matching framework in context of the firearms identification problem. Section 3.3 then describes the data used; there exist both 2D reflectance data, and 3D topographic data. Section 3.4 analyzes the former, and methodology as well as evaluation results are described in detail. The same is done for 3D in Section 3.5. Section 3.6 concludes.

3.1 Introduction and Background

Firing a gun leaves marks on the bottom surface of the cartridge case. In Figure 3.1 notice that the bottom surface of the cartridge is in contact with the breech block of the gun and the firing pin. During the firing process the cartridge is hit by the firing pin, which causes it to break up into two components, the bullet which goes out the barrel, and the cartridge case that is subsequently ejected from the side. This process leaves at least two kinds of marks: the firing pin impression caused by the firing pin hitting the cartridge, and breechface marks that are caused by the bottom surface of the cartridge pressing against the breech block of the gun. These can be seen clearly in Figure 3.2. This chapter focuses exclusively on breechface marks. These are impressed on the primer of the cartridge by the breech block, which is made of harder material than the primer. Any microscopic patterns or imperfections on the breech block may be reproduced in the breechface impression, and this is thought to individualize each gun (see e.g., Lightstone, 2010).

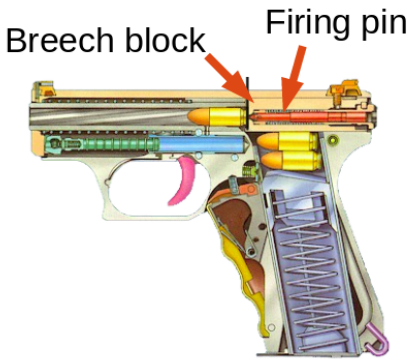


Figure 3.1: Gun that is about to be fired, showing the internal parts.

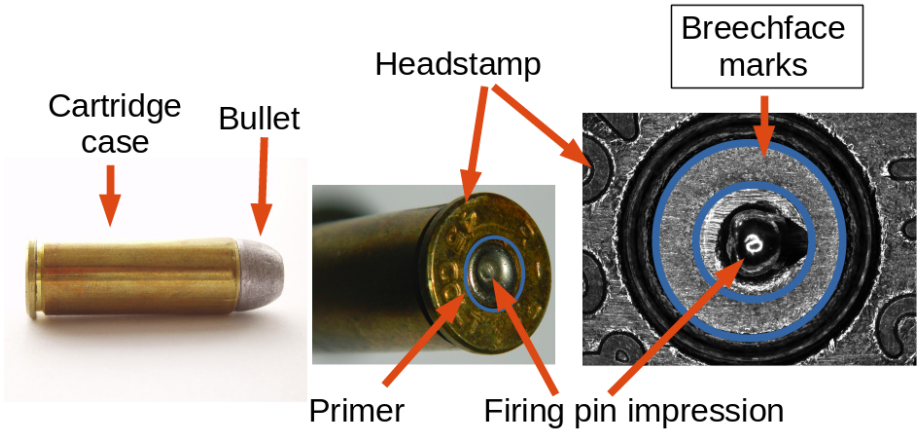


Figure 3.2: On the far left is a cartridge before firing. In the middle is the bottom surface of a cartridge case after firing. On the far right is an image of such a bottom surface, taken using a reflectance microscope.

The forensic challenge then is to determine if pairs of cartridge cases come from the same gun. This could aid investigations by enabling evidence from different crime scenes to be combined; alternatively investigators might attribute retrieved evidence to a source, potentially implicating a suspect.

Current technology allows for images of fired cartridge cases to be captured and stored for later comparison. Reflectance microscopes measure the amount of light reflected off the object's surface, producing 2D optical (grayscale) images. 3D profilers (such as confocal microscopes) measure surface contours directly, producing 3D topographies. Both these types of data are 2D matrices, with entries being either reflectance or depth values. In this chapter, methods to compare both 2D and 3D images are developed, to produce a similarity score which can be used to determine if pairs of cartridge cases come from the same gun. This methodology is both fully automatic and open source, and is tested on over 1,000 publicly available images.

In the remainder of this section, I introduce the current system and briefly review the literature. Section 3.2 goes over the general steps in the analysis, in relation to the framework introduced in Figure 2.2. The data are described in Section 3.3. Section 3.4 explains the methodology for analyzing 2D data, and includes evaluation results. Section 3.5 does the same for 3D data. Section 3.6 concludes.

3.1.1 Current Practice and PCAST

Law enforcement officers routinely collect guns and cartridge cases from crime scenes, because of their potential usefulness in investigations. The Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) maintains a national database called the National Integrated Ballistics Information Network (NIBIN). This contains around 3.3 million images of cartridge cases retrieved from crime scenes, and an additional 12.7 million test fires (as of May 2018) (Bureau of Alcohol, Tobacco, Firearms and Explosives, 2018). In current practice, cartridge cases may be entered into NIBIN through a computer-based platform called the Integrated Ballistics Identification System (IBIS), which was developed and is maintained by Ultra Electronics Forensic Technology Inc. (FTI), a company based in Montreal, Canada. This platform captures an image of the questioned cartridge case and runs a proprietary search, returning a list of top ranked potential matches from the database. Firearms examiners may then examine the list of associated images, to make a judgment about which potential matches warrant further investigation. If there are promising matches, the physical cartridge cases associated with these images are then located and examined under a comparison microscope. The firearms examiner decides if there are any matches, based on whether there is "sufficient agreement" between the marks (AFTE Criteria for Identification Committee, 1992), and may bring this evidence to court. It is the examiner that performs the evaluation or identification step, in which the evidence is attributed to a particular source. The NIBIN system is simply used as an investigative tool to generate leads; it is also up to the discretion of firearms examiners whether or not to enter retrieved cartridge cases into NIBIN. In addition, the NIBIN database is proprietary and unavailable for study.

There has been much public criticism in recent years about the current system. For example, PCAST (President's Council of Advisors on Science and Technology, 2016) questioned the scientific validity of firearms analysis. In particular, they expressed concern that there had been insufficient studies establishing the reliability of conclusions made by examiners, and the associated error rates had not been adequately determined. The database search is also problematic, because FTI's method is proprietary, and image acquisition is also affected by subjectivity. This was highlighted in De Kinder et al. (2004), where the frequency of manual corrections applied to the automatic imaging of the breechface was reported. For example, in their study breechface positioning (the outer circle in Figure 3.2) was manually corrected 18.5% of the time. The PCAST report suggested two directions for the path forward. The first is to "continue

to improve firearms analysis as a subjective method," by conducting further studies, and the second is to "convert firearms analysis from a subjective method to an objective method."

3.1.2 Overview of Literature

There have been efforts by various groups, both commercial and academic, in line with this second recommendation. As described, one can make a distinction between a database search (currently done using NIBIN) and an evaluation or identification (currently done by examiners). It is common in the literature to develop methods for one or the other. NIBIN's system for example is designed to be a database search, which is to be followed by a manual examination for identification. Other methods might focus on the identification step.

Ultra Electronics Forensic Technology Inc. (FTI), maintains the system used in NIBIN. The software extracts a numerical signature from each region of interest and does a database search using these signatures (Ultra Electronics Forensic Technology, 2018). Methods exist for both 2D and 3D data; the methodology is proprietary and few details are known. Cadre Forensics is a rival manufacturer of microscopes; they have similarly developed software for the identification problem for 3D images. For breechface marks, they extract geometric feature points that "a trained firearms examiner would identify," such as ridges, peaks, gouges, and concavities. These features are then aligned, as a set, to the features from another image. Logistic regression is used to produce a similarity score, with covariates such as number of matched features as input (Lilien, 2017).

Researchers at NIST have developed methodology for identification for both 2D and 3D topographies, using a method named Congruent Matching Cells (Song, 2013, 2015; Ott et al., 2017; Song et al., 2018). The idea is that the breechface area is split into a grid of cells and cells are aligned independently. Pre-specified criteria are used to determine if a pair of cells qualify as matching or not, and if the number of matching cells exceeds 6, the image pair is declared to be a match. Separately, Roth et al. (2015) focus on identification for 2D images. They treat each pixel as a feature and use a supervised classification method, gradient boosting, to classify pairs of images as being matched or not. Riva and Champod (2014) work on identification for 3D images. They compute six different similarity metrics, apply Principal Components Analysis, and then estimate likelihood ratios. Thumwarin (2008) use Fourier decomposition and unsupervised clustering for the identification problem using 2D images. Finally, Geradts et al. (2001) focus on the database search for 2D images, using wavelet transforms and location of prominent marks as features.

The shortcomings of currently available methods are that none of them are both fully automatic and open source. Automatic methods ensure that results are objective and reproducible, and open source software allows proper testing and validation. With respect to these attributes, Cadre Forensics's software is the most

promising – it is fully automatic but not open source. All other surveyed methods involve some manual input to select breechface marks. As far as I know, there are no open source methods available.

Large bodies of work exist in related areas, for example pre-processing and image alignment are well-studied in the image processing and computer vision literature. I do not go into the full details here. Related problems also exist in other domains that have similarities to the cartridge case problem, such as bullet matching (Hare et al., 2016) and iris recognition (Daugman, 2004).

3.2 Applying Matching Framework

I first describe the application of the matching framework introduced in Figure 2.2, in the context of the firearms identification problem. I do so generally and briefly in this section, including additional references to related work. The specific details of the methodology implemented for 2D and 3D data are in Sections 3.4 and 3.5 respectively.

3.2.1 Features for Each Record

Here an individual record is a 2D or 3D image. The goal of this step is to pre-process each image, and extract features that can be used for comparison. Pre-processing frequently involves selecting relevant areas of interest (in this case the breechface marks), highlighting certain features, and removing outliers. A comprehensive review of imaging systems and processing techniques is in Gerules et al. (2013). There seems to be little consensus on the appropriate order for performing the various pre-processing steps. Breechface marks are often first manually selected. In other work, commonly used automatic methods include the Canny edge detector, circular Hough transform, and active snake method (Li, 2003; Zhou et al., 2001; Brein, 2005; Tunali et al., 2009; Kamalakannan et al., 2011). Pixels that are judged to be outliers are removed and interpolated (Vorburger et al., 2007; Roth et al., 2015). Filters are used to smooth the image, and highlight individual as opposed to class characteristics (Roth et al., 2015; Vorburger et al., 2007; Riva and Champod, 2014). Other techniques include conversion to polar coordinates (Roth et al., 2015; Thumwarin, 2008), or wavelet transforms (Geradts et al., 2001).

Now, after pre-processing the pixel values can be used directly; some work also generates "signatures" or features that are used instead. Some examples are FTI's system which uses signatures produced using undisclosed methods, and Cadre Forensics which uses geometric feature points. Both are mentioned in the preceding section.

3.2.2 Similarity for Each Pair of Records

Given features for individual records, the next step is to generate meaningful similarity scores for pairwise comparisons. In this particular application, it is necessary to first align the two images or features to one another. Aligning images typically involves finding the best rotation and translation parameters (horizontal and vertical), where "best" means some metric is optimized, over all pixels or some subset of pixels. This optimization can be done either using a grid search or some other approximate means. The metric used might be correlation, mean-squared error, or something similar. Frequently, the optimized value of the metric is used as a similarity measure; this might be used on its own or combined with other similarity measures to produce a feature set for comparing two images.

Some examples follow. NIST aligns image cells using correlation (Song, 2015); Cadre Forensics aligns the geometric feature set extracted in the previous individual feature extraction step (Lilien, 2017); Roth et al. (2015) align rolled out versions of the images using correlation. The metric produced after maximizing correlation over rotations and translation is known in the literature as CCF_{max}, or maximum cross-correlation function. This is perhaps the most widely used measure of similarity (see e.g., Vorburger et al., 2007; Roth et al., 2015; Riva and Champod, 2014; Geradts et al., 2001). To be precise, the cross-correlation between two zero-mean images I_1 and I_2 (or more generally, 2-dimensional matrices), is defined as

$$CCF_{I_1,I_2}(k,l) = \frac{\sum_{i,j} I_1(i,j)I_2(i+k,j+l)}{\sqrt{\sum_{i,j} I_1(i,j)^2}\sqrt{\sum_{i,j} I_2(i,j)^2}}, \tag{3.1}$$

where (k,l) represents spatial lag (translation; with k and l being the vertical and horizontal lags respectively), i indexes the rows and j indexes the columns. The computation is repeated for different rotation angles, and CCF_{max} is then the maximum correlation, taking into account rotations and translations.

Some other pairwise features that have been used include metrics from the difference image ($D(i,j) = I_1(i,j) - I_2(i,j)$, for each i,j). Riva and Champod (2014) use the median of the squared differences, Vorburger et al. (2007) use the mean of squared differences, and Geradts et al. (2001) use the variance of the differences. Cadre Forensics uses metrics such as the number of matched features and the average difference in feature appearance (Lilien, 2017).

The problem of aligning images is also studied in the computer vision literature. For example, the Lucas-Kanade algorithm is commonly used for tracking objects in video (Baker and Matthews, 2004). Here instead of maximizing correlation, mean-squared error is minimized, typically using gradient descent. In fact, if the mean of the pixel values in the image is zero and the sum of squares is one, maximizing correlation is the same as minimizing mean-squared error. This can be seen in the following calculation.

Let $x \in \mathbb{R}^n$ be the first image, rolled out as a vector (for example if an image is 64×64 pixels, then the rolled out vector is $x \in \mathbb{R}^{64^2}$). Let y be the second image after rotation and translation (to be precise,

$y_i = f(x_i; \theta, k, l)$, where f is a transformation function with rotation parameter θ, vertical translation k and horizontal translation l). Then if $\bar{x} = 0$ and $\sum_{i=1}^{n} x_i^2 = 1$, $\bar{y} = 0$ and $\sum_{i=1}^{n} y_i^2 = 1$.

Now,

$$\begin{aligned}
\text{corr} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \\
&= \sum_{i=1}^{n} x_i y_i
\end{aligned}$$

and

$$\begin{aligned}
\text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i^2 - 2x_i y_i + x_i^2) \\
&= \frac{2}{n} - \frac{2}{n}\sum_{i=1}^{n} x_i y_i
\end{aligned}$$

Hence, maximizing correlation and minimizing mean-squared error are equivalent for estimating the translation and rotation parameters.

3.2.3 Classification

After the pairwise features have been extracted, the classification task is then standard in the sense that methods commonly used in record linkage or in classification problems can be applied directly. Both unsupervised and supervised methods can and have been used in the firearms identification literature. A match or non-match conclusion is not necessarily produced, instead some methods report a final similarity score, likelihood ratio or probability of matches. NIST and Riva and Champod (2014) use unsupervised, threshold-based methods. NIST uses a cutoff on a single similarity score (Song, 2013), while Riva and Champod (2014) combine multiple similarity measures into two principal components and compute a likelihood ratio. Thumwarin (2008) uses a clustering approach where for a new sample, only one match is found. This is the known sample that has the minimum Euclidean distance from the new sample. Roth et al. (2015) use a supervised boosting-based classification method.

3.2.4 Hierarchical Clustering

The final step in the framework is to generate transitive closures using hierarchical clustering. A final similarity score, produced in the previous step, can be converted into a distance measure in the same way that was described in Section 2.5.1, by taking $d = 1 - s$, and hierarchical clustering can then be used.

As far as I know, the issue of intransitivity has not been seriously considered in the forensic matching literature. As described in Section 2.2, one reason generating a linked data set may be desirable is to compare results with examiner proficiency tests. In these tests multiple samples of the reference firearm are provided as reference samples. Examiners need to determine if various questioned samples come from the same source as the reference samples. Results that examiners produce are necessarily transitive in terms of pairwise comparisons, so to compare these accurately with results from an automated method would require that the latter also have transitive links.

Additionally, reference data sets are often used in developing automated methods (more in Section 3.3). These typically contain multiple test fires from several firearms, and researchers perform all pairwise comparisons within a data set, and make evaluations on a pairwise basis. These pairwise predictions might contain intransitive links, meaning that overall model predictions are inconsistent. To perform a fairer assessment one might be interested in resolving these inconsistencies.

A final argument for imposing the constraint of transitivity is that it can lead to better classifier performance, as is seen in the remainder of this chapter.

It is possible that the above are not a concern depending on the goals of the analysis. For example, if it is only of interest to generate investigative leads by producing a ranking of similarities of pairs, then it is entirely reasonable to skip the step of generating transitive closures.

3.2.5 Evaluation

Given pairwise predictions and ground truth labels, standard evaluation metrics like misclassification rate, precision and recall can be directly estimated. In the firearms identification literature evaluation is not done in any standardized way. Most commonly, plots of scores from match and non-match distributions are displayed as histograms (Vorburger et al., 2007; Ott et al., 2017; Song et al., 2018; Roth et al., 2015; Riva and Champod, 2014; Lilien, 2017), and a comment is made about the separation between the distributions.

Vorburger et al. (2007) also estimate the overlapping area between distributions, and report the number of true matches in among the top 10 pairs in terms of similarity scores. Ott et al. (2017) additionally report the mean and standard deviation of the distributions of scores for matching and non-matching pairs. Song et al. (2018) further consider false positive and false negative rates, as well as false discovery and false omission rates.

Roth et al. (2015) use ROC curves in addition to plotting match and non-match score distributions. They also fit Gaussian distributions to the scores for matching and non-matching pairs, and estimate the overlapping area to quantify the expected error. Riva and Champod (2014) further analyze the distribution of likelihood ratios, and Lilien (2017) report false positive rates.

I propose standardizing the evaluation process by reporting precision and recall, as well as the area under the precision-recall graph as a numerical summary. This gives a better characterization of classifier performance, as opposed to visually comparing the overlap between match and non-match distributions. It is an improvement over estimating the overlap region, since the latter often requires assumptions to be made about the data. It is more informative than other summary measures like mean and standard deviation of the distributions.

3.2.6 Weight of Evidence*

Depending on the method used in the classification step, a likelihood ratio might be a by-product and can be reported as a measure of the weight of evidence. For example, methods like Fellegi-Sunter or the Naive Bayes classifier involve estimating a likelihood ratio.

In the firearms literature, score-based likelihood ratios (introduced in Chapter 2) have been proposed by NIST and Riva and Champod (2014). Both groups compute these as an additional step, distinct from classification. NIST uses a parametric distribution applicable to all guns (Song, 2015), while Riva and Champod (2014) use case-specific distributions for each questioned cartridge case and sample being compared to. To be specific, in the latter the distribution of scores for matches in the numerator comprises repeated test fires from the same gun, using the same ammunition. As for the denominator, the non-match distribution comprises comparisons of the questioned cartridge case with 1 test fire each from firearms of the same brand and similar models, using the same ammunition. Riva and Champod's approach is extremely specific, requiring collecting a new set of data for every questioned cartridge case and suspected firearm. The distributions are then modeled either parametrically or using kernel density estimation.

The non-match distribution in particular, in the denominator, can describe many different scenarios and may depend on the alternative hypothesis. For example, the at one end of the spectrum non-matches might come from guns of the same make and model, using the same ammunition. At the other end non-matches might come from guns of different makes, using different ammunition. One might expect that in the first case the non-matches are more similar and result in higher similarity scores on average. As far as I know, there is no consensus on how this issue should be tackled.

3.3 Data

Although a national database of firearms data exists (NIBIN), these data are not publicly available. Instead, data from NIST's Ballistics Toolmark Research Database (`https://tsapps.nist.gov/NRBTD`) are used in this thesis. This is an open-access research database of bullet and cartridge case toolmark data maintained by the National Institute of Standards and Technology (NIST). To my knowledge, this is the largest publicly

*This is not a specific step in applying the matching framework in Figure 2.2, but is discussed here for convenience.

available collection of reference data. The database contains images originating from studies conducted by various groups in the firearm and toolmark community. These cartridge cases were originally conducted for different purposes, for example the Laura Lightstone study investigated whether firearms examiners were able to differentiate cartridge cases from consecutively manufactured pistol slides (Lightstone, 2010). Majority of the data available are of cartridge cases that were sent to NIST for imaging, but the website also allows users to upload their own data in a standardized format.

The various data sets are summarized in Table 3.1, with each data set containing images from a single study. A total of 2,305 cartridge cases have been imaged (as of 3/4/2019). Among these data are sets involving consecutively manufactured pistol slides, a large number of firings (termed persistence studies because they investigate the persistence of marks), as well as a combination of different makes and models of guns and ammunition. Gun manufacturers include Glock, Hi-Point, Ruger, Sig Sauer, and Smith & Wesson, and ammunition brands include CCI, Federal, PMC, Remington, Speer, Wolf and Winchester. Metadata available for download provide additional information such as study details, the type of firing pin, material of the primer, etc.

As mentioned in Section 3.1, methods to obtain both 2D and 3D images exist, and this database contains both types of images. In recent years the National Institute of Standards and Technology (NIST) has advocated the use of 3D images because of their insensitivity to lighting conditions and traceability to the International System of Units (SI) unit of length (Song et al., 2012). The SI system comprises units of measurement built on base units that have precise standards of measurements. Specifically, base units are derived from invariant constants of nature (that can be observed and measured with great accuracy), and one physical artifact. This means that measurements of cartridge cases using any instrument can be compared to a known standard, and instruments can be calibrated to this known standard to assess and ensure precision. There remains interest in 2D images, however, because of the large amount of data that has been collected over the years in 2D, the cost of 3D equipment, as well as a lack of validated methods for 3D data.

For the 2D data, each casing was imaged using a Leica FS M reflectance microscope with ring light illumination. The objective was 2X, the resolution was 2.53 μm, and images are 1944 $\times$ 2592 pixel grayscale files in PNG format. Pixel values range from 0 to 255, where 0 represents black pixels and 255 represents white pixels. 3D data is primarily measured using a Nanofocus μSurf disc scanning confocal microscope. Various magnifications were used, for example an objective of 10X results in a lateral resolution of 3.125μm, and images that are around 1200 $\times$ 1200. Pixel values are depth values in μm (microns). More details are in Section 3.5.

In this chapter, data (2D and 3D) from all of the above studies are analyzed, except for FBI S&W M&P9.[†] For each study, all pairwise comparisons are performed. Since the metadata provided gives information on the

[†]The FBI S&W M&P9 data were added most recently to NIST's database and were not analyzed in time for the publication of this thesis.

Table 3.1: Summary of data available in NIST's Ballistics Toolmark Research Database on 3/4/2019. 2D and 3D data are available for all studies listed, with the exception of CTS, FBI S&W M&P9, and Todd Weller (Cadre), where only 3D images exist. Note that for Todd Weller 95 cartridge cases were imaged in 3D but only 50 were imaged in 2D.

Study	Cartridge cases	Firearm	Number of firearms	Slides per firearm	Cartridge	Test fires per firearm/slide
Cary Wong	91	Ruger P89	1	1	Winchester	91
CTS	74	Ruger P94DC	1	1	Federal	44
		Ruger P91DC	1	1	Federal	18
		S&W SW40VE	1	1	Federal	12
De Kinder	70	Sig Sauer P226	10	1	Remington	2
					CCI	1
					Wolf	1
					Winchester	1
					Speer	1
					Federal	1
Thomas Fadul	40	Ruger P95PR15	1	10	Federal	3-5
FBI S&W M&P9	1097	S&W M&P9	1	11	Luger	∼100
Hamby	30	Hi-Point C9	1	10	Remington	3
Kong	36	S&W 10-10	12	1	Fiocchi	3
Laura Lightstone	30	S&W 40VE	1	10	PMC	3
NIST Ballistics	144	Ruger P95D	4	1	Remington	3
Imaging					Winchester	3
Database					Speer	3
Evaluation					PMC	3
(NBIDE)		S&W 9VE	4	1	Remington	3
					Winchester	3
					Speer	3
					PMC	3
		Sig Sauer P226	4	1	Remington	3
					Winchester	3
					Speer	3
					PMC	3
FBI: Colt	90	Various Colts	45	1	Remington	2
FBI: Glock	90	Various Glocks	45	1	Remington	2
FBI: Ruger	100	Various Rugers	50	1	Remington	2
FBI: S&W	138	Various S&Ws	69	1	Remington	2
FBI: Sig Sauer	130	Various Sig Sauers	65	1	Remington	2
Todd Weller	95	Ruger P95DC	1	10	Winchester	5-9

source of each image, ground truth labels for whether pairs are from the same gun (a match or non-match) can be generated. The methodology is evaluated using these labels. In Section 3.4 I describe the analysis of 2D data, and in Section 3.5 3D data.

3.4 2D

3.4.1 Methodology

Individual Features: Pre-processing

As described in Section 3.2, it is necessary to pre-process the images before extracting individual features. In Tai and Eddy (2018) we use four preprocessing steps: 1. Automatically select breechface marks; 2. Level image; 3. Remove circular symmetry; 4. Remove outliers and filter. These steps build on methodology published in Vorburger et al. (2007) and implemented in Roth et al. (2015), and specific improvements are in steps 1 and 3.

Automatically select breechface marks We first find the primer region, and then remove the firing pin impression. Neither region is constrained to be circular, and this is especially important for the firing pin impression, which could have different shapes depending on the make and model of the gun. A rough schema for finding the primer region is given in Figure 3.3, and steps for removing the firing pin impression are in Figures 3.4 and 3.5. These make use of standard operations in image processing, such as filtering, dilation and erosion, and edge detection.

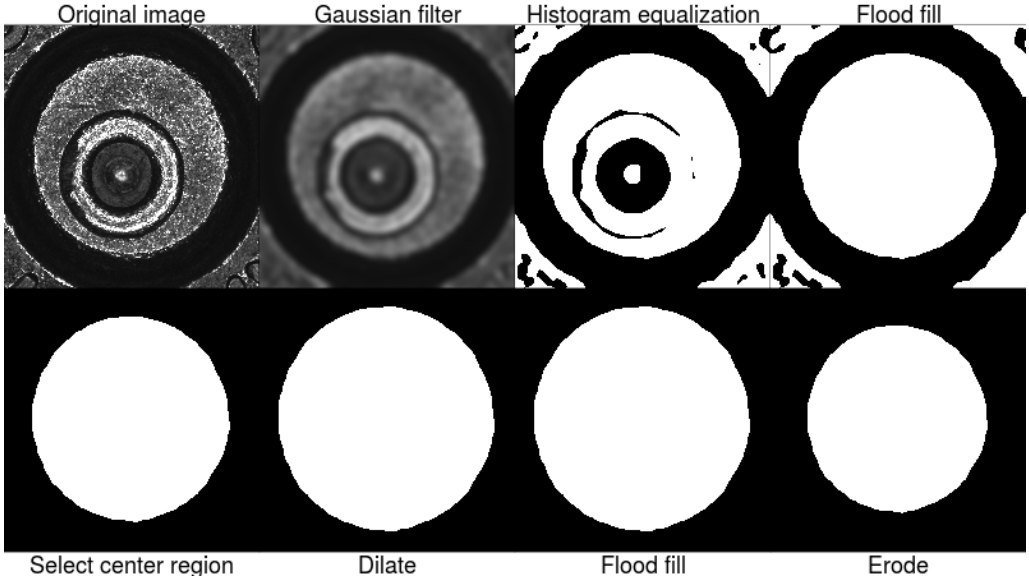


Figure 3.3: Series of steps to find the primer region for an example image. This image is from a Ruger gun, firing a PMC cartridge.

For finding the primer region in Figure 3.3, the image is first blurred using a Gaussian filter. Histogram equalization then converts the image into a binary image, where pixels below the median are set to black and pixels above are set to white. Holes are then filled, resulting in the firing pin impression being filled in. Next the center region is selected. In many images this is sufficient to select the primer area, but in some

images, the firing pin impression is so close to the edge of the primer that this center region selected ends up being irregularly shaped (not circular). This is because the flood filling step does not fill in parts of the firing pin impression that are connected to the outer black ring. To resolve such situations, the last three steps (dilating, filling and eroding) are added. Dilation increases the size of white areas, filling again fills in any remaining holes, and erosion does the opposite of dilation.

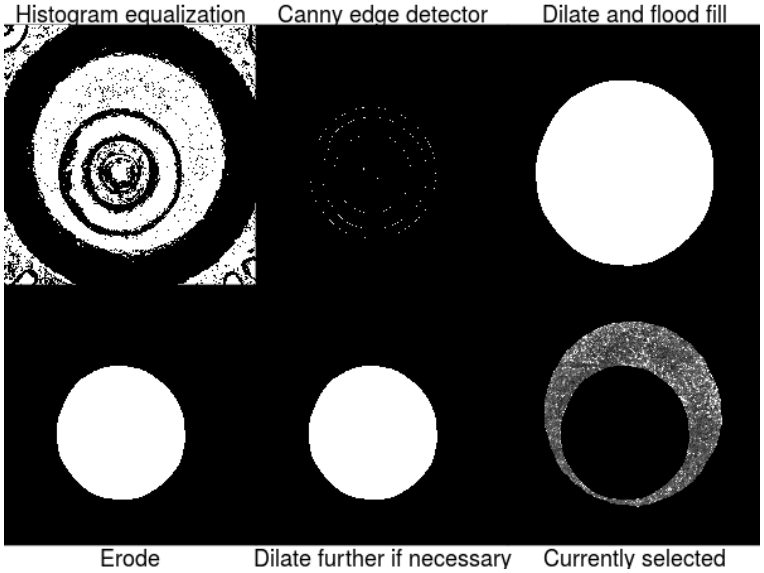


Figure 3.4: Series of steps to find the firing pin impression of the example image.

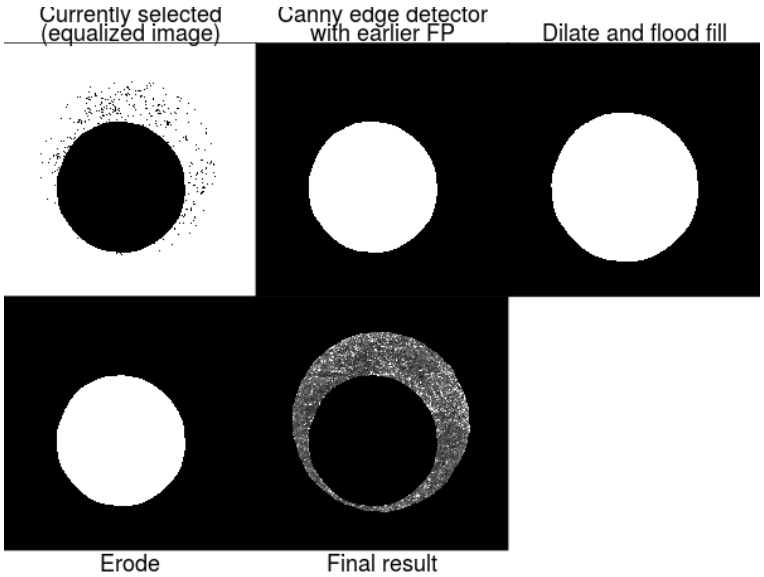


Figure 3.5: Here the edge detector is run a second time. In this particular example the entire firing pin impression has already been removed, so these steps do not produce any effect and the image remains unchanged.

In Figure 3.4 for finding the firing pin impression, the original image is first converted to a binary image (in the same manner described in the preceding paragraph), and the firing pin impression is then identified using a Canny edge detector (Canny, 1986). Briefly, image gradients are computed at each pixel location, and points with gradients of large magnitudes are designated to be edges. The Canny edge detector allows the specification of two thresholds, t_{high} and t_{low}, where $t \in [0, 1]$. Let g_{max} be the largest magnitude of the gradient among all the pixels in the image. Then any pixel with magnitude of gradient larger than $t_{high} * g_{max}$ is designated to be an edge, and pixels with magnitude of gradient larger than $t_{low} * g_{max}$ are designated to be edges as long as they are connected to an existing edge. The use of two thresholds enables weaker edges to be detected, if they are connected to a strong edge. This is useful because the entire border of the firing pin impression may not be prominently marked. Two passes of the edge detector are made; Figure 3.5 shows a second pass where the edge detector uses slightly different parameters to try to remove any remaining marks. This is necessary for some images where parts of the firing pin impression might not be as highly contrasted with the surrounding breechface impression, resulting in these less contrasted parts being missed the first time. This different set of parameters picks up such edges that are connected to the previously identified firing pin impression.

Level image This step adjusts for non-uniform lighting caused by the surface being tilted on a plane. This step is necessary because the base of the cartridge case may be tilted slightly on a plane, resulting in differences in brightness that are planar in nature. This could result in high overall similarity for two images with the same type of tilt with no similar individual features, which is undesirable since one is less concerned with overall patterns in brightness. To address this issue a plane is fit to capture these differences, and then the residuals are taken for further processing. This resulting image is free from planar differences in brightness. An example is in Figure 3.6.

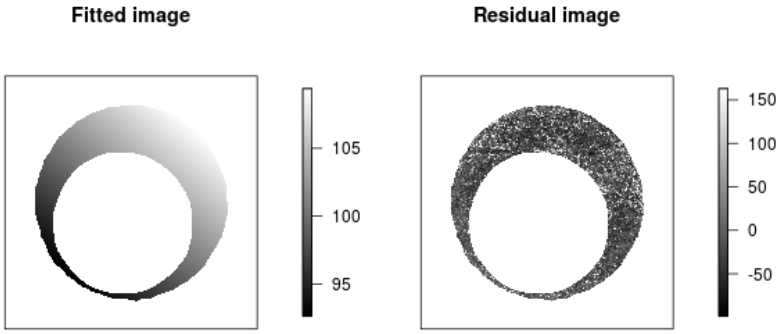


Figure 3.6: The fitted plane is on the left and the residuals are on the right. In this example the original image is slightly darker in the bottom left corner and brighter on the top right.

Remove circular symmetry This step is to adjust for non-uniform lighting caused by the surface having differences in depth that are circular in nature. Analogous to the previous step, the base of the cartridge case could have differences in depth that are circular in nature. This could arise if the base of the cartridge case slopes inwards towards the center, or has other circular differences in depth, incident to the manufacturing process. This might cause images to be darker in the center and brighter in the edges, for example. The step of removing circular symmetry corrects for these circular differences in depth, by fitting a model that captures the differences. The residuals are used for further processing.

The model that we fit is a linear combination of circularly symmetric basis functions (Zeifman, 2014). Consider a square matrix of dimension $m \times m$, where m is odd. The center entry is at $\left(\frac{m+1}{2}, \frac{m+1}{2}\right)$, and such a matrix is said to be circularly symmetric if entries located the same distance from the center entry take the same value. Since images are matrices of pixel values, pixels located the same distance from the center of a circularly symmetric image take the same value. Any image can be decomposed into a linear combination of the matrices in a circularly symmetric basis, plus residuals. The first few matrices in the basis are shown in Figure 3.7, where each panel represents one basis matrix. Each matrix takes the value 1 for pixels that are the same distance from the center, and zero otherwise. Bases are enumerated from center outwards.

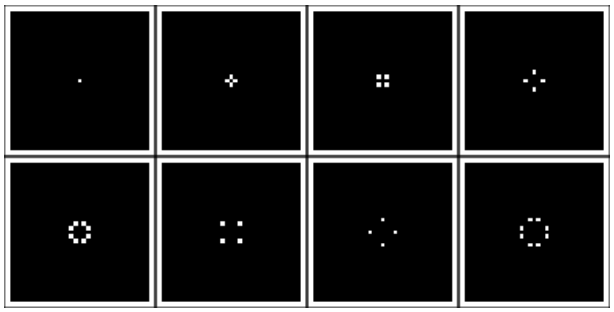


Figure 3.7: Illustration of the first eight matrices in a circularly symmetric basis. The pixels in white are the same distance from the center, and matrices are enumerated from center outwards.

To represent these matrices as basis functions, define functions f_k for each matrix k, taking the ij-coordinates as inputs, and returning the value 0 or 1 depending on whether the input pixel is the required distance, d_k, from the center, i.e., if it is white or black in the pictorial representation in Figure 3.7. An example of a basis function (the fifth one, enumerated from the center outwards) is in Equation 3.2. The decomposition of an image can then be represented as Equation 3.3, where K is the number of basis functions, f_k is the kth basis function, β_k is the basis function coefficient for f_k, and ϵ is the error.

$$f_5(i,j) = \begin{cases} 1 & \text{if } d(i,j) = d_5 = \sqrt{5} \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

$$Image(i,j) = \sum_{k=1}^{K} \beta_k f_k(i,j) + \epsilon(i,j) \tag{3.3}$$

The number of basis functions required, K, depends on the resolution of the image. For example, a 3×3 image has three possible distances from the center (including the corners), and can be represented using three basis functions. A 1769×1769 image (this is the size of the NIST images after cropping) requires 237,569 basis functions. The coefficient for each basis function is the mean of pixel values for pixels with the corresponding distance from the center.

In this analysis missing pixels are ignored. Since only the breechface marks are selected for analysis, treating the other regions as missing pixels, some fraction of the coefficients will not be computed, but the total number of coefficients is still very large. There are large local fluctuations in the model coefficients since many of them are estimated only using a small number of pixels (many basis functions only have four pixels, and many of the pixels are also missing due to the selection of the breechface area). In order to only capture a global effect, a loess regression (Cleveland, 1979) is fit to the basis coefficients to reduce the variance of the estimates.

Loess regression assumes that $y_i = f(x_i) + \epsilon_i$, where ϵ_i is the error. The function f is approximated at each point in the range of the data set, using a low-degree polynomial and weighted least squares. Weights are inversely proportional to the distance from the point of interest. The parameters required for estimating the function are the degree of the polynomial, the weighting function, and the proportion of all points to be included in the fit. Here a quadratic function is fit, and the proportion of points to be included in the fit is set to be $\alpha = 0.75$. The weight function (for local points that are included) from the point x_0 is computed using a tricubic kernel $w_i(x_0) = \left(1 - \frac{|x_i - x_0|}{h_0}^3\right)^3$, where h_0 is the $(\alpha * n)^{\text{th}}$ largest $|x_i - x_0|$. A large value of α was selected for a large degree of smoothing, as can be seen in Figure 3.8. With this large degree of smoothing, the loess regression is not particularly sensitive to the weight function or degree of polynomial, and we simply used the default values implemented in the `loess()` function in `R`.

Loess regression is a linear smoother, meaning that the fitted values are a linear combination of the original values (Buja et al., 1989). Precisely, $\hat{\mathbf{Y}} = \mathbf{SY}$, where $\mathbf{S}$ is an $n \times n$ matrix, called a smoother matrix. The smoother matrix is analogous to the hat matrix in multiple linear regression: $\hat{\mathbf{Y}} = \mathbf{HY}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T Y}$. The trace of the smoother matrix is the effective degrees of freedom of the loess regression, and using these specifications the effective degrees of freedom is less than 5 for each image. Results for an example image are in Figure 3.8.

The fitted model and the residuals for the same example image are in Figure 3.9. These residuals are free from both planar bias from the previous step, and circular symmetry.

Remove outliers and filter The last pre-processing step is outlier removal and filtering. This follows the methodology of Vorburger et al. (2007) and Roth et al. (2015). Outliers are removed and filled in so

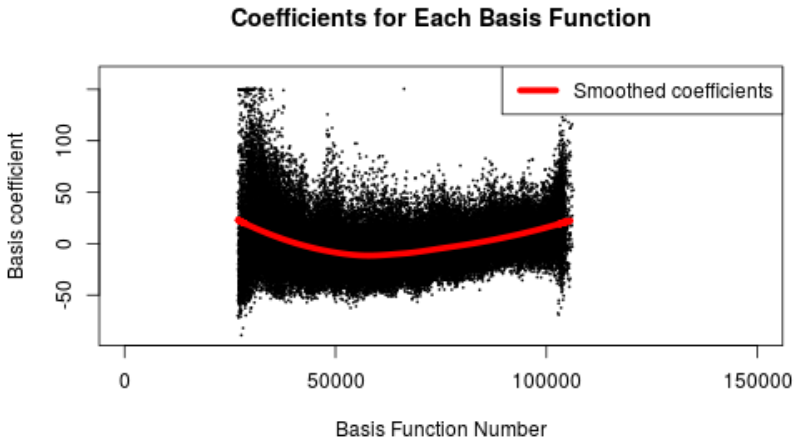


Figure 3.8: The set of circularly symmetric basis functions is fit to an example residual image after leveling.

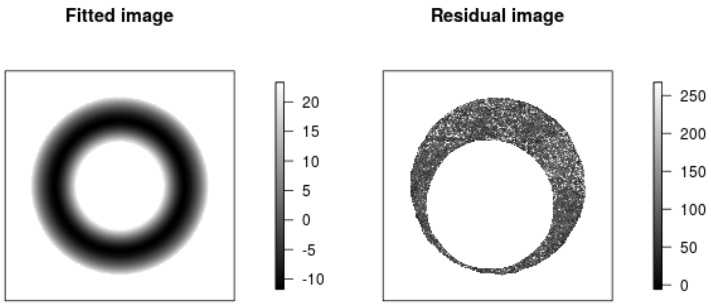


Figure 3.9: On the left is a plot of the smoothed fitted coefficients from Figure 3.8 as an image. Only pixels in the breechface area are plotted. Removing this fitted circularly symmetric model produces the residual image on the right.

as not to affect the similarity scores being computed, and filtering highlights certain features of the image. For outliers, we consider a local patch consisting of pixels in a 21×21 area. Outliers are then defined as pixels that have values larger than 3 standard deviations from the mean of the pixel values in its local patch. These outliers are replaced by values based on their neighbors (top and bottom, left and right), using the methodology in D'Errico (2004).

For filtering, a Gaussian filter is used, with short and long cutoffs approximately 15-110μm, meaning that patterns with wavelengths within this range are highlighted. The resulting image after all pre-processing is on the left in Figure 3.10.

This concludes the pre-processing, and the individual features used for each image are simply intensity values for all pixels after pre-processing.

Pairwise Features: Alignment and Feature Generation

For the alignment, a grid search is used to maximize the correlation between two images over translations and rotations. The maximized coefficient derived in this manner is an estimator of CCF_{max}, described in Section 3.2, or more precisely, $CCF^{max}_{I_1,I_2}$, where the two images being compared are I_1 and I_2. The details are in Algorithm 1.

Algorithm 1 Aligning image 2 to image 1

1: **procedure** ALIGN(I_1, I_2)
2: `thetas` $\leftarrow -175, -170, \ldots, -10, -7.5, -5, \ldots, 5, 7.5, 10, 15, \ldots, 180$
3: Scale I_1 (so $\bar{I}_1 = 0$ and $\sum_{ij} I_1(i,j)^2 = 1$) ▷ So that maximizing correlation ≡ minimizing MSE
4: **for** `theta` in `thetas` **do**
5: Rotate I_2 by `theta`
6: Scale rotated I_2
7: Compute cross-correlation, $CCF^{\theta}_{I_1,I_2}(k,l)$, as in Equation 3.1, for integer k and l
8: $CCF^{\theta}_{I_1,I_2} \leftarrow \max_{k,l} CCF_{I_1,I_2}(k,l)$
9: **end for**
10: $\theta' \leftarrow \arg\max_{\theta} CCF^{\theta}_{I_1,I_2}$ ▷ Neighborhood of best θ
11: **if** $\theta' \in [-10, 10]$ **then**
12: `fineThetas` $\leftarrow \theta' - 2, \theta' - 1.5, \ldots \theta' + 2$
13: **else**
14: `fineThetas` $\leftarrow \theta' - 4, \theta' - 3, \ldots \theta' + 4$
15: **end if**
16: **for** `theta` in `fineThetas` **do** ▷ Finer search in neighborhood of θ'
17: Lines 5-8
18: **end for**
19: $\theta^\star, k^\star_\theta, l^\star_\theta \leftarrow \arg\max_{\theta'} CCF^{\theta'}_{I_1,I_2}$
20: $CCF^{max}_{I_1,I_2} \leftarrow \max_{\theta'} CCF^{\theta'}_{I_1,I_2}$
21: **return** $CCF^{max}_{I_1,I_2}, \theta^\star, k^\star_\theta, l^\star_\theta$
22: **end procedure**

Let the correlation returned by Algorithm 1 be $\hat{c_{12}}$. Now, for a comparison of I_1 and I_2, one can either align I_2 to I_1 using ALIGN(I_1, I_2) to return $\hat{c_{12}}$, or I_1 to I_2 using ALIGN(I_2, I_1) to return $\hat{c_{21}}$. This might give slightly different results due to computational issues such as interpolation that is involved during a rotation. In Appendix A $\hat{c_{12}}$ is compared to $\hat{c_{21}}$ for each of the data sets, and one can note that the differences are minimal.

Now, the similarity score between two images is defined to be

$$\begin{aligned} \hat{s_{12}} &= \max(\hat{c_{12}}, \hat{c_{21}}) \\ &= \max(\text{ALIGN}(I_1, I_2), \text{ALIGN}(I_2, I_1)), \end{aligned} \tag{3.4}$$

and this single pairwise feature is used for classification.

To illustrate the results, Algorithm 1 is applied to a pair of example images in Figure 3.10, where both are from the same gun. A similarity score of .36 is obtained, with a rotation angle of $-15°$, meaning that the second image is rotated 15° counter-clockwise for best alignment. Plotting the two images with the second rotated, notice that the breechface marks are now lined up well. The difference between the two images can also be computed, and this is plotted on the far right of Figure 3.10.

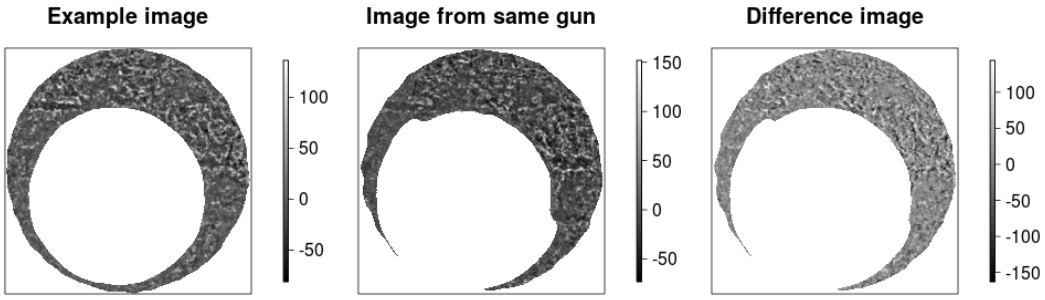


Figure 3.10: The example image from the previous figures is on the far left. In the center is a processed image from the same gun. On the far right is the difference image after alignment.

Classification, Hierarchical Clustering and Evaluation

The classification task is treated as an unsupervised problem. A threshold-based approach is used, meaning that cutoffs on $\hat{s}$ are selected, above which pairs are classified as matches. The effects of choosing various cutoffs are examined using precision-recall graphs.

Since pairs of images are evaluated independently, the classifier might produce matches that are intransitive. If it is of interest to resolve this (see discussion in Section 3.2), hierarchical agglomerative clustering is used. This is described in Section 2.5.1. The different linkage methods are implemented, and the changes in precision and recall after clustering are examined.

Random Match Probability/Likelihood Ratio

The steps above can give 1) a ranking of similarities of pairs, 2) a match or non-match conclusion, and 3) a disambiguated data set. This current methodology does not provide a measure for the evaluative context, and if it is of interest to compute a random match probability, the following additional step is proposed.

The random match probability, in context, is the probability of obtaining a higher correlation value ($\hat{s}$ defined in Equation 3.4) if the pair is a non-match, in other words by chance. In a similar manner to the computation of a score-based likelihood ratio, one needs access to some distribution of scores for known non-matches, and computing the right tail proportion would then give the required probability. Without knowing the theoretical distribution of CCF_{max} values, instead one can use a known database to derive an

empirical $\hat{s}$ distribution for non-matching pairs. Now, as described in Section 3.2, there is no consensus as to what distribution to use.

In Riva and Champod (2014), the assumption is that an unknown sample is being compared with a test fire from a specific, known firearm. The non-match distribution is then constructed from comparisons of test fires from this firearm with others of similar models, using the same type of ammunition. Here an alternative is proposed, which does not assume that either sample is known.

In Tai and Eddy (2018), we examined distributions of similarity scores for the NBIDE data set, for comparisons involving pairs of 1) the same type of gun (make and model) and same cartridge brand, 2) same type of gun, different cartridge brand, 3) different type of gun, same cartridge brand, and 4) different type of gun and cartridge. We found that the distribution for comparisons involving the same gun and cartridge brand was shifted slightly to the right, and formal statistical tests showed significant differences in distributions. Hence, we propose a conservative approach, where we construct the empirical non-match distribution only using comparisons involving the same type of gun and ammunition. Probabilities might then be interpreted as an upper bound: the probability of obtaining a larger similarity score by chance is less than p, where p is the probability computed in this manner.

3.4.2 Evaluation Results

Classifier Accuracy

Distributions of similarity scores $\hat{s}$ for matches and non-matches by data set are presented in Figure 3.11, and precision-recall graphs are in Figure 3.15. These are summarized in Table 3.2.

If the method achieves perfect discrimination between matches and non-matches, there would be a separation between the match and non-match distributions of the scores in Figure 3.11. Any vertical line between the two could then serve as a fixed cutoff, beyond which it can be concluded that the images are a match. From Table 3.2 notice that the consecutively manufactured studies (Lightstone, Weller, Fadul and Hamby) have good performance, with perfect or almost perfect separation between the match and non-match distributions. Kong's and De Kinder's data have very poor performance, with almost overlapping distributions. The FBI and NBIDE studies have a whole range of performances, from poor to middling to good. These are corroborated by the area under the curve (AUC) computed for the precision-recall graphs (black lines) in Figure 3.15. Data sets with almost perfect separation between the match and non-match histograms have areas under the curve of close to 1, the NBIDE set with good separation has an AUC of .84, those with some separation have scores around .4, and finally overlapping distributions have scores around .1 or lower.

Before discussing the data sets in more detail, I state several observations reported either in the literature or anecdotally by examiners, with regards to the ease of different types of comparisons. The first anecdotal

Table 3.2: Summary of 2D results. PR-AUC refers to area under the precision-recall curve. Results from the Cary Wong study are not listed because only one firearm is involved and all pairs are matches.

Study	Cartridge cases	Firearm	Cartridge	Histograms	PR-AUC
Lightstone	30	S&W 40VE	PMC	Perfect separation	.98
Weller	95	Ruger P95DC	Winchester	Almost perfect	.98
Fadul	40	Ruger P95PR15	Federal	Almost perfect	.97
Hamby	30	Hi-Point C9	Remington	Almost perfect	.98
Kong	36	S&W 10-10	Fiocchi	Overlapping	.14
De Kinder	70	Sig Sauer P226	Remington CCI Wolf Winchester Speer Federal	Overlapping	.11
NIST Ballistics Imaging Database Evaluation	144	Ruger P95D S&W 9VE Sig Sauer P226	Remington Winchester Speer PMC	Good separation	.84
FBI: Colt	90	Various Colts	Remington	Overlapping	.07
FBI: Glock	90	Various Glocks	Remington	Overlapping	.07
FBI: Ruger	100	Various Rugers	Remington	Some separation	.42
FBI: Sig Sauer	130	Various Sig Sauers	Remington	Overlapping	.04
FBI: S&W	138	Various S&Ws	Remington	Some separation	.43
Cary Wong	91	Ruger P89	Winchester		

piece of evidence is that Sig Sauers tend to mark poorly, in the sense that marks are not pronounced and are therefore difficult to compare accurately. Next, there are ammunition effects: examiners observed in Lightstone (2010) that breechface marks were most prominent on PMC, Federal and Winchester ammunition, while CCI and Remington ammunition marked very poorly (when a S&W 40VE firearm was used). Vorburger et al. (2007) noted for 3D data that matching scores are higher if the casings are the same ammunition brand, especially for PMC (for the NBIDE data set). George (2004) reached the same conclusion that comparisons involving the same gun using the same ammunition were more successful than those with different ammunition. Ott et al. (2017) further observed from the 3D CTS data, the two Ruger firearms had "inconsistent marks" between firings, in the sense that there were "large regions where poor contact was made with the breech face."

Now, I discuss the different data sets available. Four out of 13 of these involve consecutively manufactured pistol slides firing the same ammunition. The pistol slide is a component of the firearm that includes the breech block, and is hence responsible for the breechface marks. It had been suggested that pistol slides that were produced successively might have similar patterns and imperfections that persist through the manufacturing process (see e.g., Lightstone, 2010), so cartridge cases obtained using such slides might be more difficult to differentiate. Lightstone (2010) found that firearms examiners did not in fact have trouble comparing such

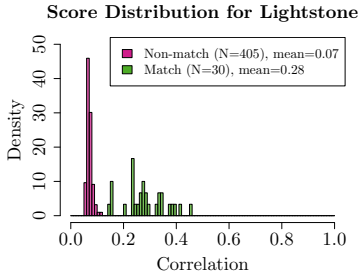
cartridge cases, because even though the molds that were used to produce the slides did have markings that carried over through the manufacturing process, the individual characteristics on each slide were sufficient for examiners to make correct identifications to the individual slide. Here one can come to the same conclusion: in Figures 3.11(a) to 3.11(d) the method produces higher similarity scores for all comparisons belonging to the same pistol slide, than those involving different slides. All the plots show a very good separation between the true match and non-match distributions. Referring to Table 3.1, the four studies with consecutively manufactured slides are Lightstone, Weller, Fadul and Hamby, and involve the following firearm-ammunition combinations respectively: S&W 40VE/PMC, Ruger P95DC/Winchester, Ruger P95PR15/Federal, Hi-Point C9/Remington. Hence it can be concluded that at least for these combinations, having pistol slides that are consecutively manufactured does not appear to pose a problem for automatic identification.

Now consider the Kong and De Kinder studies, which involve multiple copies of the same firearm. Kong uses 12 Smith & Wesson 10-10's, firing Fiocchi ammunition. One might expect results to mimic those that involve consecutively manufactured slides, but this does not turn out to be the case. A visual examination of the images reveals distinct circular patterns that are unusual and could have negatively impacted the results; some examples are shown in Figure 3.12. De Kinder on the other hand, uses 10 Sig Sauer P226's, firing six different brands of ammunition, and the results also show extremely poor separation between match and non-match scores.

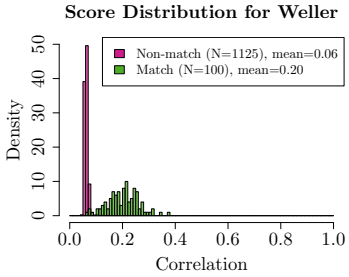
Next, results for the NBIDE and FBI studies are in Figures 3.11(g) to 3.11(l). These all involve either multiple ammunition or firearm types. The three FBI studies use various Colts, Glocks, Rugers, Sig Sauers and Smith & Wessons, and only Remington ammunition, while NBIDE uses different firearms with Remington, Winchester, Speer and PMC ammunition. For all these studies, there is some separation between the match and non-match scores, but results are clearly poorer than in the studies using consecutively manufactured slides, where only a single firearm and ammunition were used. Gun brands that have the best performance are Ruger and Smith & Wesson. The FBI Glock study deserves special mention. With the exception of this data set, the non-match distributions appear to be similar in all studies, having a low mean, relatively low variance and a longer right tail. Differences in performance are mainly driven by the match scores. As for the Glock study, looking at the processed images in Figure 3.13, note that the rectangular firing pin impressions unique to Glock pistols were removed poorly by the automatic procedure. This has effects on all the remaining pre-processing steps, and could cause extremely poor results. The example in Figure 3.13 is a non-match comparison and has a similarity score of .45, which as can be seen from Figure 3.11, is a high score.

Finally, the Cary Wong data set similarly involves a single type of firearm (Ruger P89) and ammunition (Winchester), but the gun is fired 2000 times. Cartridge cases 25-1000 at intervals of 25 test fires are imaged, as well as the first 10 and the 1001st test fire, giving a total of 91 images. The purpose of this study was to investigate if marks persisted over multiple fires, say, over the lifespan of a gun that is heavily used. The

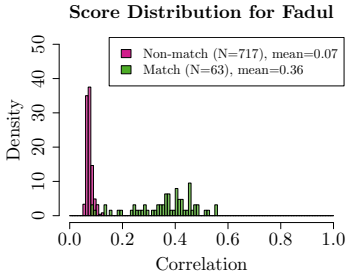
Figure 3.11: Distribution of $\hat{s}$ for matches and non-matches by study using 2D data.



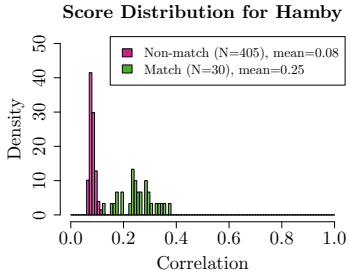
(a) Lightstone (consecutively manufactured)



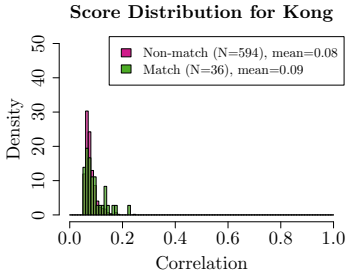
(b) Weller (consecutively manufactured)



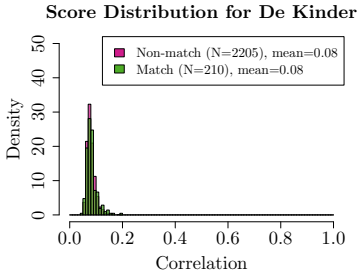
(c) Fadul (consecutively manufactured)



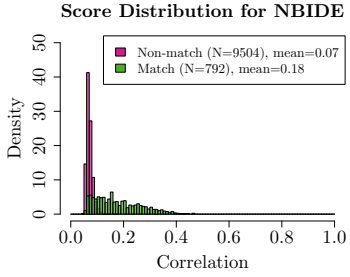
(d) Hamby (consecutively manufactured)



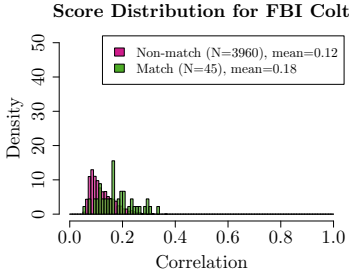
(e) Kong



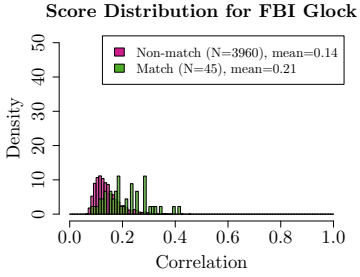
(f) De Kinder



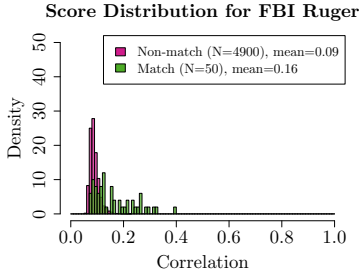
(g) NBIDE



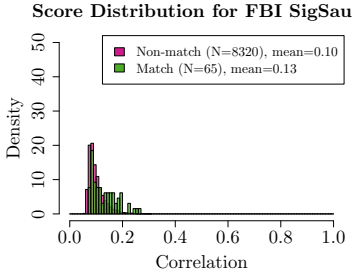
(h) FBI Colt



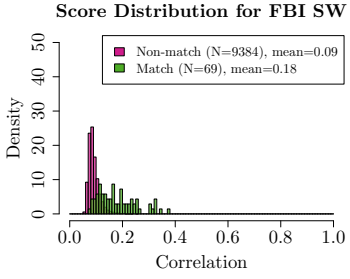
(i) FBI Glock



(j) FBI Ruger



(k) FBI Sig Sauer



(l) FBI Smith & Wesson

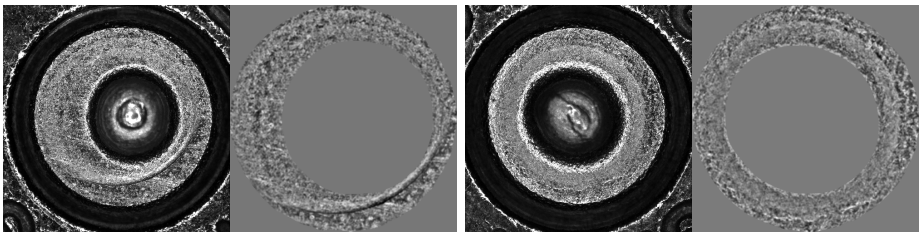


Figure 3.12: Example images from the Kong study, showing distinct circular patterns. The original images are on the left and the processed images on the right.

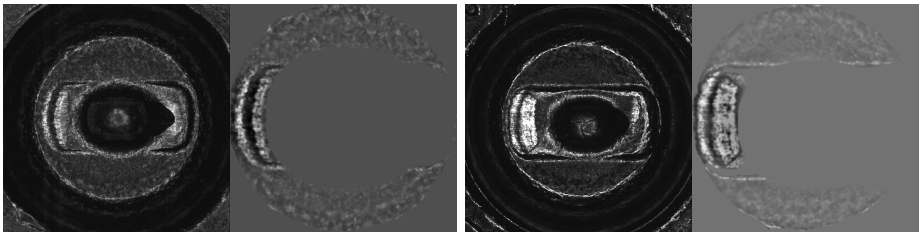
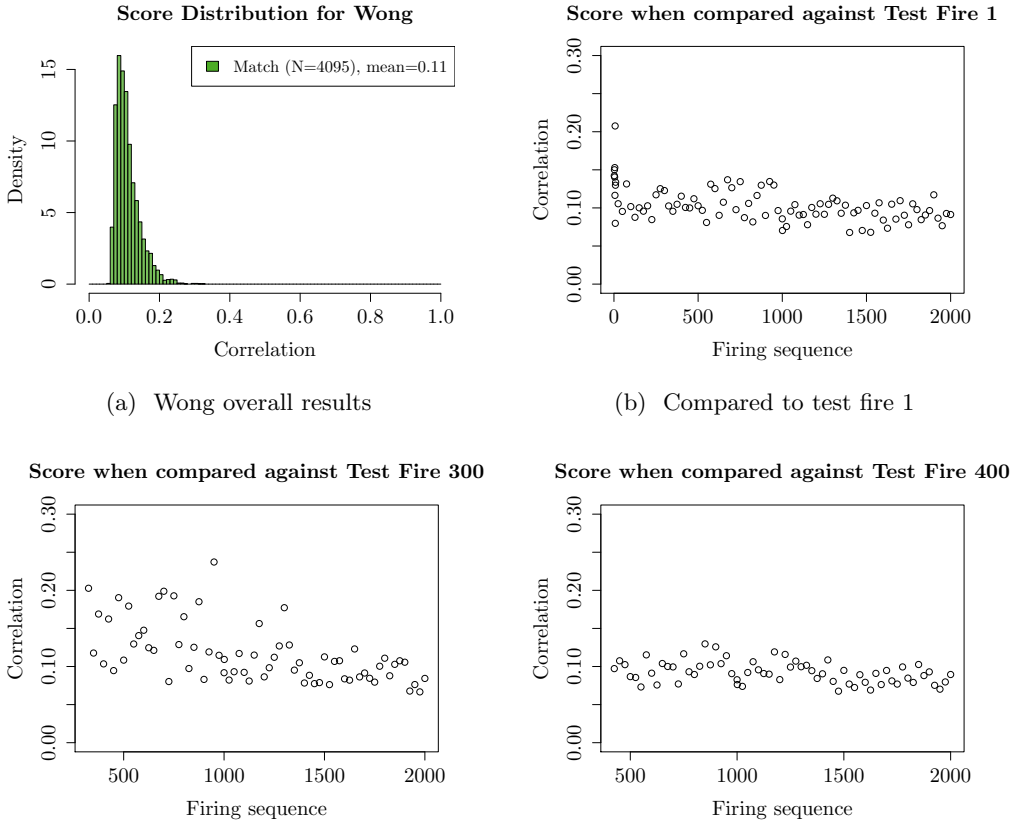


Figure 3.13: Examples of images from Glock pistols, with the firing pin impression removed very poorly. The original images are on the left and the processed images on the right. This is a non-matching pair and the comparison has a similarity score of .45.

results in show a range of $\hat{s}$ values, but most are small. This suggests that it is very unlikely that all cartridge cases can be identified to this particular firearm, in its entire lifespan. Similarity scores for comparisons against the first, 300th and 400th images are also plotted. These generally show a downward trend, although there is considerable variability between test fires that are close in firing sequence. These suggest that breechface marks do not necessarily persist for each test fire, over many test fires. Plots of scores against the 300th and 400th images further explore the possibility of a "settling-in" or "breaking-in" period: when tools are first used, some observe that their working surfaces change rapidly, before eventually stabilizing (National Research Council, Division on Engineering and Physical Sci, National Materials Advisory Board, 2008). In the context of firearms identification, this has been documented mainly for striations on bullets, and not breechface marks, but the latter might be affected by the same phenomenon. Here it is difficult to reach any definitive conclusions on whether this "breaking-in" period exists, because of the considerable shot-to-shot variability. For example, the 300th test fire has relatively high similarity scores with some test fires from 300-1000 in firing sequence, while the first does not, supporting the "breaking-in" hypothesis, whereas the same trends are not observed for the 400th firing sequence. It could simply be that the 300th test fire happened to leave prominent marks that were amenable to comparisons.

In summary, the results seem to corroborate the general observations of poor performance for Sig Sauer firearms, as well as when multiple ammunition brands are used. This was most obvious in the De Kinder data set, where performance was extremely poor. In terms of gun brands, Ruger and Smith & Wesson seem to have the best performance. Generally, pairs of images involving the same firearm and ammunition, fired in

Figure 3.14: Results for Cary Wong study in 2D.



(a) Wong overall results

(b) Compared to test fire 1

(c) Compared to test fire 300

(d) Compared to test fire 400

close succession have a better chance at producing high similarity scores, although this is not guaranteed due to shot-to-shot variability.

Clustering Accuracy

As discussed in Section 3.2, in some cases it is of interest to perform the final clustering step, and here precision-recall graphs are presented in Figure 3.15, which illustrate the results both before and after this step, for the different linkage methods.

In data sets with poor and almost perfect performance, the hierarchical clustering step does not affect AUC much; the largest differences are for NBIDE (.84 to .92), FBI Ruger (.42-.47) and FBI S&W (.43-.54). All of these are improvements. These are summarized in the final column in Table 3.3.

Table 3.3: Summary of 2D results with an additional column for the results after hierarchical clustering. The maximum area under the precision-recall graph is reported, among all linkage methods tested.

Study	Cartridge cases	Firearm	Cartridge	Histograms	PR-AUC	Max. PR-AUC after clustering
Lightstone	30	S&W 40VE	PMC	Perfect separation	.98	.97
Weller	95	Ruger P95DC	Winchester	Almost perfect	.98	.98
Fadul	40	Ruger P95PR15	Federal	Almost perfect	.97	.97
Hamby	30	Hi-Point C9	Remington	Almost perfect	.98	.97
Kong	36	S&W 10-10	Fiocchi	Overlapping	.14	.12
De Kinder	70	Sig Sauer P226	Remington CCI Wolf Winchester Speer Federal	Overlapping	.11	.10
NIST Ballistics Imaging Database Evaluation	144	Ruger P95D S&W 9VE Sig Sauer P226	Remington Winchester Speer PMC	Good separation	.84	.92
FBI: Colt	90	Various Colts	Remington	Overlapping	.07	.08
FBI: Glock	90	Various Glocks	Remington	Overlapping	.07	.11
FBI: Ruger	100	Various Rugers	Remington	Some separation	.42	.47
FBI: Sig Sauer	130	Various Sig Sauers	Remington	Overlapping	.04	.08
FBI: S&W	138	Various S&Ws	Remington	Some separation	.43	.54
Cary Wong	91	Ruger P89	Winchester			

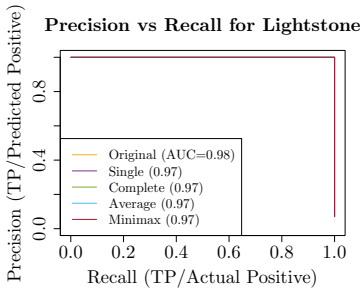
Final Clusters

It might be of interest to select a particular cutoff to produce a disambiguated data set where images are clustered based on whether they share a common source. Figure 3.15 shows the range of values of precision and recall that can be achieved using various cutoffs. The selection of a cutoff depends on the goals and desired tradeoff between false negatives and false positives. Here for the purposes of illustration I treat these equally, and choose a cutoff that maximizes precision and recall; this corresponds to the value at the bend of the precision-recall curve.

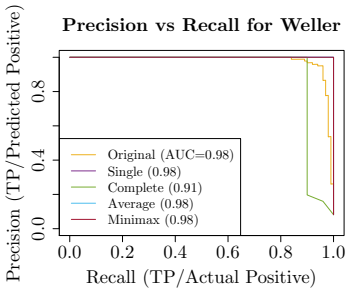
As for the linkage method, examining results from the different studies, it appears that average linkage has good performance across the board, so this is used to generate final clusters. Figure 3.16 plots the precision-recall curves using average linkage, colored by value of cutoff. A cutoff that works reasonably well for all data sets seems to be around .1.

Here I illustrate the results using the NBIDE data set; one can repeat this analysis for any data set. For the NBIDE data set a precision of almost 1 and recall of .9 is achieved at a cutoff of .08, which is used to generate final clusters. The NBIDE study consists of a total of 144 images from 12 different guns. I reproduce the relevant lines from Table 3.1 in Table 3.4. If the algorithm worked perfectly, one would expect to see 12

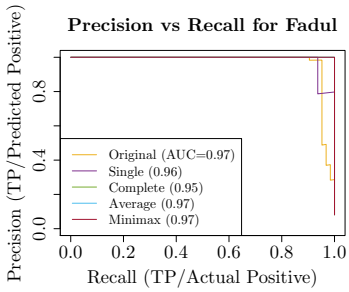
Figure 3.15: Precision-recall plots by study for 2D data. Area under the curve is reported in parentheses.



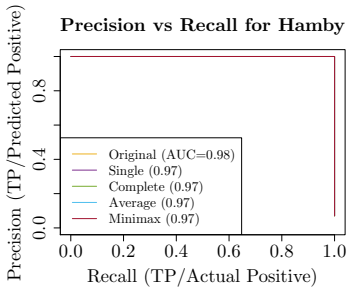
(a) Lightstone (consecutively manufactured)



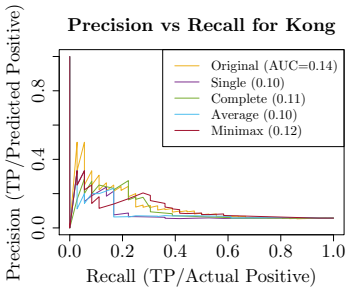
(b) Weller (consecutively manufactured)



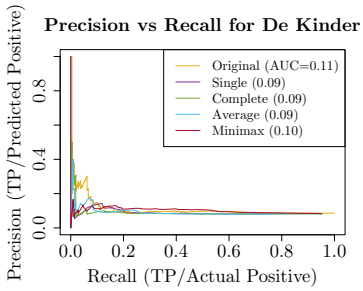
(c) Fadul (consecutively manufactured)



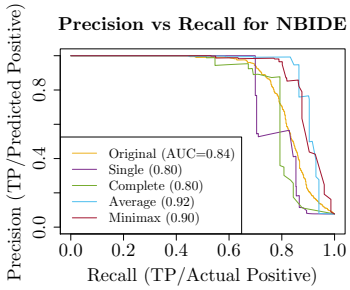
(d) Hamby (consecutively manufactured)



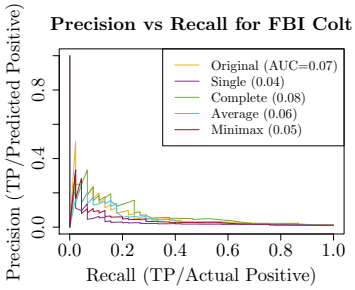
(e) Kong



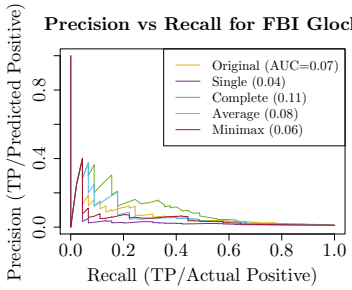
(f) DeKinder



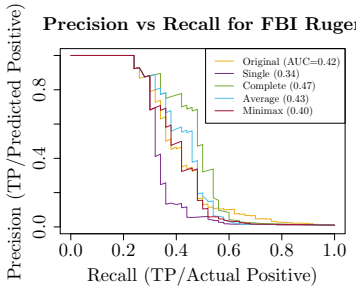
(g) NBIDE



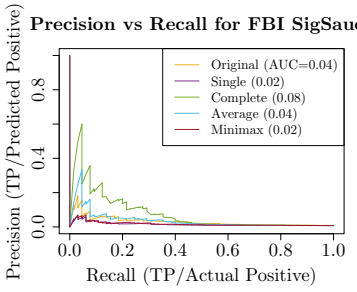
(h) FBI Colt



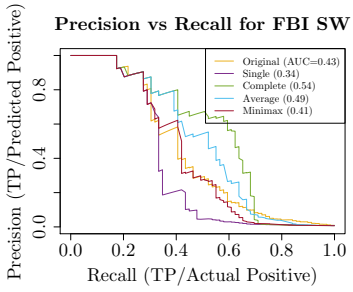
(i) FBI Glock



(j) FBI Ruger

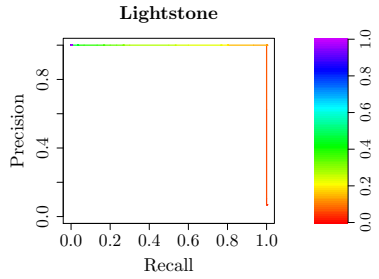


(k) FBI Sig Sauer

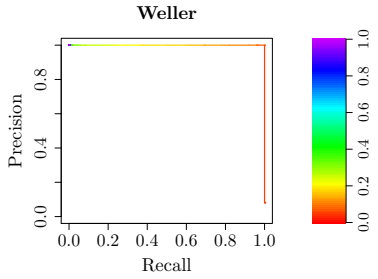


(l) FBI Smith & Wesson

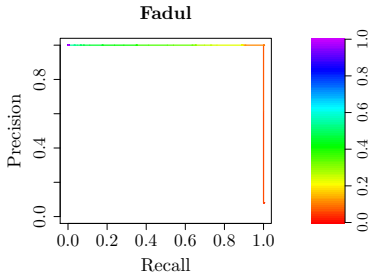
Figure 3.16: Precision-recall plots after hierarchical clustering using average linkage by study, colored by cutoff.



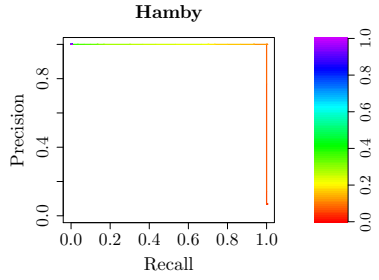
(a) Lightstone (consecutively manufactured)



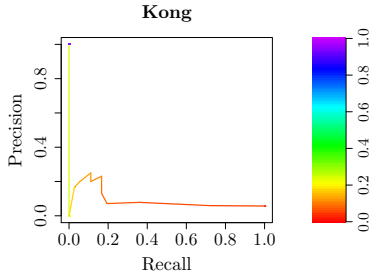
(b) Weller (consecutively manufactured)



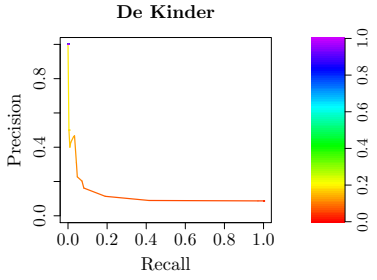
(c) Fadul (consecutively manufactured)



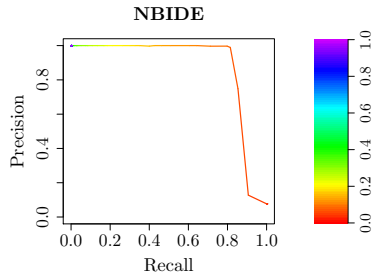
(d) Hamby (consecutively manufactured)



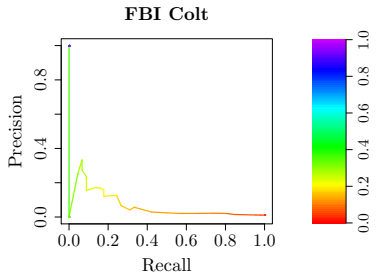
(e) Kong



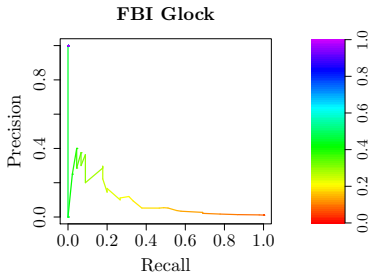
(f) De Kinder



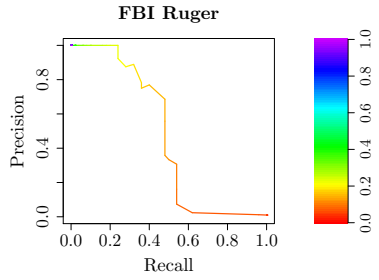
(g) NBIDE



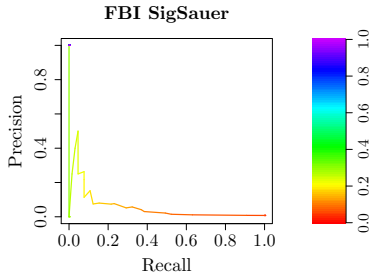
(h) FBI Colt



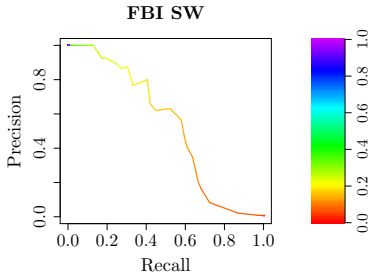
(i) FBI Glock



(j) FBI Ruger



(k) FBI Sig Sauer



(l) FBI Smith & Wesson

clusters of size 12. Instead, the cluster sizes are shown in Table 3.5. Now, comparing to ground truth, all 6 clusters of 12 are correct, and correspond to 3 Ruger firearms, and 2 Smith & Wesson, and 1 Sig Sauer. The remaining 6 firearms comprise of combining two or more of the smaller clusters, with one exception: one cluster of size 5 mistakenly contains cartridge cases from three different firearms.

Table 3.4: Information about NBIDE data set, reproduced from Table 3.1

Study	Images	Firearm	Number of firearms	Slides per firearm	Cartridge	Test fires per firearm/slide
NIST Ballistics Imaging Database Evaluation (NBIDE)	144	Ruger P95D	4	1	Remington	3
					Winchester	3
					Speer	3
					PMC	3
		S&W 9VE	4	1	Remington	3
					Winchester	3
					Speer	3
					PMC	3
		Sig Sauer P226	4	1	Remington	3
					Winchester	3
					Speer	3
					PMC	3

Table 3.5: Cluster sizes for the NBIDE data set, after hierarchical clustering using average linkage with a cutoff of .08. Perfect results would be 12 clusters of size 12.

Cluster Size	1	2	3	5	9	11	12
Count	4	3	2	1	2	3	6

In terms of performance by gun brand, Rugers perform the best, followed by Smith & Wesson and then Sig Sauer. This is in line with anecdotal evidence that Sig Sauers mark poorly, and also matches performances by study: comparing the FBI Ruger, Smith & Wesson and Sig Sauer data sets (see e.g., Figure 3.11), FBI Ruger performs the best, followed by Smith & Wesson and then Sig Sauer. This suggests that there may be firearm brand effects, with some firearms performing better than others.

3.4.3 R Package

I have developed an R package, `cartridges`, available on GitHub, to analyze 2D images in png format. The primary functions of the package are `allPreprocess()` and `calculateCCFmaxSearch()`. The former performs all pre-processing steps, while the latter does both alignment and computation of a similarity score for a pair of images. If analyzing a data set in the same manner as in the preceding sections, `linksAnalysis()` can be used to perform hierarchical clustering and `getClust()` to get final clusters. If desired, the random match probability can be computed using `computeProb()`, but the distribution of non-matches will have to be supplied by the user.

3.4.4 Comparison with Other Work

To my knowledge only the Fadul and Weller data sets have been analyzed in detail in 2D by others (Tong et al. (2014) and Roth et al. (2015) respectively). For both these data sets, the methods proposed in other papers achieved perfect separation of true match and non-match scores, whereas we have close to perfect separation. (Since these studies selected only one data set each from the NIST database, it is unclear if these perfect results would hold if all the data sets were used.)

In addition, Vorburger et al. (2007) reports results returned through the NIBIN system in 2007 for the NBIDE and De Kinder data sets, in the form of the number of matches returned in the top 10 pairs returned. In Tai and Eddy (2018) we compared the results for NBIDE using this metric, showing that our method produced superior results. This was a non-standard metric. It was not generated for the De Kinder set, but could be the subject of future work.

3.4.5 Comparison with Published Methodology

In Tai and Eddy (2018) we explored the effects of the automatic selection of breechface marks, and the removal of circular symmetry, for the NBIDE data set. In particular, we compared these to a baseline which used four pre-processing steps: manual selection of breechface marks, leveling, outlier removal, and filtering, and found that 1) removing circular symmetry has little impact on the similarity scores for true matches, but reduces scores for non-matches, 2) automatic selection of breechface marks generally increases both the mean and variance of scores, but has no obvious effect on the separation between scores for matches and non-matches. These are illustrated in Figure 3.17.

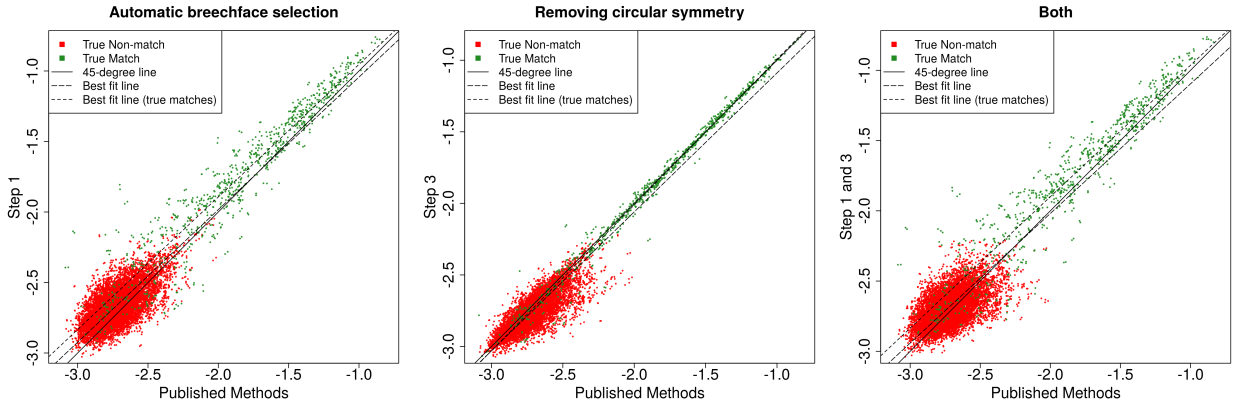


Figure 3.17: Plots reproduced from Tai and Eddy (2018), illustrating the effects of the addition of the automatic selection of breechface marks (Step 1), and the removal of circular symmetry (Step 3). The logarithmic scale is used on both axes to highlight the differences in the lower values.

In particular, we do a closer analysis of the effects of circular symmetry. Circular symmetry was added to pairs of images, causing them to be darker in the center and getting brighter toward the edges. This

dramatically increases correlation scores. An example is in Figure 3.18. In actual pairs of images, some have similar patterns, though much less drastic than in the synthetic example. Removing these reduces similarity scores for non-matches. For the true matches, the individual marks are similar enough that even after any circular symmetry is removed, the images remain highly correlated.

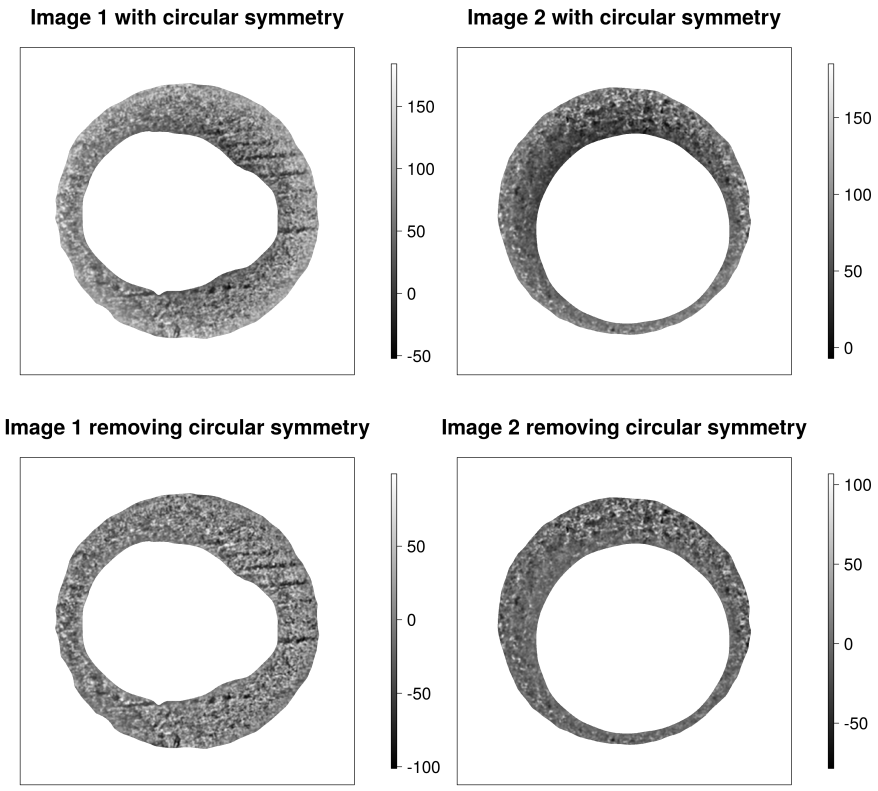


Figure 3.18: Plot reproduced from Tai and Eddy (2018), illustrating the effects of the removal of circular symmetry. The images on the left are compared respectively with those on the right. The first row has the same circular symmetry added to both images, producing a similarity score of .72. The second row removes this circular symmetry and the score drops to just .04, because the individual marks are not very similar.

3.5 3D Topographies

In this section I attempt to replicate the analysis using 3D topographic data. 2D and 3D data differ in that the two measure different properties of the same physical object: 2D images record reflectance while 3D topographies record surface contours directly. 2D images have been used in NIBIN systems for almost two decades, while 3D topographies are measured using newer microscopes. While 2D images are a matrix of intensity values, 3D topographies are a matrix of depth values, stored in an x3p (XML 3-D Surface Profile) file (see `https://sourceforge.net/p/open-gps/mwiki/X3p/`). This is a container format, containing a matrix

of depth values, together with associated metadata, including the instrument and lateral resolution used to capture the topography.[‡]

2D and 3D data differ in some other characteristics. The 2D data in NIST's database all have the same dimension and resolution, while the 3D data are of varying dimensions and resolutions. The dimensions are 600×600, 1200×1200, or 2500×2500, corresponding to lateral resolutions of 6.25 microns, 3.12 microns, and 1.56 microns respectively. Lenses with different magnifications were used for different measurements, and some images were downsampled before being uploaded into the database. (For example the 1.56 micron resolution images are taken using a 20X objective with no downsampling, while the 3.12 micron resolution images could be a 20X objective downsampled by half, or using a 10X objective.) Another difference between 2D and 3D data is the presence of dropouts (missing values). When there is a large change in depth, for example a steep slope at the walls of the firing pin impression, the measuring instrument is unable to capture the depth values accurately and they are recorded as missing or NA values. These points are important to note as they affect the subsequent pre-processing.

3.5.1 Differences in Methodology

As a reminder, for 2D data I introduced four pre-processing steps in Section 3.4.1: 1) Automatically select breechface marks; 2) Level image; 3) Remove circular symmetry; 4) Remove outliers and filter. There are a few differences for 3D data. First, to standardize the different resolutions all images are resized to the lowest resolution of $6.25\mu m$, and it was at this resolution that all pre-processing was done (for 2D this was $2.53\mu m$). Second, a different procedure for automatic selection of breechface marks is used, and this is described shortly. Third, this selection of breechface marks takes care of outliers, so Step 4 now involves only the Gaussian filter. The parameters of the Gaussian filter are slightly different; in the 2D case we used a filter with cutoffs approximately 15-110μm, while here the cutoffs are 20-150μm.[§] The comparison procedure is the same as for 2D, however the resolution used for comparisons is $25\mu m$, compared to the $10\mu m$ for 2D images. This results in quicker comparisons, and Section 3.5.2 shows that accuracy is not sacrificed. The only major difference in methodology is in the selection of breechface marks, which is described in more detail as follows.

For 2D data we used edge detection and image processing operations to select the breechface marks. In 3D, one can instead make use of physical characteristics to achieve better performance. The breechface region is relatively flat and has low depth, while the firing pin impression, being an indentation on the primer surface, is typically much deeper. The solution then is to fit a plane through the breechface region, ignoring the firing pin impression, and to select only points lying on or close to the plane.

[‡]This is the standard file format that has been agreed upon by the Open Forensic Metrology Consortium (OpenFMC), "a group of academic, industry, and government Firearm Forensics researchers whose aim is to establish file formats, means of data exchange, and best practices for researchers using metrology in the forensic sciences."

[§]Empirically these produced the best results among some test examples; as far as I know there is no there is no established guidance on what wavelengths are relevant for the comparison. Some cutoffs that have been used in other work are 2.5-250μm, 37-150μm (Vorburger et al., 2007), and 16-250μm (Song et al., 2018).

To achieve this, I use an algorithm called RANdom SAmple Consensus (RANSAC) (Fischler and Bolles, 1981), which is designed to fit models in the presence of outliers. The procedure used is outlined in Algorithm 2. Briefly, planes are repeatedly fit by sub-sampling points, and the plane with the largest number of inliers is chosen, where inliers are defined as points within a selected threshold from the fitted plane. Points within the selected threshold are determined to be part of the breechface area.

Algorithm 2 RANSAC to find best-fitting plane through Image I

```
procedure bestPlane(I)
    for i in 1:iter do
        Sample 3 points from I
        Fit plane to sampled points
        Count the number of inliers within preset threshold
    end for
    Select model with largest number of inliers
    Re-fit model using only inliers from selected model
    return fitted plane, selected inliers
end procedure
```

There are three parameters to be selected: s, the number of points sampled, δ, the threshold above which points are considered to be outliers, and N, the number of iterations `iter` the algorithm runs for (in other words the number of samples). s is typically the number of points required to fit the model; in this case 3 points are required to fit a plane. $\delta = 10\mu m$ is selected: Figure 3.19 shows a typical breechface image (without a firing pin impression captured) and the associated histogram of depth values – one can see that most of the depths are within $10\mu m$ of each other, making this a reasonable choice.

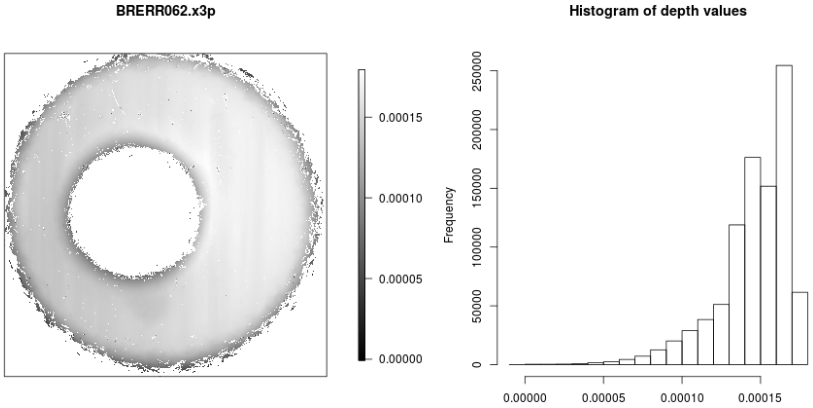


Figure 3.19: Typical 3D breechface image (without a firing pin impression captured) and associated histogram of depth values in microns.

As for N, it should be set to be large enough such that the probability of at least one random sample being free from outliers is at least p. Now, let X be the random variable denoting the number of samples, out

of N, having no outliers. Let e be the proportion of outliers in the data, and s be the number of sampled points. Then $X \sim \text{Binom}(N, (1-e)^s)$, and the goal is to have $\mathbb{P}[X \geq 1] \geq p$.

Now, $\mathbb{P}[X \geq 1] = 1 - \mathbb{P}[X = 0] = 1 - [1 - (1-e)^s]^N$, so the goal becomes

$$1 - [1 - (1-e)^s]^N \geq p.$$

Taking logs,

$$\log(1-p) \geq N \log[1 - (1-e)^s]$$
$$N \geq \frac{\log(1-p)}{\log[1 - (1-e)^s]}$$

Consider $p = .99$, $s = 3$ and $e = .6$. This last argument says that 60% of points are outliers. This is set to be a relatively large proportion, since some images capture parts of the firing pin impressions, and all these points would be considered to be outliers. Nevertheless, $e = .6$ is still likely to be an overestimate. This gives $N \geq 70$. In the actual implementation I use 75 iterations. Two examples of the resulting breechface impressions being selected are in Figure 3.20. Both circular and rectangular firing pin impressions can be removed reasonably well using Algorithm 2.

In Section 3.4 it was noted that performance on the FBI Glock study was poor, and that the automatic selection of breechface marks had poor performance particularly on Glock pistols, due to their unique rectangular shape. Figure 3.13 showed an example of a pair of Glock cartridge cases with very high similarity scores and poorly removed firing pin impressions. The same pair in 3D before and after all pre-processing is in Figure 3.21, and performance is much better.

Similar to 2D, I have a developed an R package with analogous functions, in `cartridges3D`.

3.5.2 Evaluation

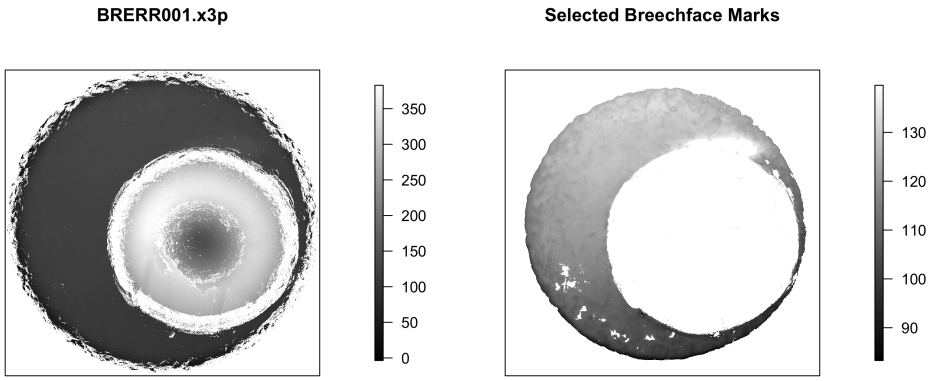
Classifier and Clustering Accuracy

Results in terms of match and non-match distributions are presented in a fashion similar to Section 3.4.2, in Figures 3.22 and 3.23. Precision-recall graphs are in Figure 3.25, with the yellow lines representing results after the classification step. In addition, the individual correspondences of 2D and 3D scores for each pairwise comparison is in Figure 3.24. The results are summarized in Table 3.6.

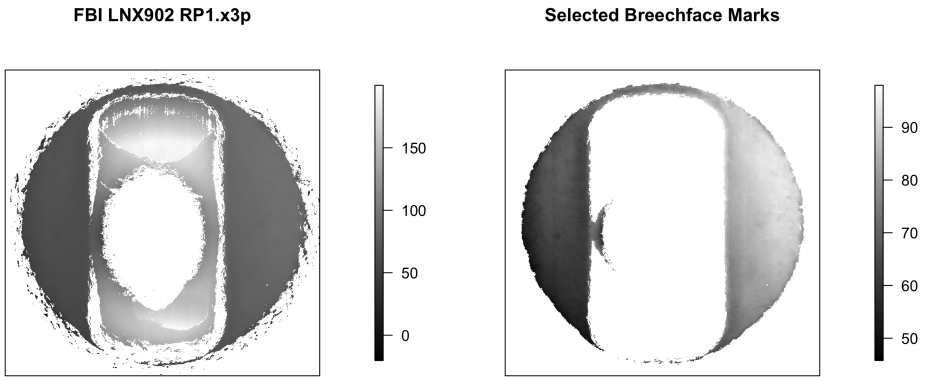
From Table 3.6, 3D results are better than in 2D. ln particular there are marked improvements for all of the FBI studies: Colt, Glock, Ruger, Sig Sauer and S&W. There are also some improvements in NBIDE results. Performance on Kong and De Kinder data remain very poor.

From Figure 3.23 the Cary Wong persistence results exhibit a similar downward trend as the number of test fires increase, but scores generally are much higher. It still remains unlikely that *all* cartridge cases can be identified to the firearm in its lifespan, but it is likely that many of them can, and more compared to 2D.

Figure 3.20: Examples of selected breechface areas after applying Algorithm 2. The scale is in microns.



(a) Example image with circular firing pin impression



(b) Example image with rectangular firing pin impression

Similarly to the 2D case, the clustering step results in better AUCs for data sets that are not at the extremes in terms of performance (either very poor or close to perfect): the larger differences are for Hamby (.87 to .93), NBIDE (.86 to .94), CTS (.87 to .94), FBI Colt (.6 to .64). Overall, very good performance is achieved in 9 of the 13 data sets, where AUCs are over .9 using at least one linkage method. One might hypothesize as to why clustering improves performances. Minimax linkage, for example, selects a cluster center for each cluster, and members of the cluster need to be similar to this center but not necessarily to other members of the cluster. It is plausible that the cluster center is a particularly well-marked cartridge case, while other cluster members might have areas that are less well-marked. If this is true, clustering can achieve the same effect as grid-based methods like Congruent Matching Cells (see Section 3.5.3). Further investigation would need to be done to determine if it is in fact the case that cluster centers in minimax linkage tend to be better-marked examples.

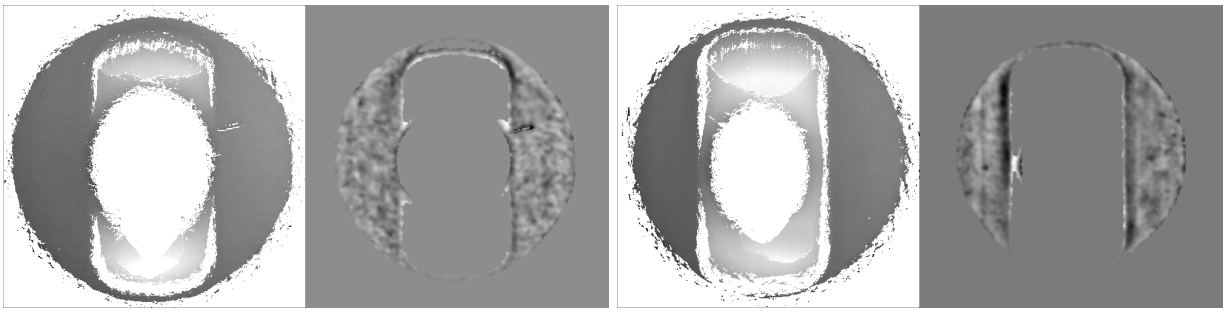


Figure 3.21: The same Glock examples as in Figure 3.13. The original images are on the left and the processed images on the right. This is a non-matching pair and the comparison now has a similarity score of .25, which is low.

Table 3.6: Summary of 2D and 3D results. Max. PR-AUC after clustering refers to the maximum area under the precision-recall graph, among all linkage methods tested.

Study	2D Histograms	2D PR-AUC	2D Max. PR-AUC after clustering	3D Histograms	3D PR-AUC	3D Max. PR-AUC after clustering
Lightstone	Perfect separation	.98	.97	Almost perfect	.91	.97
Weller	Almost perfect	.98	.98	Perfect separation	1	.99
Fadul	Almost perfect	.97	.97	Perfect separation	.99	.95
Hamby	Almost perfect	.98	.97	Almost perfect	.87	.93
Kong	Overlapping	.14	.12	Overlapping	.09	.08
De Kinder	Overlapping	.11	.10	Overlapping	.14	.12
NBIDE	Good separation	.84	.92	Good separation	.86	.94
CTS				Some separation	.87	.94
FBI: Colt	Overlapping	.07	.08	Good separation	.6	.64
FBI: Glock	Overlapping	.07	.11	Almost perfect	.89	.93
FBI: Ruger	Some separation	.42	.47	Almost perfect	.95	.97
FBI: Sig Sauer	Overlapping	.04	.08	Some separation	.44	.47
FBI: S&W	Some separation	.43	.54	Almost perfect	.92	.94

From Figure 3.24 there is no straightforward correspondence between the scores (such as a linear relationship, for example). Some other observations are that 3D scores are generally higher than 2D scores, and high 2D scores generally correspond to high 3D scores. If the 2D score is low however, 3D scores span the entire range of values. For non-matches this is undesirable, but for matches the opposite is true.

Final Clusters

Finally, it might again be of interest to select a cutoff and generate final clusters. For 3D data, it appears that minimax and average linkage have good performance across the board. In addition, minimax linkage has the advantage of being interpretable, so minimax linkage is selected to generate final clusters. Figure 3.26 plots the precision-recall curves using minimax linkage, colored by value of cutoff. A cutoff that works reasonably well for all data sets seems to be around .4. I again illustrate the results for the NBIDE data set.

Here, similar to in 2D, all 6 clusters of 12 are correct. These correspond to the 4 Ruger firearms, and 2 Sig Sauers. The remaining 6 firearms comprise of combining two or more of the smaller clusters, with one exception: one cluster of size 4 mistakenly puts a cartridge case from a Smith & Wesson firearm in a cluster containing cartridge cases from a different Smith & Wesson firearm.

Table 3.7: Cluster sizes for the NBIDE data set in 3D, after hierarchical clustering using minimax linkage with a cutoff of .4. Perfect results would be 12 clusters of size 12.

Cluster Size	1	3	4	5	6	7	9	12
Count	2	5	4	1	3	1	1	6

In terms of performance by gun brand, it appears that from best to worst are Ruger, Sig Sauer, and then Smith & Wesson (earlier Sig Sauer had the worst performance). This is slightly surprising and suggests that Smith & Wesson firearms may not always be easy to identify. 2D results in Section 3.4.2 seemed to suggest that it was indeed the case that Sig Sauers tended to have poorer performance, but this does not necessarily seem to be the case for 3D.

3.5.3 Comparison with Other Work

For 3D, among the 14 data sets used in this paper, work has been published on 5 of them: Fadul (Song, 2013; Lilien, 2017; Song et al., 2018), Weller (Weller et al., 2012; Song et al., 2018), NBIDE and De Kinder (Vorburger et al., 2007), and CTS (Ott et al., 2017). These report perfect separation between match and non-match score distributions for Fadul and Weller, which is similarly achieved using the methodology developed in this thesis.

Vorburger et al. (2007)'s analysis of NBIDE and De Kinder data similarly use CCF_{max} to compute a similarity measure, producing good but not perfect separation for NBIDE, and almost completely overlapping distributions for De Kinder. These are not reproduced in this thesis (due to copyright issues), but the interested reader is encouraged to refer to Vorburger et al. (2007). Based on a visual examination the results look comparable to the results in Figure 3.22. In addition, the number of matches in the top 10 pairs with highest scores was reported. I have not generated this metric for comparison, but this will be the subject of future work.

As for the CTS data, Ott et al. (2017) uses the Congruent Matching Cells (CMC) algorithm (described in Section 3.1.2), and display match and non-match score distributions by firearm. I do the same in Figure 3.27 for comparison. Visually comparing the two sets of results, Ott et al. (2017)'s results appear comparable to slightly worse for Firearm 2, and better for Firearms 1 and 3. The idea behind CMC is that the breechface area is split into a grid of cells that are compared independently. The motivation for using a grid is that "a firearm often produces characteristic marks, or individual characteristics, on only a portion of the bullet or cartridge case surface, depending on its degree of contact with the firearm during firing" (Song et al., 2018).

Using a grid might lead to better results if it is indeed the case that only certain areas on the breechface region are similar due to a non-uniform transfer of marks. It could be the case that Firearms 1 and 3 only had similar marks in some areas of the breechface region, leading to better performance for these firearms using CMC. Nevertheless, CMC was not able to perfectly classify matches and non-matches for Firearms 1 and 2 (the two Ruger firearms), and Ott et al. (2017) state that the reason for this imperfect performance was "large regions where poor contact was made with the breech face." In these cases CMC was not able to overcome the problems caused by the union of two poorly marked images, resulting in poor performance.

3.5.4 Comparison with Published Methodology

Finally, I compare the results to that of my implementation of published methodology. Similar to Section 3.4.5, I consider the NBIDE data set, and a baseline using a manual selection of breechface marks, leveling, filtering, and comparison using CCF_{max}. Here I have eliminated the outlier removal step, as described in Section 3.5.1. Corresponding plots are in Figure 3.28. One can make the same observation that removing circular symmetry reduces similarity scores for non-matches. The effect of automatic selection of breechface marks again seems to be to increase variance.

In addition, precision-recall plots are in Figure 3.29. From this plot it is clear that just switching from a manual to an automatic selection of breechface marks results in poorer performance, which is to be expected. The removal of circular symmetry more than makes up for this, resulting in slightly improved overall performance. It is possible to repeat this exercise for all the data sets, and this will be the subject of future work.

3.6 Conclusion and Discussion

I have applied the matching framework proposed in Chapter 2 to develop fully automatic and open source methods to compare breechface marks on cartridge cases, using both 2D photographic images and 3D topographies. These produce similarity scores for pairs of images, suitable for generating investigative leads, or for blind verification with examiner conclusions. I propose methodology to generate match or non-match conclusions, as well as a linked or disambiguated data set, that can be used for comparing performance with examiner proficiency tests. I suggest using precision and recall to evaluate performance, and area under the precision-recall curve to summarize distributions of match and non-match scores. I also propose methods to estimate a random match probability for the evaluative context. A full evaluation of the methodology was conducted, using over 1000 images from 13 publicly available data sets. Comparisons include those with other work, as well as to a baseline using published methodology.

Compared to other work, our methods have the advantage of being open source and fully automatic, resulting in full transparency and reproducibility. Removing circular symmetry during pre-processing reduces the likelihood of spurious correlations. I propose the consideration of resolving intransitive links, and suggest the use of hierarchical clustering in cases where this is desired. Finally, I present results using AUC for precision-recall graphs as an evaluation metric, providing a better quantification of performance compared to visually examining distributions of match and non-match similarity scores.

In terms of the results, overall performance for 3D topographies is superior to 2D optical images. Methods for 2D images generally transfer to 3D, with the exception of the automatic selection of breechface marks. In particular for Glock firearms with rectangular firing pin impressions, locating the breechface area is much more successful in 3D compared to 2D. In 3D, a lower lateral resolution is used, resulting in much smaller image sizes and faster computation times, while still achieving comparable or better results for all of the data sets analyzed. This provides evidence supporting NIST's recommendation of moving from 2D to 3D imaging.

Performing the clustering step generally improves performance for both 2D and 3D, particularly for data sets with middling to good (but not excellent) results. In terms of area under precision-recall graphs, 3D methods achieve excellent performance (AUC of over .9) on 9 of the 13 data sets, and middling performance on a further two (FBI Colt and Sig Sauer). Performance is generally comparable to other published work, although some evidence suggests that the Congruent Matching Cells method developed by NIST is an improvement over the standard measure of correlation that we used, for some subsets of data.

Finally, I make conclusions in relation to gun and ammunition brands, and observations reported either in the literature or anecdotally by examiners. To make this easier I summarize the results of the individual studies in Table 3.8, combining this with information about the firearm and ammunition brands. (Other work published is also summarized in this table.)

The first conclusion that can be made is that consecutive manufacturing does not appear to be a problem for the firearm-ammunition combinations studied: S&W 40VE/PMC, Ruger P95DC/Winchester, Ruger P95PR15/Federal and Hi-Point C9/Remington. This corroborates Lightstone (2010), where examiners observe that individual characteristics on each slide were sufficient for examiners to make correct identifications. The second conclusion is that it is not possible to sustain blanket statements such as "Sig Sauers mark poorly." The statement on Sig Sauers seemed to have some support using 2D data, but this was less true using 3D topographies. The Sig Sauer P226s in the De Kinder data set were associated with extremely poor performance and among the FBI studies, Sig Sauers had the poorest performance. However the P226s in the NBIDE study had good performance particularly in 3D, despite also involving different ammunition brands. Considering other firearm brands, the performance of Ruger firearms also varied widely between the different data sets. Poor performers were Ruger P94DC and P91DCs (CTS), but good performers included P95D (NBIDE), P95DC (Weller), P89 (Cary Wong), and P95PR15 (Fadul). Smith & Wessons had generally good performance SW40VE (Lightstone, CTS), with the 9VE being slightly worse in NBIDE. The final observation

is that there is some evidence of ammunition effects. Only two data sets (De Kinder and NBIDE) involved different ammunition brands, and this is confounded with other factors such as the firearm brand used, so it is hard to make a definitive conclusion. De Kinder's data had very poor performance, while good performance was achieved on NBIDE.

Table 3.8: Summary of data and results in 2D and 3D for data in NIST's Ballistics Toolmark Research Database.

Study	Cartridge cases	Firearm	Cartridge	Max. 2D PR-AUC	Max. 3D PR-AUC	Published results/ Other comments
Lightstone	30	S&W 40VE	PMC	.98	.97	
Weller	95	Ruger P95DC	Winchester	.98	1	2D: Roth et al. (2015) 3D: Weller et al. (2012)
Fadul	40	Ruger P95PR15	Federal	.97	.99	2D: Tong et al. (2014) 3D: Song (2013); Lilien (2017)
Hamby	30	Hi-Point C9	Remington	.98	.93	
Kong	36	S&W 10-10	Fiocchi	.14	.10	
De Kinder	70	Sig Sauer P226	Remington CCI Wolf Winchester Speer Federal	.11	.14	2D: Vorburger et al. (2007) 3D: Vorburger et al. (2007)
NIST Ballistics Imaging Database Evaluation	144	Ruger P95D S&W 9VE Sig Sauer P226	Remington Winchester Speer PMC	.92	.94	2D: Vorburger et al. (2007) 3D: Vorburger et al. (2007) Ruger had best performance
CTS	74	Ruger P94DC Ruger P91DC S&W SW40VE	Federal Federal Federal		.94	3D: Ott et al. (2017) Rugers had poor performance
FBI: Colt	90	Various Colts	Remington	.08	.64	
FBI: Glock	90	Various Glocks	Remington	.11	.93	
FBI: Ruger	100	Various Rugers	Remington	.47	.97	
FBI: Sig Sauer	130	Various Sig Sauers	Remington	.08	.47	
FBI: S&W	138	Various S&Ws	Remington	.54	.94	
Cary Wong	91	Ruger P89	Winchester			Reasonably high scores for 3D

Some areas of future work are to do a more comprehensive comparison of algorithm performance with that of examiners. For example, the CTS data originated from an examiner proficiency test, and examiner responses (both multiple choice responses and detailed comments) are available online.[¶] These responses could be analyzed more closely. In addition, the PCAST report (President's Council of Advisors on Science and Technology, 2016) mentions one study, Baldwin et al. (2014), which they found to be acceptable for estimating examiner error rates. This study reported both false positive and false negative rates for 218

[¶] https://cts-forensics.com/program-3.php

participating examiners. A team at Iowa State University are in the process of imaging a subset of these cartridge cases, which can be then be analyzed.

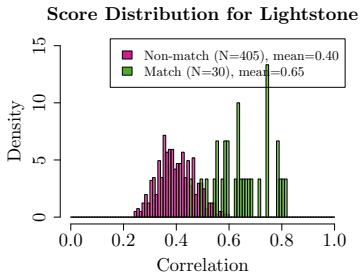
Next, more could be done in terms of comparisons with baselines. In terms of published results, as explained in Section 3.2, results are not published in any standardized way; most commonly distributions of match and non-match similarity scores are presented. In this thesis I have visually compared our results with these distributions where available, but in some work other metrics have been computed, such as the number of matches in the top 10 returned pairs. In particular detailed results were published in Vorburger et al. (2007) for the NBIDE and De Kinder data sets, which could be used to do a fuller comparison. In terms of published methodology, results were presented for the NBIDE data set, but the same analysis could be done for all of the data sets. This was not undertaken because of the manual effort required to select breechface marks on all of the images, but this could be done for a more comprehensive analysis.

Additionally, to make our work more accessible, especially to examiners, it would be beneficial to create an interface (for example using Shiny (RStudio, Inc, 2013)) where users can upload images and get processed images and comparison results as an output. Another feature that would be extremely useful would be to create sliding scales that adjust parameters involved both in the pre-processing and comparison steps. This would result in a better understanding and more interpretability for all the steps involved. Related to this, it would also be useful to undertake a more careful study of all parameters involved, as well as the order of pre-processing steps.

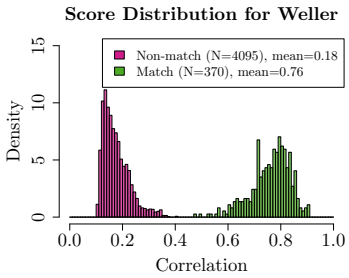
Finally, I invested a non-trivial amount of time on improving the optimization process, in particular finding alternatives such as gradient descent to the grid search (that I eventually stuck to). In particular, the Lucas-Kanade algorithm from the computer vision literature (Baker and Matthews, 2004) seemed to be a promising direction, but my implementation frequently got stuck in local minima and I was unable to achieve comparable performance to the grid search, despite significant speedups in computation. More work could be done in this direction in the future.

To conclude, automatic methods have a lot of potential in terms of corroborating and enhancing examiner testimony. We have produced good results on many studies, but to err on the side of caution, these methods should be tested much more extensively before being used in real cases.

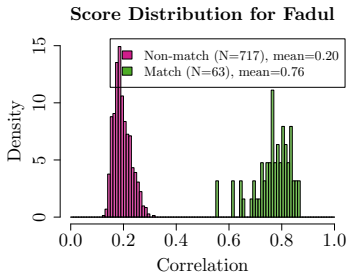
Figure 3.22: Distribution of $\hat{s}$ for matches and non-matches by study using 3D topographies.



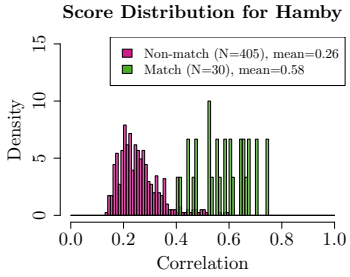
(a) Lightstone (consecutively manufactured)



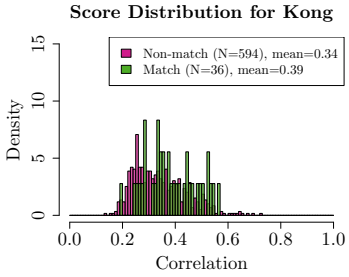
(b) Weller (consecutively manufactured)



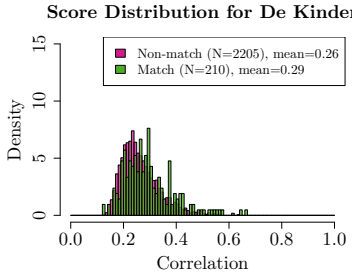
(c) Fadul (consecutively manufactured)



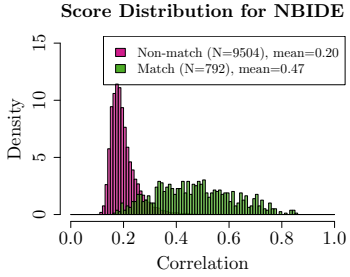
(d) Hamby (consecutively manufactured)



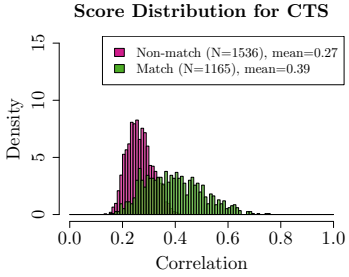
(e) Kong



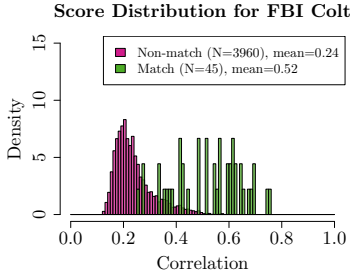
(f) De Kinder



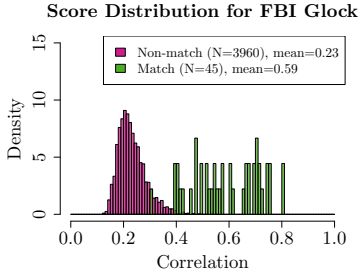
(g) NBIDE



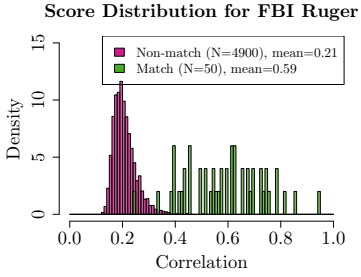
(h) CTS



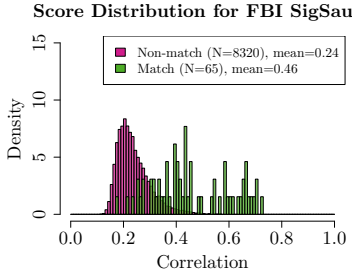
(i) FBI Colt



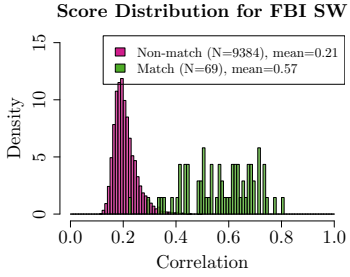
(j) FBI Glock



(k) FBI Ruger

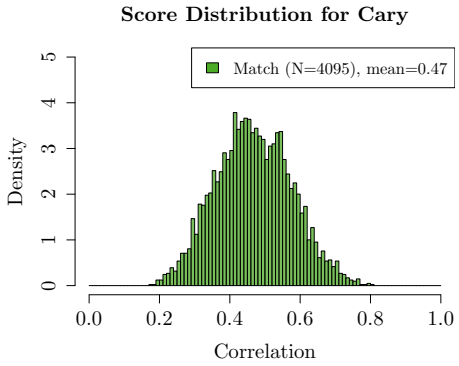


(l) FBI Sig Sauer

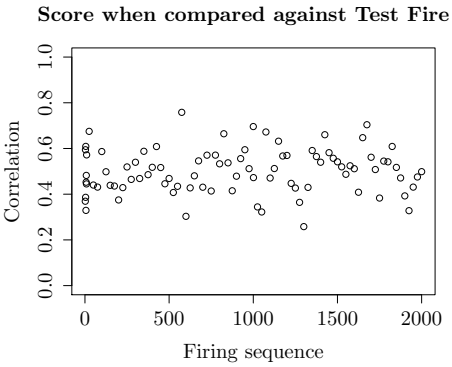


(m) FBI Smith & Wesson

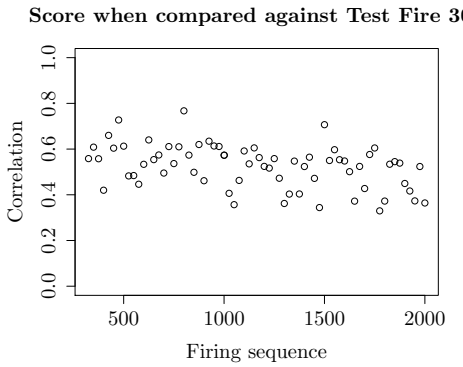
Figure 3.23: Results for Cary Wong study using 3D topographies.



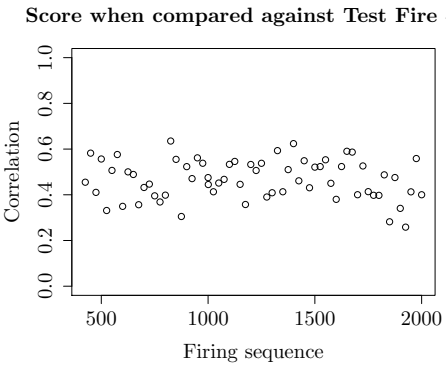
(a) Wong overall results



(b) Compared to test fire 1

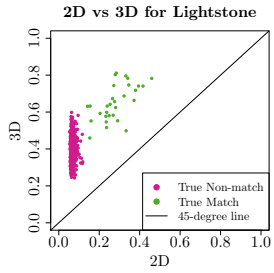


(c) Compared to test fire 300

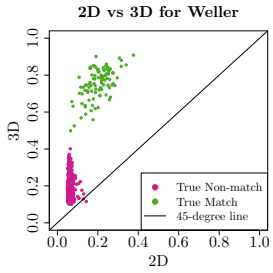


(d) Compared to test fire 400

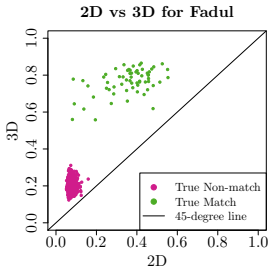
Figure 3.24: 2D scores vs 3D scores by study.



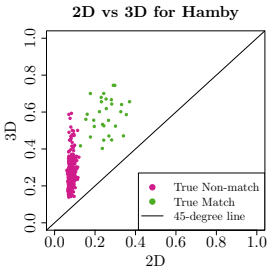
(a) Lightstone (consecutively manufactured)



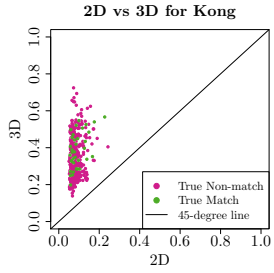
(b) Weller (consecutively manufactured)



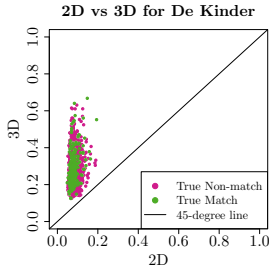
(c) Fadul (consecutively manufactured)



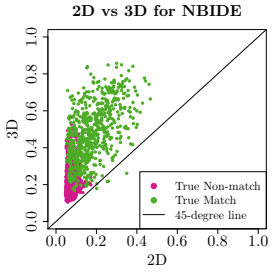
(d) Hamby (consecutively manufactured)



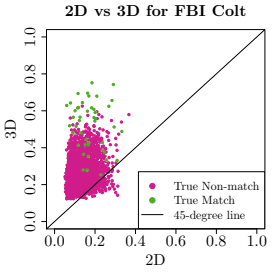
(e) Kong



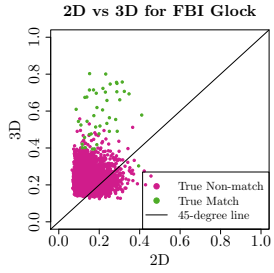
(f) De Kinder



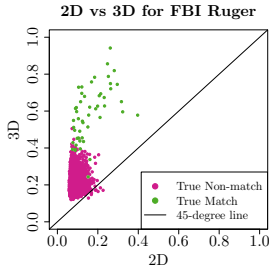
(g) NBIDE



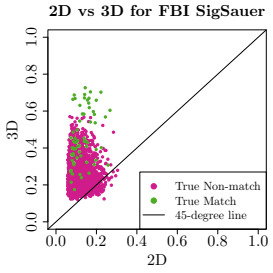
(h) FBI Colt



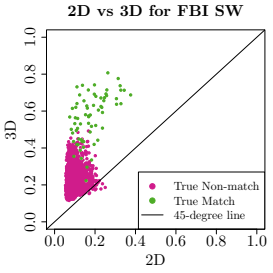
(i) FBI Glock



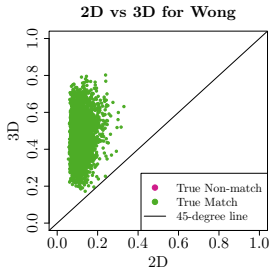
(j) FBI Ruger



(k) FBI Sig Sauer

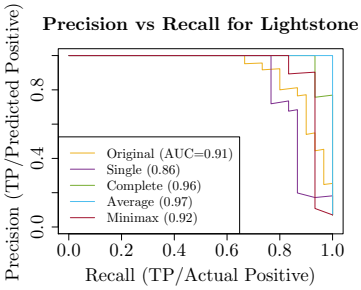


(l) FBI Smith & Wesson

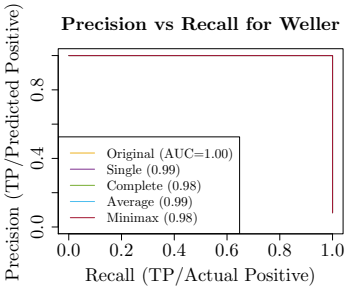


(m) Cary Wong

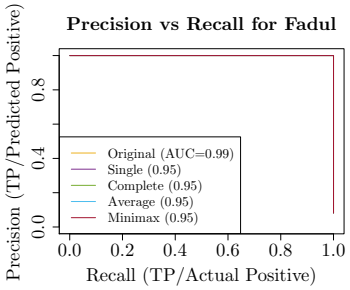
Figure 3.25: Precision-recall plots by study using 3D topographies. Area under the curve is reported in parentheses.



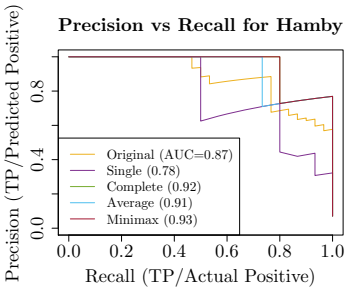
(a) Lightstone (consecutively manufactured)



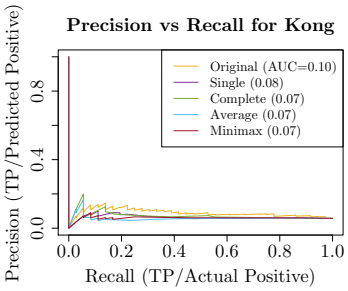
(b) Weller (consecutively manufactured)



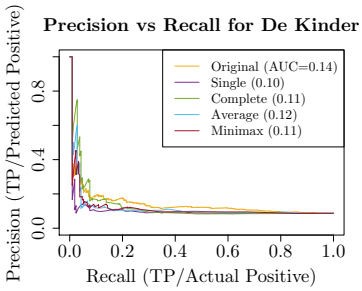
(c) Fadul (consecutively manufactured)



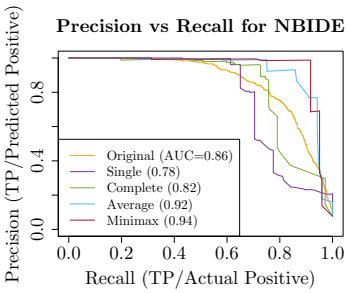
(d) Hamby (consecutively manufactured)



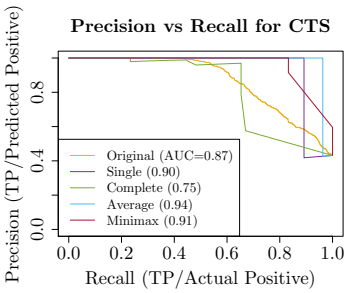
(e) Kong



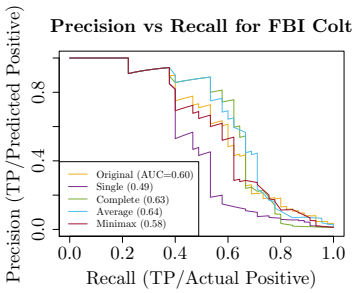
(f) De Kinder



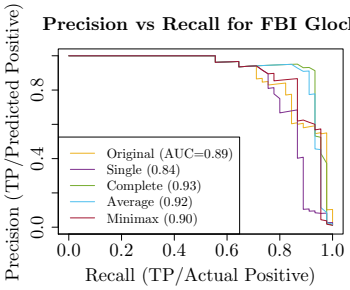
(g) NBIDE



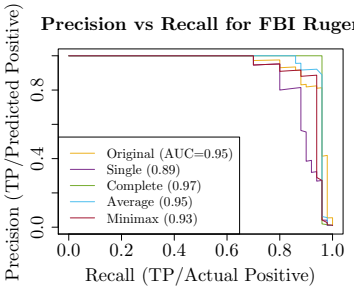
(h) CTS



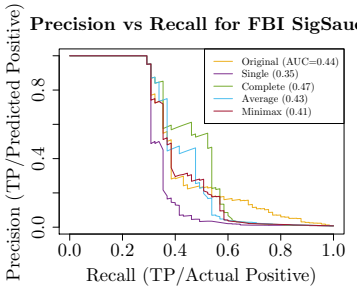
(i) FBI Colt



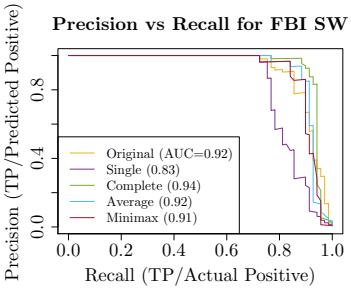
(j) FBI Glock



(k) FBI Ruger

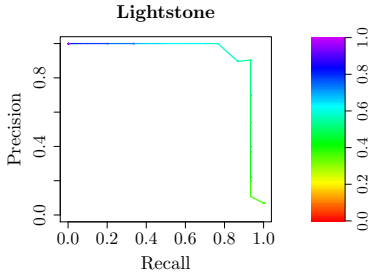


(l) FBI Sig Sauer

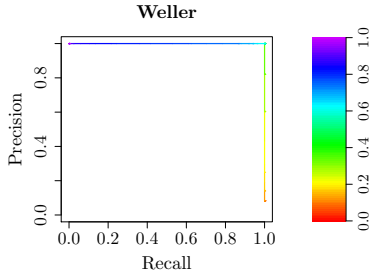


(m) FBI Smith & Wesson

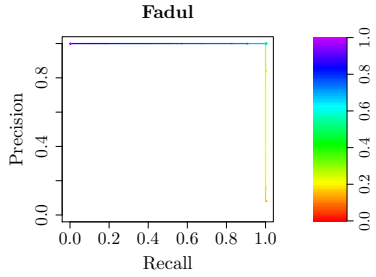
Figure 3.26: Precision-recall plots after hierarchical clustering using minimax linkage by study, colored by cutoff, for 3D topographies.



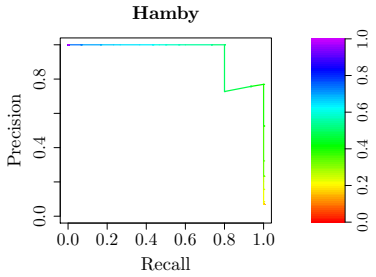
(a) Lightstone (consecutively manufactured)



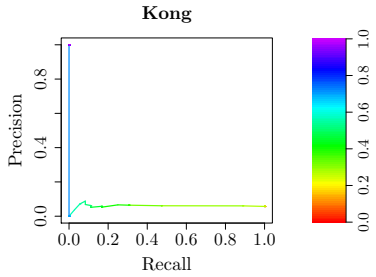
(b) Weller (consecutively manufactured)



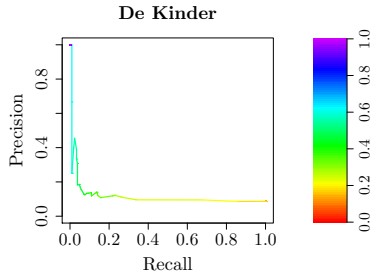
(c) Fadul (consecutively manufactured)



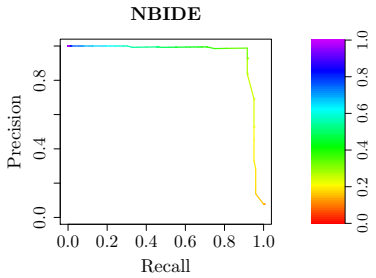
(d) Hamby (consecutively manufactured)



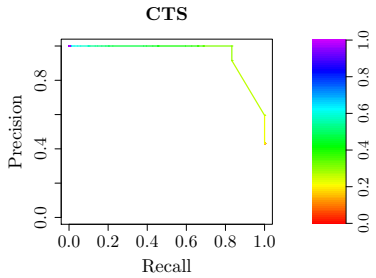
(e) Kong



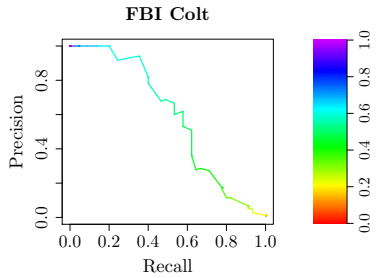
(f) De Kinder



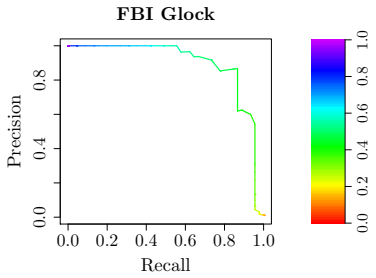
(g) NBIDE



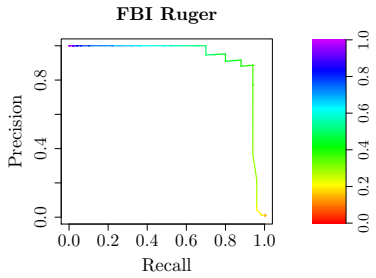
(h) CTS



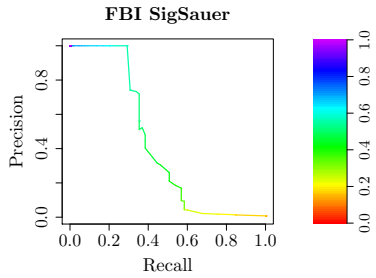
(i) FBI Colt



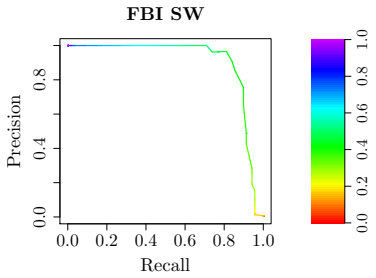
(j) FBI Glock



(k) FBI Ruger



(l) FBI Sig Sauer



(m) FBI Smith & Wesson

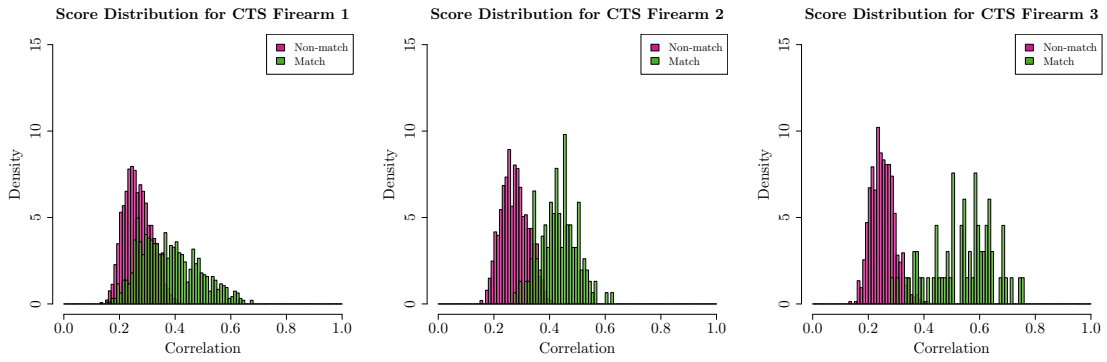


Figure 3.27: CTS results by firearm, for comparison with Ott et al. (2017). The corresponding plots are the top set of results in Fig. 6 of Ott et al. (2017).

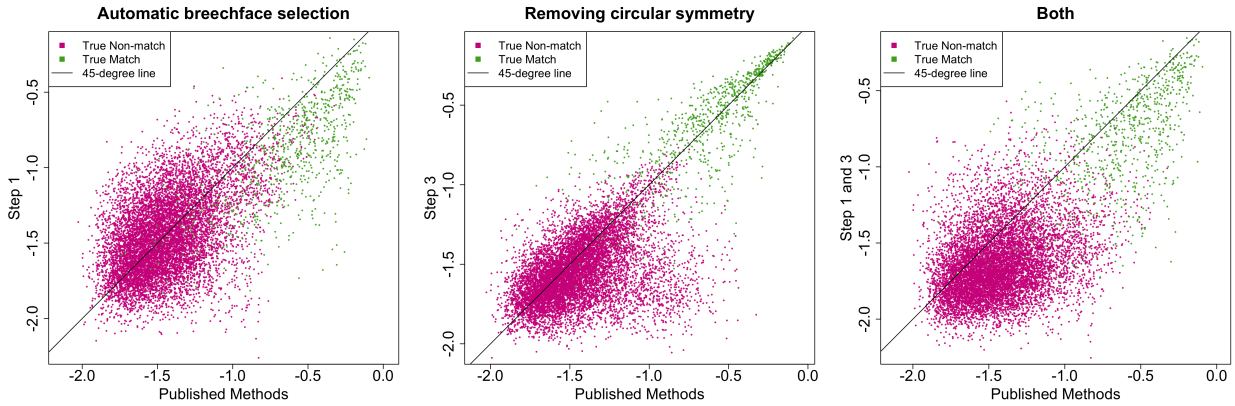


Figure 3.28: Plots of scores using known methodology against scores from adding the automatic selection of breechface marks (annotated as Step 1) and removal of circular symmetry (annotated as Step 3), for the NBIDE study. The logarithmic scale is used on both axes to highlight the differences in the lower values.

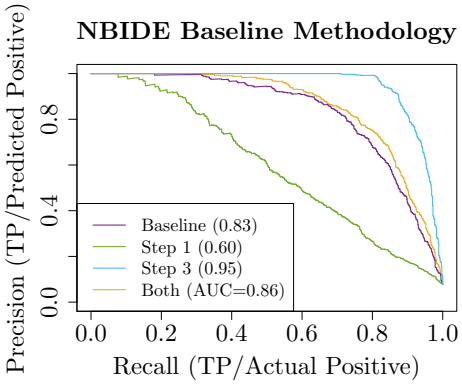


Figure 3.29: Comparison of precision and recall using my implementation of published methodology for the NBIDE study. The baseline uses a manual selection of breechface marks, leveling, filtering, and comparison using CCF_{max}. The addition of automatic selection of breechface marks is annotated as Step 1, and removal of circular symmetry is annotated as Step 3. Both refers to the addition of both steps, and is the methodology described in Section 3.5.1.

Chapter 4

Matching Seller Accounts on Anonymous Marketplaces

This chapter turns to a different problem, matching seller accounts on anonymous marketplaces. Perhaps due to the relative recentness of anonymous marketplaces (the first started in 2011), this has not been given much attention in the forensics community. I give some background in Section 4.1, and explain why this is a worthwhile problem to tackle. I describe the data in Section 4.2, and then apply the matching framework introduced in Chapter 2. We (my collaborators and I) use supervised methods for this problem; Section 4.3 describes how the data are labeled. Section 4.4 describes the methodology in detail, Section 4.5 evaluates the results and Section 4.6 briefly introduces an R package used to analyze the data. A discussion is in Section 4.7, and Section 4.8 concludes.

4.1 Introduction and Background

The dark web refers to a part of the web that is not indexed by search engines, and requires specialized software, such as Tor, to access. Online anonymous marketplaces or darknet markets are commercial marketplaces that run on the dark web. Buyers and sellers make use of cryptocurrencies such as bitcoin, and use encryption methods such as Pretty Good Privacy (PGP) to preserve anonymity. As a result these marketplaces are used primarily for the sale of illicit products such as drugs, weapons, forged documents, and even illegal wildlife. Silk Road was the first such marketplace, and operated from 2011 to 2013 when its founder was arrested. This did not result in the demise of anonymous marketplaces; instead new marketplaces appeared. This has continued over the subsequent years, becoming a game of whack-a-mole.

Sellers have done business on multiple marketplaces concurrently, as well as moved between marketplaces. By conducting such business, they invariably leave evidence of criminal activity. These marketplaces are in

the public domain, and anyone with some level of technological skills would be able to browse user profiles and products sold through these marketplaces and accounts. To put it simply, these sellers hide in plain sight, and the forensic challenge then is to track down the real-world individuals behind these accounts. In earlier sections I described how matching forensic evidence such as cartridge cases aids investigations by enabling evidence from different crime scenes to be combined. Linking seller accounts does the same; in particular, this could help with eventually linking accounts to real-world identities. To give an example, in perhaps the most well-known marketplace-related arrest, the creator of Silk Road was known by a pseudonym, Dread Pirate Roberts. Authorities linked this pseudonym to another account on a message board, Frosty, and Frosty's real-world identity was known to be Ross Ulbricht, thus leading to his arrest (Popper, 2015). Another reason linking accounts is useful to law enforcement is that the combined sales on accounts operated by the same individual could factor into prosecutorial decisions. Arrests are frequently accompanied by a report of the various handles used by sellers, as well as the number of transactions or sales revenues.

4.1.1 Current Practice

Law enforcement has made over 100 anonymous marketplace-related arrests from as early as 2012. More recently, these marketplaces have been in the spotlight due to the sale of drugs. For example, a recent report tied them to the nation's opioid epidemic, due to the proliferation of deadly synthetic opioids such as fentanyl (Popper, 2017). When purchases are linked to overdose deaths, this becomes a legal challenge, since law enforcement often seeks to prosecute sellers, but anonymity protections on these marketplaces make it very difficult to link accounts to real-world identities. Authorities have increased their efforts in cracking down on all parties involved in anonymous marketplaces, specifically operators, sellers as well as buyers. There have been some successes. A coordinated effort by global law enforcement resulted in the takedown of two large marketplaces in July 2017. There was also a string of arrests of top marketplace vendors around that period (e.g., United States District Court, Eastern District of California, 2016, 2017). In 2018 the Department of Justice put together a multi-agency team called Joint Criminal Opioid and Darknet Enforcement (J-CODE), consisting of the Federal Bureau of Investigation, Drug Enforcement Administration (DEA), United States Postal Inspection Service (USPIS), U.S. Immigration and Customs Enforcement Homeland Security Investigations (HSI), U.S. Customs and Border Protection (CBP), Department of Justice (DOJ), and the Department of Defense (DOD). This team has subsequently targeted both sellers and buyers in two major recent operations. Operation Disarray in March 2018 identified 19 overdose deaths of persons of interest (Federal Bureau of Investigation, 2018). Operation SaboTor in early 2019 targeted 50 accounts and led to 61 arrests (Federal Bureau of Investigation, 2019).

Investigators have recognized the utility of linking online accounts. Based on descriptions in court records, they have used techniques such as manually matching account handles, cryptographic public keys (frequently

advertised on accounts) and items sold. They have also searched forum discussions for account mentions (United States District Court, Eastern District of New York, 2016; United States District Court, Eastern District of California, 2016), and used captured login credentials to seize accounts on other marketplaces (Dutch National Police, 2017). All of these methods rely on manual investigation, which can be lengthy and time-consuming. Unlike in other forensic disciplines, there does not currently appear to be an automatic way of matching accounts or doing database searches on a large scale. To my knowledge a national database of seller account information (such as CODIS for DNA or NIBIN for firearms) does not exist, but marketplaces are publicly available and can be scraped to generate the required information. Automated techniques can be especially useful in searching such databases, generating leads in cases where there are no obvious signs pointing towards multiple account ownership. As law enforcement increases their efforts targeting anonymous marketplace-related activity, one might expect such work to become more important.

Additionally, even though the misuse of forensic evidence and overstatement of forensic results has so far been most problematic in the pattern matching disciplines, it is not implausible to see this playing out in the form of digital evidence. Imagine a hypothetical situation in which a victim of a fentanyl overdose is known to have purchased the product from a particular online account (call this the questioned account). Now, an investigator may testify with absolute certainty that a different account that is operated by a known individual has the same ownership as the questioned account, implicating this known individual. This statement could be backed up by a seemingly scientific analysis, for example the style of writing or use of particular symbols. This could result in wrongful convictions in the same way as in other pattern matching disciplines. Use of automatic methods in digital evidence can similarly remove this potential subjectivity. Needless to say, they should first be comprehensively tested, and error rates should be properly quantified.

4.1.2 Overview of Literature

Several attempts to automatically match seller accounts exist in the literature, but most of these rely on exact matching schemes that are inflexible and vulnerable to impersonation attacks. Soska and Christin (2015) used PGP keys, aliases and information from the Grams (a marketplace search engine) seller directory. Broséus et al. (2016) analyzes vendor activity across eight different marketplaces, and used PGP keys and aliases, as well as manual comparison of profile information to link seller accounts. Kruithof et al. (2016) similarly used exact matching on PGP keys, aliases and profile descriptions. Dolliver and Kenney (2016) found an 8% overlap between aliases between two marketplaces, but made no further attempt to infer if they belonged to the same entities. Buskirk et al. (2017) used processed aliases, resulting in a 52% reduction from the number of accounts to unique entities. The problem with these exact matching schemes is that they assume that there are no errors in the variables used for matching.

Wang et al. (2018) use a probabilistic matching scheme, relying on item images to link accounts. Their methodology is based on the premise that photography styles are distinct and reveal user identities. Some drawbacks are that images might be normalized by marketplace operators before publication, who might modify or remove distinguishing features. An attacker could also relatively easily copy image features.

Apart from this, there seems to be little additional research, however the use of "sockpuppet" accounts (a user account that is controlled by an individual or puppetmaster who controls at least one other user account) (Kumar et al., 2017) have been studied in other contexts. In particular, Kumar et al. (2017) characterize sockpuppet behavior on online discussion communities. They also briefly tackle the actual matching task, although this is not the focus. The context they analyze is different and calls for features involving community interactions such as responses to posts.

4.1.3 Applying Matching Framework

I first briefly describe the application of the matching framework introduced in Figure 2.2, in the context of matching seller accounts. Specific details are in Section 4.4.

Here we make use of data scraped from publicly available sites in the period from 2011 to 2018. An individual record is all the captured information relating to an individual account. This includes profile information, items sold and their associated information, such as item descriptions, prices and feedback received. More information about the data is in Section 4.2. Data are pre-processed and fields are standardized. Pairwise comparisons are then generated, by extracting string and numerical similarity measures. We use a supervised random forest approach (Breiman, 2001), and Section 4.3 is dedicated to explaining the labeling process.

The final step in the framework is to generate transitive closures. The random forest classifier produces a prediction for whether a pair is a match or not, say p, where $p \in [0, 1]$. This can be converted into a distance measure in the same way as was described in Section 2.5.1, by taking $d = 1 - p$. We then use hierarchical clustering. Again, this is an optional step depending on the goals of the analysis. One might consider the case for imposing transitivity, to generate a linked data set where accounts belonging to the same seller are clustered together. Data involved here are scraped data rather than reference data, and so can be thought of as a census of anonymous marketplace vendors. Linking the data gives an estimate of the number of sellers involved in such marketplaces. Examining the distribution of cluster sizes and accounts in the cluster might also give an idea as to whether the results make intuitive sense. Researchers studying marketplace ecosystems might be interested in conducting analysis related to seller behavior, which is more suitably addressed on a seller-level than an account-level. Some examples of this include seller longevity and reputation. Characteristics of matched pairs can also shed light on the motivations for using multiple accounts.

Similar to in firearms identification, another argument for imposing the constraint of transitivity is that it can lead to better classifier performance. Again, if one is simply concerned with generating investigative leads, as is demonstrated in Section 4.5, then the clustering step might be skipped.

Finally, for evaluation, given pairwise predictions and some set of labels, we again report match and non-match distributions, precision-recall curves, and compute the area under the precision-recall curve as a numerical summary.

To reiterate and expand on the points made in Section 2.5, from a record linkage perspective, the uniqueness of this particular application area is that in the literature, records often refer to single entries in demographic, bibliographic, and other databases (Christen, 2012), while here records are an entire account profile, consisting of a history of user pages, inventories and sales. Further, there is an adversarial component that to my knowledge has not been studied in detail in the literature. To be specific, there could be malicious reasons for sellers wanting accounts to appear to be matched or unmatched. For instance, a seller might hope to impersonate a well-known account to increase their credibility or defraud customers. Sellers could also want anonymity due to participation in illegal activities, creating different personas, or compartmentalizing different lines of businesses. Dishonest sellers might not want their accounts to be associated with one another.

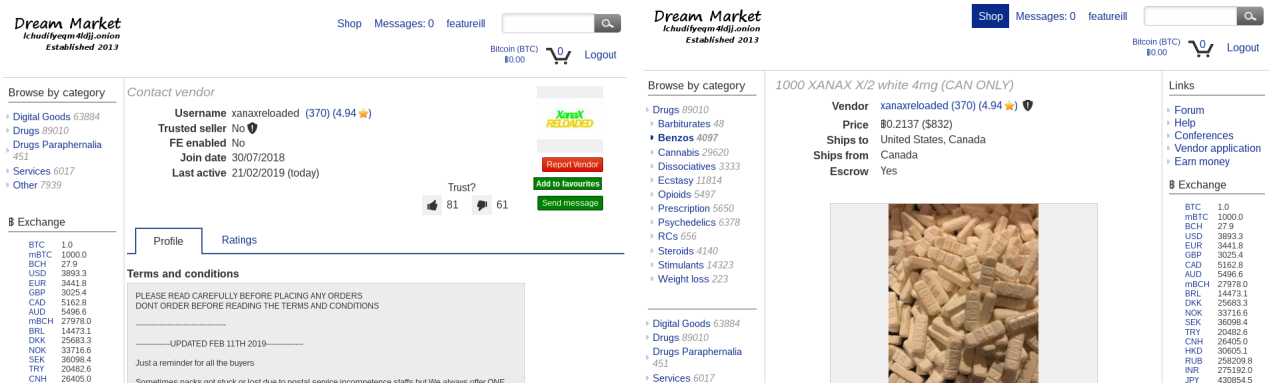
4.2 Data

4.2.1 Marketplace Scrapes

Anonymous marketplaces are publicly available, and one can simply open an account on a marketplace, and access seller profiles as well as item listings. Examples are in Figure 4.1. Seller profiles typically contain a description and available items. Frequently cryptographic (PGP) public keys are available, which are used to encrypt communications with the seller. Item listings include item titles, descriptions, prices, and the shipping origin and destination. PGP keys are also sometimes advertised on these pages. Associated feedback, such as reviews left by buyers, is also listed. Feedback can be used as a proxy for the number of sales, since feedback is often mandatory on such marketplaces, and can only be left by an account that has made an associated purchase (Christin, 2013; Soska and Christin, 2015). Feedback information includes the approximate date the feedback was left and the buyer's comment.

Information from a large number of accounts can be gathered by scraping such web pages. Here the data that we are using are from Soska and Christin (2015)'s data collection effort (which includes data from Christin (2013)), combined with newer data from the AlphaBay marketplace (van Wegberg et al., 2018; Möser et al., 2018), as well as from the Dream, Berlusconi, Valhalla, and Traderoute marketplaces. These pages were scraped and parsed regularly, so for a particular user, there would be multiple captures of their profile page and item listings, each with an associated timestamp. A discussion of the technical and ethical details,

Figure 4.1: Example pages on Dream marketplace. On the left is a seller's profile page and on the right is an associated item listing.



(a) Profile page

(b) Items page

as well as completeness of the data collected can be found at Christin (2013) and Soska and Christin (2015). A large subset of these data are publicly available through the IMPACT portal (DHS S&T – CSD, 2019), and the remainder will be made available in the near future.

Table 4.1 (reproduced from Tai et al. (2019)) summarizes the number of snapshots taken, data collection interval, number of vendor accounts, and number of distinct PGP public keys extracted from the various pages present, for each marketplace. PGP keys are extracted from both profiles and item listings throughout the collection period.

4.2.2 Grams Data

Grams was a dark net search engine that was active from 2014 to December 2017 (Aliens, 2017). It allowed users to search for handles or PGP keys, returning all associated accounts on various marketplaces. Examples of the search interface and returned results are in Figure 4.2. As mentioned in Section 4.1.2, this has been used by some researchers to link accounts. The methodology behind Grams was never explained, but appeared to be crowd-sourced and manually curated (Tai et al., 2019). When Grams shut down, its administrators released their seller databases to the public.

The Grams data contain information on handles from 15 marketplaces, five of which we have data on: Agora, AlphaBay, Evolution, Valhalla, and Dream. From these, there are 28,727 handles, reportedly corresponding to 19,021 unique sellers.

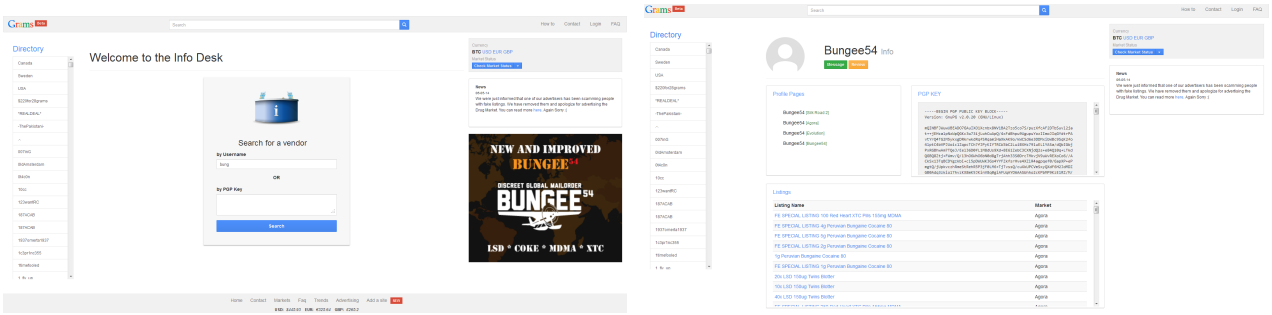
4.2.3 Court Records

When an individual is arrested for allegedly selling illicit products on anonymous marketplaces, court records such as criminal complaints are released. Some of these indicate the aliases used by a seller, and what

Table 4.1: Markets collected and analyzed. The table shows the number of snapshots (complete or incomplete) collected on the various markets present in the study, the collection interval, the number of vendor accounts observed with a sale or PGP key, and the number of distinct PGP keys observed. This table is reproduced from Tai et al. (2019).

Market	**# snap.**	**Collection interval**	**# accts w/ sale**	**# accts w/ key**	**# keys**
AlphaBay	27	03/18/15–05/24/17	6,215	8,370	9,865
Agora	161	12/28/13–06/12/15	1,956	2,563	3,246
Berlusconi	8	11/22/17–08/22/18	226	0	0
BMR	25	10/11/13–11/29/13	975	0	0
Dream	19	07/15/17–08/20/18	4,305	3,950	4,281
Evolution	45	07/02/14–02/16/15	2,338	11,586	12,288
Hydra	29	07/01/14–10/28/14	132	10	11
Pandora	140	12/01/13–10/28/14	457	0	0
Silk Road 1	164	11/22/11–08/18/13	2,327	467	574
Silk Road 2	195	11/24/13–10/26/14	1,196	1,359	1,926
Traderoute	5	07/28/17–10/11/17	1,768	2,463	2,592
Valhalla	3	07/28/17–12/06/17	268	332	341
Other	175	10/19/13–08/11/14	N/A	999	1,160
Total	*996*	*11/22/11–8/22/18*	*22,163*	*32,101*	*36,284*

Figure 4.2: Example pages on the Grams search engine (Source: `https://www.deepdotweb.com/2014/05/17/a-sneak-peek-to-grams-search-engine-stage-2-infodesk/`.)



(a) Search page

(b) Results page

marketplace they operated in. There is no fixed reporting rule in these documents, for example, account names usually need to be significantly different to be mentioned separately; case changes are unlikely to be mentioned.

Emily Rah, a collaborator, manually collected criminal complaints, indictments and sentencing statements from multiple sources (e.g., press articles, DoJ releases). This data collection effort is incomplete, and information is extremely sparse. In total, we have information on 195 individuals, 103 of which mention screen names. In 12 cases, the individual is allegedly using multiple aliases. These are listed in Section 4.5.5.

4.3 Labeling

In the subsequent analysis, we restrict our analysis to accounts that have reported at least one sale, that is, accounts that have obtained at least one piece of feedback on any of the items that they have listed. This corresponds to the 22,163 accounts in the fourth column of Table 4.1. This choice was made since accounts with no sales are unlikely to contribute to the ecosystem.

In Section 4.1.3 it was mentioned that we would be using a supervised classification method. Training this model requires labels on whether pairs of accounts belong to the same seller or not. This is also important for the evaluation step. There is no ground truth available, except through court records, which are extremely sparse and may be incomplete, as described in Section 4.2.3. Hence, multiple alternatives are explored to generate labels, using both marketplace scrapes and Grams data. We note that none of these methods is perfect, but are useful in both training a model and estimating performance.

Method 1: Common PGP Keys

A common heuristic to link accounts is to treat posted pubilc PGP keys as identifiers (Soska and Christin, 2015; Broséus et al., 2016). This is plausible since keys are unique. In such a scheme, accounts posting common PGP keys are labeled as belonging to the same sellers, or matches. However, there are some problems with this approach. The first is that usage of PGP encryption is not mandatory, and not all sellers would list a public key. Secondly, using common PGP keys could result in incorrect labels. A seller might try to impersonate a different seller by advertising their PGP key.* The same seller could also post different keys on different accounts, to assume different identities or compartmentalize their business. Furthermore, there are legitimate reasons for using multiple keys, for example people frequently lose access to their private keys, and generate new keypairs.

With these caveats in mind, public PGP keys are used to generate labels: two accounts are said to be matched if they share a common public key. A precise definition is in Definition 4.1.

Definition 4.1 (PGP labels)**.** *For any two vendor handles i and j, with associated sets of public PGP keys K_i and K_j, consider i and j as mapping to the same vendor if and only if $\exists k \in K_i, \exists k' \in K_j$ such that $k = k'$.*

Method 2: Grams Data

In this labeling method, we simply use the links reported to Grams, as described in Section 4.2.2.

*In such an impersonation attempt, a seller would post a different seller's public key. A potential buyer might encrypt a message to the seller using this public key, and the seller would have to decrypt it using the private key in the key-pair. Obviously, being an impersonator the seller would not have the private key, and would be unable to read the message. The seller might feign ignorance and have the buyer resend the message through some other means. It is not possible to actually verify if the seller is actually in possession of the associated private key, making this a plausible method of impersonation.

Formally, for each handle i, the grams database contains a `link` identifier. Different handles sharing the same link reportedly belong to the same seller, and a precise definition is in 4.2.

Definition 4.2 (Grams labels)**.** *For any two vendor handles i and j, consider i and j to map to the same vendor if and only if $link(i) = link(j)$.*

Intuitively, given the crowd-sourced manual effort involved in curating the Grams database, one might expect these to be more accurate and suitable for use as ground truth. However, these data are limited to certain marketplaces and do not cover all the marketplaces that we are considering. The time period is also limited to April 2014 to December 2017. Since Grams no longer exists, moving forward it is no longer a viable source of data. Finally, since the data are crowd-sourced rather than systematically verified, adversaries might be able to insert false information, resulting in inaccuracies.

Method 3: Profile Descriptions

We also experimented with generating labels using profile descriptions. Sellers often mention other accounts that they own in their profile descriptions. Some examples include "*We are the same OzAlpha from SR1, SR2, BMR & SM etc*," "*Also on Evo market if you're having issues with deposits*," and "*I have been on the following markets previously: Silk Road (top 1% vendor), Black Market Reloaded, Agora (still active), Sheep, Pandora, Silk Road 2.0. And now EVOLUTION.*" Unfortunately, because of the wide variety of ways in which one can express ownership of multiple accounts, simple regular expression matching fails to yield usable information. More complex techniques might involve text analysis of the provided descriptions, and have not been fully tested but will be the subject of future work. Instead, in Section 4.4 we explore some other ways of using information from profile descriptions, in particular to generate similarity measures (instead of labels).

4.3.1 Final Labeled Sets

A combination of PGP and Grams labels (Definitions 4.1 and 4.2) are used to generate two final labeled sets, that are subsequently used for training and evaluating model performance. **Labeled Set 1** is the labeled set corresponding to Definition 4.1, using only a subset of the keys from data earlier in the collection period.[†] This relation does not capture instances where a seller has three or more accounts that do not all use the same PGP key but may be linked through an intermediate key. To address this case, the transitive closure of this relation is also considered. This, together with the full PGP data, as well as Definition 4.2, results in **Labeled Set 2**.

[†]These contained 3,653 PGP matches, while using the full PGP data involves 7,564 matches. Even with this restricted set of labels, when we evaluate using the full set of PGP keys and Grams labels, including transitive closures, results are reasonable, as shown in Figure 4.6.

Table 4.2: Accounts with PGP keys. Distribution of the number of PGP keys associated with each account, for the 22,163 accounts.

Number of PGP keys	0	1	2	3	4	5	6	7	8	9
Accounts	10042	9757	1849	381	81	38	9	4	1	1

Number of Labeled Matches

Labeled Set 1 has 3,653 pairwise matches on the 12,121 out of 22,163 accounts that posted a key in Table 4.1. The distribution of number of PGP keys posted per account is in Table 4.2.

Labeled Set 2 has 8,918 matches involving 18,023 out of the 22,163 accounts. This larger number of accounts that we have information on is unsurprising, since Grams data includes information on sellers that might not have posted a PGP key.

4.4 Methodology

This section describes the details of our implementation of the steps described in Section 4.1.3. We use the 22,163 accounts that reported at least one sale, and the labels generated in Section 4.3.

4.4.1 Individual Features: Account-level Information

For each seller account, we extract account-level information, such as the ID, most sold category,[‡] diversity coefficient (a metric $\in [0, 1]$ to evaluate the diversity across product categories of the goods the account is selling (Soska and Christin, 2015)), as well as information collected throughout the time period, such as all profile and item descriptions, item titles, and feedback received. We include characteristics that are difficult for an adversary to control, such as item prices and days in which sales were made.

Per-account inventories of items that had at least one sale throughout the period are also extracted. For each item, we extract the predicted category (Soska and Christin, 2015), the dosage (number and unit, e.g. "8 grams"), and the quantity e.g., number of pills, tabs, tablets, blotters, etc. To infer dosages and quantities, the regular expressions from Christin (2017) are used, searching within item titles. Specifically, for dosages we look for a number followed by any of the following strings: g, mg, kg, kilo, lb, pound, oz, ug and mcg, and their variants (for example grams instead of g). Similarly, for quantities we look for numbers followed by words such as hits, stamps, tabs, caps, pieces, units, pack, pills, and their variants.

[‡]Prior work (Soska and Christin, 2015) has shown that items sold on anonymous marketplaces can generally be characterized into a few high level categories: Cannabis, Ecstacy, Stimulants, Dissociatives, Psychedelics, Benzos, Opioids, Prescription, Digital Goods, and Others.

4.4.2 Pairwise Features

We design features that are derivable from publicly-available data, that both properly discriminate between matches and non-matches, and are resilient to adversarial behavior. That is, an adversary should not be able to easily produce misleading feature sets or copy feature sets from a different seller. For example, it is straightforward to copy an account handle in a different marketplace, or create an account with a minimally different handle, such as `taco` and `tacos`. Mimicking items sold, sales dates or volumes is significantly more difficult. With these in mind, from account-level information, pairwise comparisons are computed for the 22,163 accounts, resulting in approximately 245 million pairs.

For each pair, the following similarity measures are computed. Some of these are adapted from standard metrics used in the record linkage literature, described in Section 2.4.

- **IDs** Edit (Levenshtein) distance between the IDs (Levenshtein, 1966)
- **Marketplace** Same or different marketplace
- **Profile and item text** Jaccard similarity ($J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ for sets A and B) (Jaccard, 1908) between
  - bag-of-words representation of profiles
  - bag-of-words representation of item titles
  - bag-of-words representation of item descriptions.

  All of the above exclude any extracted PGP keys.
- **Profile and item text length** Absolute difference between number of tokens in
  - bag-of-words representation of profiles
  - bag-of-words representation of item titles
  - bag-of-words representation of item descriptions
- **Inventories** Jaccard similarities between
  - unique categories of all items sold
  - (category, dosage) pairs
  - (category, unit) pairs
  - (category, dosage, unit) tuples
- **Feedback** Absolute difference between
  - number of item listings with feedback
  - number of feedback
  - number of feedback normalized by days active and marketplace total
- **Diversity of product categories** Absolute difference between diversity coefficients (Soska and Christin, 2015). Again, diversity coefficients are a metric $\in [0, 1]$ to evaluate the diversity across product categories of the goods the account is selling.

- **Sales prices** Absolute difference between five-number summary of sales prices (minimum, first quartile, median, third quartile, maximum)
- **Sales days**
  - Hamming distance ($d_H(x, y) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \neq y_i)$, where x and y are length n vectors) between binary vectors encoding days in which sales occur
  - Fraction of overlapping sales days (number of days where both accounts have sales / size of union of sales days)
  - Sum of number of sales days for both accounts
- **Days active** Absolute difference between number of days active (defined as the period between which sales are recorded)

In total, there are 25 similarity measures. As a set, these are costly for an adversary to forge. For example, because only items which received feedback are considered, an impersonator would not only have to post similar items, but receive sales on these items, or have the ability to generate fraudulent feedback to pretend sales occurred. While the latter is not difficult to do in principle, such "feedback padding" is usually noticed by marketplace operators and customers (Ormsby, 2016), possibly by spotting similar reviews by the same "buyer." Offenders are subsequently banned.

4.4.3 Classification

We treat the problem as a supervised classification task, and use both Labeled Sets 1 and 2 (see Section 4.3) to train a classifier. As described, Labeled Set 1 contains 3,653 matches (the pair of accounts shares at least one PGP key), 73,449,607 non-matches (the pair of accounts does not share any PGP key), and 172,134,943 pairs have missing labels (at least one of the two accounts considered is not associated with any PGP key). Labeled Set 2 contains 8,918 matches, 162,396,335 non-matches, and 83,182,950 missing.

Crucially, these labels are not taken to be ground truth when generating predictions, since as discussed in Section 4.3, they could be incorrect. Instead, as an additional step we split the training set into 10 disjoint subsets and obtain predictions by training on 9 of these and predicting on the tenth (in the same way that cross-validation is typically done; these subsets correspond to folds in cross-validation). In this manner, if a training example was a pair consisting of an impersonator copying another vendor's PGP key and that said vendor, the training label would be "match," but it could still have a low model prediction.

A random forest classifier (Breiman, 2001) is used, trained on the extracted features and the generated labels. A random forest averages the predictions from a collection of decision trees, each built from a bootstrap sample of the data. At each split, only a random subset of the variables in the model are considered. Some parameters involved are the number of trees trained, and the number of variables considered at each split. When developing the model on a smaller subset of data, some experimentation was done in terms of optimizing

these parameters. For example, a smaller number of trees generally had more errors. Comparing 50 and 100 trees however, results were not significantly different, despite the much longer training time and higher memory requirements. Due to the large size of the data set, increasing the number of trees beyond 100 was somewhat infeasible. As a result, 50 trees were eventually used. In terms of number of variables tried at each split, increasing the number of variables tended to increase the number of predicted matches, while not significantly improving performance. We eventually stuck to the default setting in R's `randomForest` package of the square-root of the number of variables available (Liaw et al., 2002), in this case resulting in 5 variables tried at each split. Other classifiers such as boosting and logistic regression were also considered, and these produced similar to worse results. Ultimately, in the absence of ground truth labels, it is difficult to properly judge the attempts at model and parameter selection, so this has not currently been fully explored.

Due to the large class imbalance, we also experimented with down-sampling the number of non-matches in the data. Section 4.5.1 presents results sampling equal numbers of matches and non-matches, up to sampling 10 million non-matches.

The output of the random forest classifier, for any pair of accounts (i, j) is a proportion of votes p_{ij} for the accounts correspond to the same seller, where p_{ij} = Number of trees voting for match label / Total number of trees in random forest.

4.4.4 Hierarchical Clustering

Since pairs of accounts are evaluated independently, the classifier might produce matches that are intransitive. If it is of interest to resolve this (see discussion in Section 4.1.3), hierarchical agglomerative clustering is used; this was described in Section 2.5.1. Here we compute the distance (or dissimilarity) d_{ij} between both accounts as $d_{ij} = 1 - p_{ij}$. The same four linkage methods are used (single, complete, average and minimax). We examine how precision and recall changes after clustering.

4.4.5 Random Match Probability/Likelihood Ratio

Similar to what was described in Section 3.2, the above steps give us similarity scores that can be used for ranking, generating match and non-match conclusions, and a disambiguated data set. If additionally measures for the evaluative context (see Section 2.5.2) are of interest, one could use score-based likelihood ratios or random match probabilities.

Similarly to the analysis of cartridge cases, one might think about using a conservative approach. For example, for the non-match distribution similarity scores from pairs that have similar characteristics might be used (for example pairs selling the same category of products). Labeled Set 2 could also be used instead of Labeled Set 1, since Labeled Set 2 has additional links that 1 does not have. Future work will include a more careful analysis of this topic.

4.5 Evaluation and Case Studies

First, distributions of random forest predictions for matches and non-matches are presented. Performance is also evaluated using precision-recall graphs. As discussed we vary the number of non-matches sampled, and explore how this affects model performance. The same is done for the different linkage methods. Evaluations are conducted using both Labeled Sets 1 and 2. Finally, clusters are produced by selecting a particular model and cutoff.

We note that the labeled sets generated in Section 4.3 do not represent ground truth, but still have some utility in examining model performance and the effect of modeling choices. Finally, since we do have limited ground truth in the form of court records, these are examined as case studies, to see how well the model performs in practice.

4.5.1 Classifier Accuracy

Using Labeled Set 1, we train the model sampling all of the labeled matches, and between 3,653 to 10 million labeled non-matches. We then predict on all remaining non-sampled pairs. We additionally split the sampled pairs into 10 disjoint subsets, and obtain predictions on these by training on 9 of them and predicting on the tenth. We evaluate performance first with respect to Labeled Set 1. As detailed in Section 4.3, this consists of 73,453,260 pairs (3,653 labeled matches), coming from the 12,121 accounts that posted at least one PGP key. The remaining pairs are discarded in this evaluation.

Distributions of the match and non-match random forest predictions for the model sampling 10 million non-matches are in Figure 4.3. Precision-recall curves for the various number of non-matches sampled are in Figure 4.5. In all of these figures the area under the curves are given in parentheses in the legend, but should be taken with a grain of salt, because they were computed by connecting the ends of each line to the top-left corner (recall = 0 and precision = 1), and to the bottom-right corner (recall = 1 and precision = 0). These points are not necessarily achieved by the models. Using the red line (just using a threshold-based approach on edit distance of the IDs for classification) in Figure 4.5(a) as an illustration, moving from left to right more pairs are being classified as matches, meaning that points with larger edit distances are classified as matches. At the leftmost point, only points with edit distance 0 are classified as matches, meaning that handles have to be identical for the pair to be classified as a match. Even so, the model achieves only a recall of 0.56 and a precision of 0.83, because there are both matches and non-matches with an edit distance between their IDs that are zero; these are indistinguishable to the model. The AUCs can still give some information, but should be interpreted carefully, keeping the fact that the end points have been connected to the corners in mind. For example, in Figure 4.6 the leftmost end points of the two precision-recall curves are close both to one another and to the top-left corner, so a difference in the AUC might give meaningful information. On the other hand, comparing the yellow line (model sampling 3653 non-matches) and the black line (model

sampling 10 million non-matches) in Figure 4.5(a) is less meaningful since the yellow line's end point is far from the top-left corner.

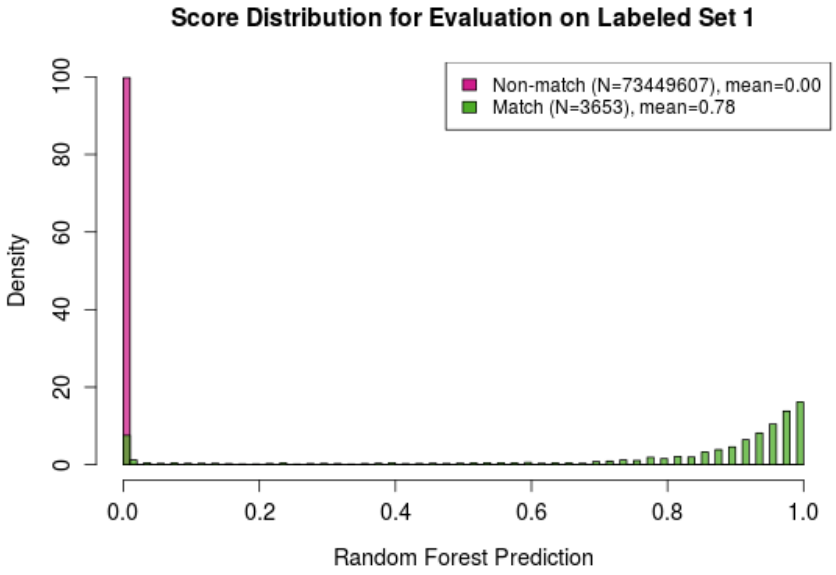


Figure 4.3: Distributions of random forest predictions for matches and non-matches for the model sampling 10 million non-matches, trained and evaluated using Labeled Set 1.

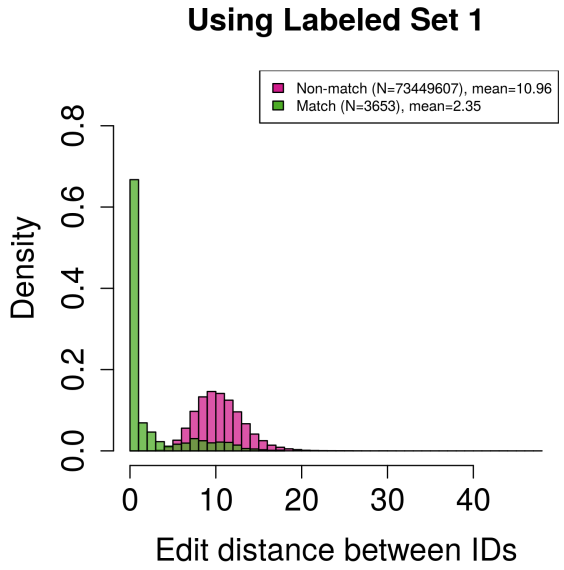
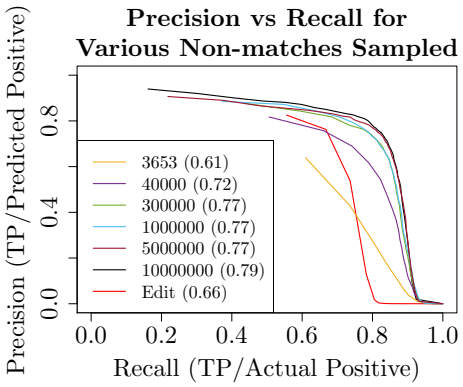
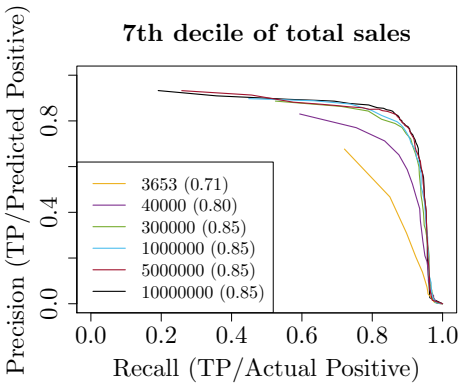


Figure 4.4: Distributions of edit distances between IDs for matches and non-matches using Labeled Set 1.

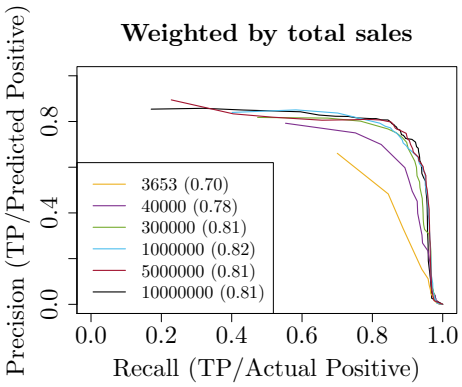
Figure 4.5: Precision vs. recall varying the number of non-matches sampled, using 50 trees. The left hand plot shows results for all accounts, with ID distance as a baseline; the middle plot only considers the top 30% of accounts in sales volume (i.e., those who have sold more than $11,617.81 worth of product); the right plot weighs each point by the accounts' sales volume.



(a) All seller accounts.



(b) Top 30% of seller accounts by volume.



(c) All seller accounts (weighted by volume).

Figure 4.5(a) compares the results of the classifier to a baseline of simply using a threshold-based approach on the edit distance between handles ("ID distance"). Edit distance of IDs was chosen as a baseline since as described in Section 4.1.2, matching the same or similar IDs is a common heuristic used to match seller accounts. Section 4.7.1 also shows that this was the most important feature in the random forest.

Distributions of the edit distances for matched and non-matched pairs (according to Labeled Set 1) are in Figure 4.4. As described, in this approach we simply predict pairs below some cutoff on edit distance to be matches. Looking at the red line (representing the baseline of using edit distance of the IDs) in Figure 4.5(a), going down and to the right the threshold on edit distance below which to classify a pair as a match is increasing, meaning that more pairs are predicted to be matches. The line reaches the horizontal axis at .8 (0 precision and .8 recall), meaning that at this point the threshold is so high that the predicted matches are overwhelmingly false positives. Even so, the recall is only .8, meaning that the model is unable to predict

20% of the actual matches. On the other hand, this kink happens at around .9 to .95 for the other models in the same plot, meaning that now they are completely unable to predict only 5-10% of the actual matches. In other words, using a much more complicated classifier buys an additional 10-15% accuracy on predicting these matches. Separately, a large enough sample of non-matches is required to achieve reasonable performance.

Now, looking at the steep slopes on the colored lines, one can infer that trying to increase recall past 80-85% decreases precision dramatically. In other words, trying to correctly predict all actual matches results in true non-matches overwhelmingly being predicted as matches. This means that the last 15% or so of actual matches may be difficult to predict, in fact the last 5-10% are impossible with the model in Figure 4.5(a). The interpretation is that these pairs of accounts behave very differently from each other, yet share a common PGP key. There could be several reasons for this. One possibility is that a seller opens an account on a different marketplace to reserve the handle, and does not end up using this account much. This has been reported anecdotally, and is investigated further in Figures 4.5(b) and 4.5(c).

Figure 4.5(b) plots pairs only where both accounts are in the top 30% of sales volumes, corresponding to accounts exceeding roughly $11,000 in sales. This eliminates dormant accounts as described. Alternatively, Figure 4.5(c) weighs each pair by the smaller of the sales volumes in the pair, hence down-weighting pairs involving dormant accounts. Both these plots show marked improvements in recall, and the last 5-10% of accounts that were impossible to predict are much closer to being eliminated.

Dormant accounts are not the only reason for false negatives. Some of these could simply be mislabeled as matches, and some examples are described in the case studies described below. On the other hand, pairs incorrectly labeled as non-matches (i.e., same sellers posting different keys) could also affect classifier performance, in the sense that the entire range of behaviors associated with a pair of accounts belonging to the same seller are not captured by the model.

Turning to false positives, through manual examination of model errors, we see that many of these accounts actually belong to the same seller, although they posted different PGP keys. These sellers might have used different marketplaces in non-overlapping time periods, even years apart, and their PGP keys might have expired or they might have lost their private keys. More can be done in terms of quantifying precisely the extent of this problem, and this is elaborated upon in Section 4.7.2.

As a secondary evaluation, Figure 4.6 shows the same precision-recall plot with respect to Labeled Set 2. To be specific, the red line represents results when the model is trained using Labeled Set 1, and evaluated using Labeled Set 2. As described in Section 4.3, Labeled Set 2 includes PGP labels in Labeled Set 1, additional PGP labels from later in the data collection period, Grams labels (Definition 4.2), and transitive closures. This set of labels involves pairs from 18,023 accounts, compared to the earlier 12,121, and involves labels generated in a different way from what was used to train the model (using additional Grams data as well as transitive closures). This results in a pessimistic estimate of the generalization error of the classifier.

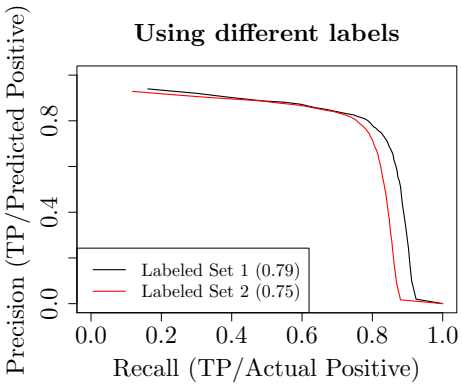


Figure 4.6: Precision vs. recall for model trained using Labeled Set 1 and sampling 10 million non-matches, then evaluated using both Labeled Set 1 and Labeled Set 2.

Figure 4.6 shows that precision is not much poorer, but recall does suffer. The issues with false negatives that were described earlier are exacerbated by the additional links reported.

4.5.2 Clustering Accuracy

Next, we re-evaluate performance after the clustering step, using the four types of linkages as described in Section 4.4. For this, all 22,163 accounts are used for hierarchical clustering to generate clusters, and then evaluation is done only on the 12,121 accounts that reported a PGP key in Labeled Set 1.

Predictions from the classifier that samples 10 million non-matches are used, since from Figure 4.5(a) this produces the best performance. Hierarchical clustering is then run using dissimilarity cutoffs at regularly spaced intervals from 0 to 1, and the precision and recall are computed at each. The results are in Figure 4.7. Minimax and average linkage have superior performance, but minimax linkage has the further advantage of interpretability (see Section 4.4). For this curve, the bend occurs at around a cutoff of 0.74, optimizing the trade-off between precision and recall. Both of these are around 0.8 with this cutoff.

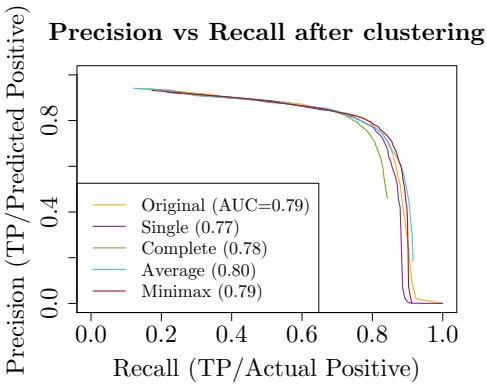


Figure 4.7: Precision vs. recall after hierarchical clustering.

4.5.3 Using Labeled Set 2

The entire exercise is repeated using Labeled Set 2 instead of Labeled Set 1. As previously discussed, even though this set of labels is likely to be more accurate and comprehensive, it is no longer feasible moving forward due to the shutdown of Grams. The following evaluation is done on the 18,023 accounts with either Grams or PGP information, which forms Labeled Set 2. The match and non-match distributions of random forest predictions are in Figure 4.8, match and non-match distributions of edit distances between IDs (the baseline) are in Figure 4.9, and precision-recall graphs are in Figure 4.10. Here match and non-match refer to labels in Labeled Set 2.

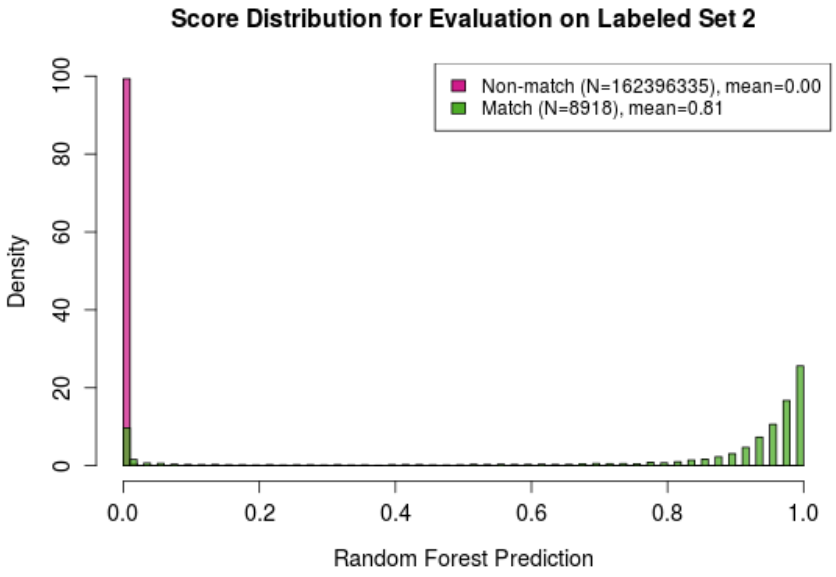


Figure 4.8: Distributions of random forest predictions for matches and non-matches for the model sampling 10 million non-matches, trained and evaluated using Labeled Set 2.

From Figure 4.10 it can be seen that the performance with respect to number of non-matches sampled as well as cutoffs are largely unchanged. Note that the smallest number of non-matches sampled is 8,918 (instead of 3,653 using Labeled Set 1), since the number of matches using Labeled Set 2 is 8,918 and we select the smallest number of non-matches sampled to be equal to the number of matches for a balanced sample. Comparing Figure 4.10(a) to Figure 4.5(a), precision is similar (to very slightly poorer), but recall is worse. The remaining plots are also very similar, with generally slightly poorer overall performance using Labeled Set 2 than Labeled Set 1.

The poorer recall is unsurprising due to the issues with false negatives that were described earlier. Specifically, Grams adds additional true matches to the data set, possibly involving dormant accounts, and these pairs might be indistinguishable from non-matches and would be hard to predict to be matches. As for

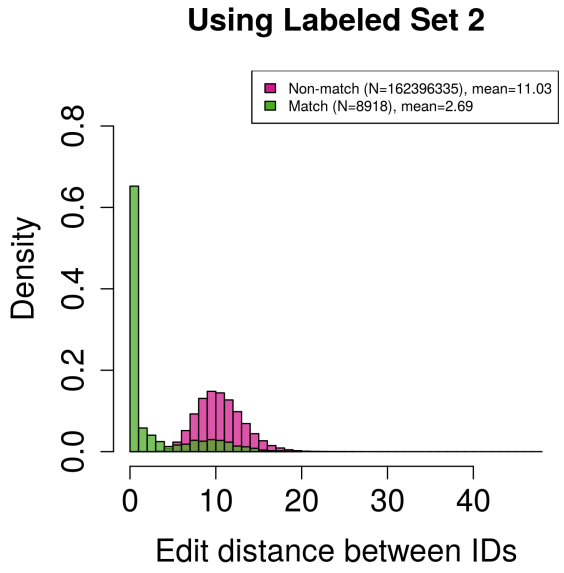


Figure 4.9: Distributions of edit distances between IDs for matches and non-matches using Labeled Set 2.

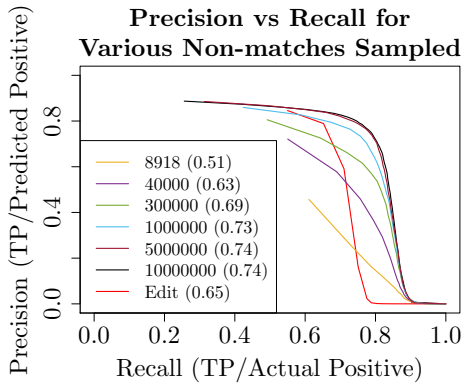
the poorer precision, these additional true matches that might have unexpected behavior can be thought of as additional noise in the training data, resulting in the model making more mistakes in predicting matches. This translates to the slightly poorer precision.

4.5.4 Final Clusters

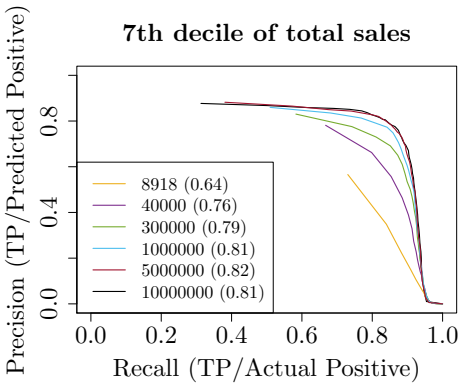
Again, the final choice of cutoff and/or linkage method depends on the type of performance desired. For example, if an individual might be implicated in a crime, false positives could be extremely undesirable, in which case one would want a very high level of precision. If it is of interest to generate investigative leads for a particular account, and if one will be reviewing potential matches manually, one might instead prefer high recall. In this case, as described in Section 4.1.3, transitive closures may not be a concern either, and one might simply select pairs for which the classifier produces higher predictions, for manual review.

To produce matching clusters, we run the entire algorithm on all 22,163 accounts available, using Labeled Set 1, sampling 10 million non-matches. We then use minimax linkage with a cutoff of 0.74 (in Section 4.5.2 this produced good results). The above modeling choices result in assigning the 22,163 accounts to 15,652 distinct sellers. The number of accounts operated by each seller is in Table 4.3. 12,155 sellers operate a single account, while two sellers operate as many as 11 accounts. Hundreds of sellers operate four or more accounts, which underscores the need for proper matching methods.

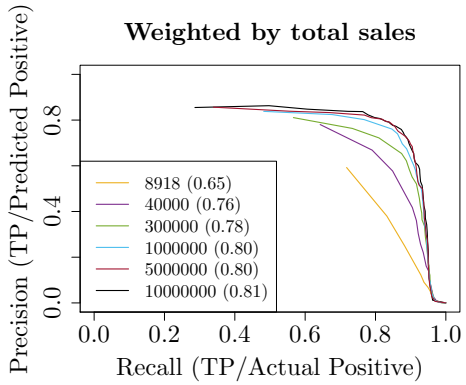
Figure 4.10: Repeating the analysis using Labeled Set 2. Again the results presented are for all accounts with ID distance as a baseline, top 30% of accounts in sales volume, weighted by sales volume, and after hierarchical clustering.



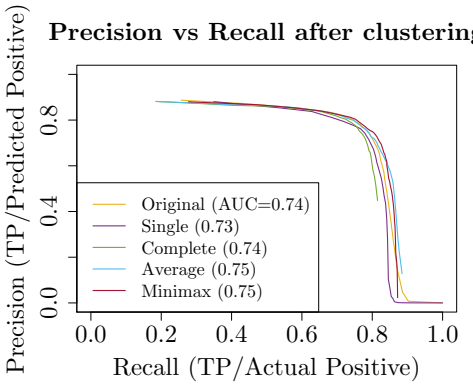
(a) All seller accounts, with ID distance as a baseline.



(b) Top 30% of seller accounts by volume.



(c) All seller accounts (weighted by volume).



(d) After hierarchical clustering.

Table 4.3: Number of accounts operated by a single unique seller. Distribution of the number of accounts associated with the same seller, using a particular set of parameter choices.

Number of Accounts	1	2	3	4	5	6	7	8	9	10	11
Number of Sellers	12155	1909	882	358	151	93	56	31	13	2	2

4.5.5 Case Studies

Manual evaluations are performed on a series of case studies, to examine if the method succeeds or fails in specific situations of interest. This is intended to supplement the earlier evaluation, since as noted the labeled sets do not represent ground truth. In this section three cases are examined. In the first we see if the model is able to predict multiple accounts documented in publicly available court records (Section 4.2.3). In the second we look specifically at adversarial examples, where sellers either try to evade detection or impersonate other accounts. Here online discussions as well as a manual examination of profiles are used to infer ground

truth. Finally we look at cases where the model and label disagree. News reports are used as an additional source of information.

Court Records

It is of interest to check if the model independently infers matches documented through court records. As described in Section 4.2.3, there are 12 cases in which multiple aliases were mentioned. Out of these, five mention accounts that do not appear in our data, possibly because the accounts did not receive any feedback, or due to the incompleteness of our data.

We verify if the models find these same matches, depending on the parameter specifications used. Three different specifications are used: the first is the conservative cutoff of 0.74 that was used earlier, using minimax linkage; the second reduces the cutoff slightly to 0.5 and still uses minimax linkage; the third just looks at random forest predictions (without hierarchical clustering), using a cutoff of .2.[§] As described in Section 4.1.3, when using an automated method to generate investigative leads, one might not be concerned about the hierarchical clustering step. One might also choose parameter specifications designed to produce high recall instead. The model used here samples 10 million non-matches (this corresponds to the outermost curve in Figure 4.5(a)). Models sampling a smaller number of non-matches produce predictions that are even less conservative, but that is not discussed in detail here.

The results are summarized in Table 4.4. The first and strictest model (minimax 0.74) does not find any of these 6 matches, and the last, least conservative model finds at least one alternate account being matched in 4 out of the 7 cases. Specifically, `HumboldtFarms` is matched to `PureFireMeds`, `NarcoBoss` to `DNMKingpin`, and `caliconnect` to `the real caliconnect`. In the remaining three cases, sellers did an excellent job compartmentalizing their businesses over multiple accounts, for example selling different categories of products on different accounts, and the algorithm was unable to find matches.

Adversarial Examples

As described, we have designed a system to detect some subset of adversaries. Two scenarios are examined. First, we are interested in accounts belonging to the same seller, but having different screen names and PGP keys. These might be sellers trying to evade detection. Second, we look at impersonators copying a screen name and/or PGP key.

In the first scenario, a large number of examples of accounts with different screen names and PGP keys ended up in the same cluster, even using the conservative modeling choice of a cutoff of 0.74 and minimax linkage. At a pairwise level, this resulted in 12,320 pairs being predicted to be matches. Of these, 2,910 had common PGP keys, 757 did not have common keys, and the remaining 8,653 pairs had a missing label,

[§]While this might seem like a low threshold, it is less than 0.01% of all pairs.

Table 4.4: Cases where multiple aliases were mentioned in court records. RF .2 refers to the random forest model, using a cutoff of .2. This model was selected for the purposes of generating investigative leads (while the minimax models were not), and successfully finds at least one alternate account being matched in 4 out of the 7 cases.

	Aliases[¶]	Found in data	Minimax .74	Minimax .5	RF .2
1	Area51, darkapollo	✓		✓	✓
2	NarcoBoss, DNMKingpin	✓			✓
3	BTH-Overdose, Blime-Sub	✓			
4	darkexpresso, bonappetit	✓			
5	richierich, happyman, bitcoins	✓			
6	PureFireMeds, HumboldtFarms	✓			✓
7	caliconnect, caliconnect2, the_real_caliconnect, caliconnect4life	✓			✓
8	drbechen, Donkey				
9	Mountain, Goldmountain, Aexpharma, TheGoldenDawn, TheLink, SilentWisdom, Darkwebtycoon, Pandora, Pandora91				
10	GoldCard, slacker, slackerxxx, slackerX, slackerplastics				
11	thehater90, No1Benzos				
12	UnderGroundSyndicate, BTCMaster				

meaning that one or both accounts in the pair did not post a key. The distribution of edit distances of IDs for these predicted matches is in Figure 4.11. Over 30% of pairs had different IDs. Broken down by labeled match status (using Labeled Set 1), pairs posting different PGP keys tend to have larger differences in screen names, more strongly suggesting adversarial intent.

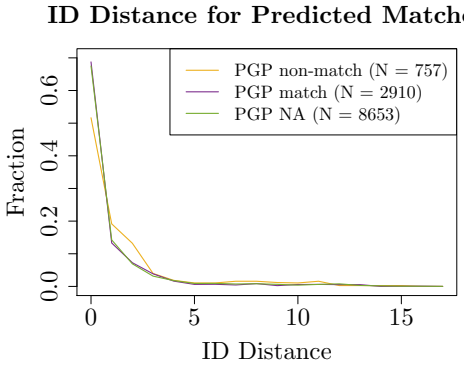


Figure 4.11: Edit distance of IDs for model predicted matches.

[¶]Citations for each of the entries are (1) United States District Court, Eastern District of New York (2016) (2) United States District Court, Western District of Pennsylvania (2017) (3) United States District Court, Eastern District of New York (2017) (4) United States District Court, Middle District of Florida (2016) (5) ITV (2016) (6) United States District Court, Eastern District of California (2017) (7) United States District Court, Eastern District of California (2016) (8) United States District Court, District of Oregon (2013) (9) United States District Court, District of North Dakota (2018) (10) United States District Court, District of New Jersey (2014) (11) Branwen (2012) (12)United States District Court, Middle District of Florida (2014).

Additionally, some predicted matches consisting of accounts with different screen names and/or PGP keys were verified manually; some sellers do not attempt to conceal their identity, providing information about multiple account ownership in their profile descriptions, while others do not explicitly do so. We attempted to verify such cases using online discussion forums such as Reddit. Some examples are `FTB` on Black Market Reloaded being matched to `fredthebaker` on other marketplaces, `kaypee911` on Black Market Reloaded being matched to `aaapee911`, `evopee911`, etc. on other marketplaces, and `Sheld0nC00per` on Pandora being matched to `KeithLemon` elsewhere. All these matches are found despite not having a common PGP key.

As illustrated in the previous subsection, sellers might intentionally take more extreme measures to avoid detection, such as selling different categories of products, adopting different personas, and so forth. In this situation, accounts owned by the same seller are extremely difficult to detect, and there are examples where the algorithm failed (see Table 4.4).

As for impersonators, we found several examples of accounts with the same screen name or the PGP key of a different user, that our algorithm successfully placed in different clusters. For instance, `LowLands` on Silk Road 2 tried to impersonate `LowLands` on other marketplaces, at some point being called out in one of the latter's profiles (*"ATTENTION !!! ON SILK ROAD 2 THERE IS A SELLER LOWLANDS CLAIMING TO BE US....THIS IS NOT TRUE!!!.. BE AWARE !!!!!!"*); similarly `gotmilkreplica` copied `gotmilk`'s PGP key, and posted on forums claiming to be `gotmilk`.[‖] The former ships knockoffs from Hong Kong, while the latter is a large seller shipping prescription medication from India, and it seems unlikely that they are in fact the same seller.

Model and Label Disagreements

Finally, we examine some examples where the model and label strongly disagree. To be specific, notice that in Figure 4.6, there is a kink in the bottom right where even using the smallest similarity cutoff for matches, the model is unable to correctly predict all matches. This was discussed in Section 4.5.1, and it was noted that the problem is worse when evaluating pairs using Labeled Set 2, where the model is unable to predict close to 10% of labeled matches. Restricting to pairs where both accounts are in the top 30% of sales volumes (as in Figure 4.5(b)), there are 117 pairs which are labeled matches (according to Labeled Set 2), but have model predictions of 0.

Looking at these pairs manually, there were 26 pairs (22% of the 117 pairs) involving a cluster of accounts on Dream (`cannab1z`, `GlazzyEyez`, `ibulk`, `MarcoPolo420`, `MissJessica` and `mushroomgirl`), and some of `MissJessica`'s accounts on other marketplaces. Further investigation revealed that during the seizure of Hansa and AlphaBay marketplaces in 2017, the Dutch National Police gained control of at least a dozen

[‖] `https://bitcointalk.org/index.php?topic=834362.0`

accounts on Dream marketplace, and posted their PGP key on all of the account pages.[**] Many of the accounts in this cluster were victims of this takeover. This case study highlights the problems with solely using PGP keys or Grams labels for matching, and suggests that the error rates reported when evaluating our model against these labels are an overestimate.

4.6 R Package

I have developed an R package, `heisenbrgr`, which allows users to do analysis similar to what was described in Sections 4.3 through 4.5, given appropriate scraped and parsed data. For these purposes, there exist publicly available marketplace scrapes.[††] Alternatively, there are various tools that have been developed for scraping pages on the dark web (see e.g. Christin (2013); Soska and Christin (2015)), and one can undertake one's own data collection effort, and process and analyze the data using `heisenbrgr`.

In particular, `runStep1()` and `runStep2()` pre-process the individual level data, `runStep3()` produces pairwise comparisons, and the classifier can be trained using `predictRF()`. Hierarchical clustering can then be done using `linkAnalysis()`.

I encourage users to train a model that is suitable for their own purposes. If this is not an option, the package contains an example model that has been pre-trained using accounts that had at least one feedback received in August 2014. This is a subset of the data used in the preceding sections. There are a total of 2395 accounts from the following marketplaces: Agora, Alphabay, Black Market Reloaded, Dream, Evolution, Pandora, Silk Road 1, Silk Road 2 and Traderoute. 2135 of these accounts had posted at least one PGP key, and Labeled Set 1 was used. There are 2,278,045 labeled pairwise comparisons, of which 610 are matches. 50 trees were used, and no down-sampling was done.

The model fit is stored as a `randomForest` object in `exampleFit`. To generate predictions, users need to have the pairwise similarities as input (the 25 pairwise features listed in Section 4.4.2, or a subset of these). Predictions can then be generated using `predict()`, specifying `type = "vote"`.

Unlike in Chapter 3 where inputting data is straightforward (information on a cartridge cases is contained in a single image), here information on each account contains an entire history of user pages, inventories and sales. As described, the currently included pre-trained model fit allows users to input a comparison vector to generate a random forest prediction. Some of these pairwise similarities can be derived simply using a manual examination, for example, the edit distance between the handles, whether or not the pair is from the same marketplace, and the difference between the amount of feedback that each account has. Missing values are also allowed. Future work will include the ability to input raw data associated with an account, which can be typed or copied from relevant webpages.

[**] `https://www.deepdotweb.com/2017/08/07/dutch-police-taken-12-dream-accounts-likely/`, `https://www.bleepingcomputer.com/news/security/crooks-reused-passwords-on-the-dark-web-so-dutch-police-hijacked-their-accounts/`.

[††] For example, `https://arima.cylab.cmu.edu/markets/cybercrime.php` contains anonymized scrapes

4.7 Discussion

4.7.1 A Closer Look at Classifier Performance

Looking at variable (or feature) importances from the random forest classifier, one can get a better idea of which features specifically are important for an adversary to stage a successful attack. Using the classifier sampling 10 million non-matches, variable importances are presented in Table 4.5, measured using mean decrease in Gini impurity. The Gini impurity for K classes is defined as $\sum_{k=1}^{K} p_k(1-p_k)$, where p_k is the fraction of items labeled with class k. The Gini impurity decreases after each split. For a single tree, summing these whenever a particular feature is used in the split gives the decrease in Gini impurity for that feature. Taking the mean over all trees gives the mean decrease in Gini impurity, and this provides a measure of feature importance. As one can see, the items sold through each account, and their associated information plays a large role in determining if a pair of accounts is classified as a match or not. The implication is that for an impersonation attack to succeed, an impersonator would have to sell products that have item titles and descriptions that are very similar to the account that they are impersonating. This is hard to forge, as sales actually need to be confirmed by feedback for the classifier to consider the associated accounts.

Table 4.5: Variable importances of random forest classifier. The top 10 (out of 25) pairwise comparison features and their associated importance (measured using mean decrease in Gini impurity) are listed in decreasing order.

Variable	Mean Gini decrease
Edit distance between IDs	3690
Jaccard similarity between item title tokens	1109
Jaccard similarity between item description tokens	697
Jaccard similarity between profile tokens	206
Same or different marketplace	160
Difference between number of item title tokens	109
Difference between number of item description tokens	93
Difference between fraction of daily sales	89
Difference between mean item price sold	88
Hamming distance between sales dates	88

4.7.2 Limitations

As described, the labels used are heuristics rather than ground truth labels. Using PGP labels for example, there are many inaccuracies due to the same seller using multiple keys, or different users using the same key. In Section 4.5.5 I discussed an incident in which the Dutch National Police posted the same key on multiple accounts. We have attempted to generate labels in alternative ways. This could be the subject of further work in the future. As discussed in Section 4.3, generating labels from profile information using regular

expression matching was not particularly successful, but could be improved either by manual extraction or using more sophisticated tools to extract evidence from profile descriptions. As part of the effort to generate more accurate labels, I also adapted a Shiny interface (RStudio, Inc, 2013) to manually label if pairs of accounts belonged to the same seller or not. A screenshot of the interface is included in Appendix B. The idea is to manually review some subset of pairs, for example those where the model predictions disagreed with PGP labels, or when unusually large clusters were produced in the hierarchical clustering step. These manually labeled pairs could either be used to correct errors in final clustering results, to retrain the model, or as a held out test set when exploring different models. After labeling somewhere around 500 pairs, this effort was abandoned due to the considerable manual effort required, as well as difficulty even for a manual labeler to be certain about whether pairs belonged to the same seller or not. This could be explored more in future work.

Related to the same point of unavailability of ground truth labels, it is difficult to determine the proportion of false positives that are in fact true positives, or false negatives that are in fact true negatives. This was noted in Section 4.5.1. One option would be to randomly sample cases, possibly bucketed by model score, for manual review. We did this to some extent in Section 4.5.5, noting that out of 117 selected false negative pairs, at least 22% are true negatives, but a more comprehensive analysis could be done.

Next, the algorithm itself is susceptible to adversaries that take great lengths to conceal or mimic behavior. Several examples were given in Section 4.5.5, where the model was not able to find some links documented in court records. With respect to this adversarial context, apart from generating better features and labels, an unexplored area is the methodology. To be specific, if it is impossible to generate accurate labels or features, are there things that can be done from a methodological standpoint to deal with the problem? For instance, one direction might be to examine the robustness of the model with respect to swapping labels. Another direction might be to re-weight observations based on how confident one is about label accuracy or adversarial intent.

Another limitation is that the current methodology assumes that account ownership does not change over time, which anecdotally is known to be false, for example due to sales of accounts, or due to police takeovers. Finally, scalability is a notable limitation, since like in a typical record linkage situation, the number of pairwise comparisons necessarily grows at rate n^2, where n is the number of individual items. Furthermore, generating some of the pairwise features is extremely slow and memory-intensive. Likewise, training the model is very memory-intensive. It was prohibitively expensive to sample more than 10 million non-matches. Future work will investigate methods to improve scalability, such as indexing.

4.8 Conclusion

In this chapter I describe methods developed to match seller accounts on anonymous marketplaces, demonstrating another application of the framework introduced in Chapter 2. The models produce predictions for whether a pair of accounts belongs to the same seller, which in the forensic context can be treated as a similarity score. This allows the generation of investigative leads, match or non-match conclusions, as well as a linked data set. The methodology is developed using eight years (2011-2018) of online anonymous marketplace data. Ground truth for whether pairs of accounts belong to the same seller are unavailable, and I describe several different methods that are used to generate labels. The models are evaluated using two different sets of labels generated, and the evaluation includes comparisons to a baseline of just using edit distances between account IDs. Although these labels do not represent ground truth, they give a general sense of model performance and sensitivity to modeling choices. Finally, several case studies are examined as a further evaluation on specific examples of interest, where ground truth is available through criminal complaints, or can be inferred through forum discussions and news reports.

In terms of the results, using generated labels, the models achieve more than 75% precision and recall by selecting appropriate cutoffs for the classifier. Performance is superior to a baseline of a threshold-based approach using only edit distance between IDs. The methodology works particularly well for accounts with significant sales volume (achieving around 90% recall at 75% precision for the top-selling 30% of accounts). Performing the clustering step slightly improves model performance. After the clustering step and using a set of chosen parameters, the 22,163 accounts with at least one confirmed sale map to 15,652 distinct sellers. 12,155 sellers (77%) operate only one account, while the remainder operate up to 11 accounts. Finally in case studies, links documented in court records were discovered by an appropriate model in 4 out of 7 cases. The models can automatically discover potential adversarial behavior, reported impersonation attempts, non-trivial links between accounts, and instances in which the labels are incorrect.

Apart from the usefulness of matching from a forensics perspective, in the analysis we were fortunate to have real-world data scraped from actual marketplaces, which gives the analysis additional utility: researchers studying these ecosystems might be interested in getting an accurate idea of the number of sellers involved, which we are able to estimate. For marketplace patrons, matching accounts is important to verify the identity of sellers and avoid both scams and law enforcement, and to evaluate the credibility of the seller. This is evidenced by the large number of posts on Reddit and other forums about the topic, as well as the popularity of the Grams vendor directory before its shutdown. From the record linkage perspective, this chapter contributes an application where an adversarial element is present, in the sense of sellers deliberately trying to evade matching, or impersonating other sellers. To deal with this we make extensive use of features that are costly for an adversary to implement, and generated predictions on training examples. More can

be done from a methodological standpoint, and we hope this paves the way for additional research in the context of adversarial matching.

As explained in Section 4.7.2, there remain several challenges associated with these data. These are left as future work. I explain at length the problem of incorrect labels; some ideas for dealing with this are to manually extract labels from profile information (see Section 4.3), use unsupervised methods, or manually label records. More could also be done in terms of scalability. With regards to the R package that I built, in order for it to have full utility, especially among forensic practitioners and investigators, it is important to extend its functionality to include the ability to input raw information that one could read off currently running marketplace webpages. Additionally, it would be beneficial to create an interface in which one could easily type in this information, if one is to expect practitioners to make use of such a tool. Finally, much more work remains in terms of estimating the weight of evidence.

To conclude, in the forensics literature, automatic methods in digital evidence have not been given as much attention as traditional pattern matching disciplines, such as fingerprints and firearms evidence. As far as I know, in the realm of anonymous marketplaces, manual investigative methods have been used with no attempt at automatic matching. This chapter demonstrates that this is not only possible but can be reasonably successful, despite inherent difficulties in the data. It almost impossible to achieve perfect error rates (for example, teams or franchises, or accounts belonging to the same distribution network, might have behavior that is hard to capture or predict), but it is definitely possible to make inroads into solving this problem.

Chapter 5

Concluding Thoughts

In this thesis, I explain the motivations for developing automatic methods for matching problems in forensics. An important one is the misuse of forensic evidence by practitioners, which has resulted (at least in part) in numerous wrongful convictions. This has been given much public attention, and has led to a push towards objective, automatic methods. Apart from this, such methods fulfill other objectives, and I explore some of these.

I then draw the link between matching problems in forensics, and record linkage problems in statistics and computer science. I propose a standardized framework to tackle matching problems in forensics, including suggestions for standardizing evaluation, and generating transitive closures. I apply this framework to two specific problems, developing automatic methods to match firearms and digital evidence. Motivations, methodology, contributions and future work relating to each of these topics are written about in great detail in Chapters 3 and 4. In terms of general future directions, I give some suggestions with respect to quantifying the weight of evidence, but overall this is not an emphasis in this thesis. Similarly, how to implement indexing methods in forensic problems is also a worthwhile future direction.

Obviously, good work has been and continues to be done in forensics, but methods are often developed in an ad-hoc and domain-specific manner. I hope that drawing the link to record linkage gives structure to forensic matching problems, and facilitates the generation of new ideas and consideration of issues that have not traditionally been a priority in forensics. For example, the step of generating a linked data set is a fairly standard consideration in record linkage, but as far as I know, has not been given much thought in the forensics literature. I give multiple reasons why this should deserve consideration. More generally, I hope to encourage more statistical thinking in the forensics field. Automatic methods have grown in importance in recent years and can be expected to continue to do so, and it is important to develop such methods in a principled manner.

From the record linkage perspective, I introduce a new and interesting set of problems to consider. Some unique characteristics are that forensic problems involve much more complex, unstructured data, and might have an adversarial component that is not typically present in traditional record linkage problems.

From both the forensics and record linkage perspectives, I have only scratched the surface of what the other has to offer, and I hope that others might find some of the ideas in this thesis useful, interesting, or worth pursuing further.

Bibliography

AFTE Criteria for Identification Committee (1992). Theory of identification, range of striae comparison reports, and modified glossary definitions – an AFTE criteria for identification committee report. *AFTE Journal*, 24(2):336–340. 5, 10, 27

Aliens, C. (2017). The darknet search engine 'grams' is shutting down. `https://www.deepdotweb.com/2017/12/15/darknet-search-engine-grams-shutting/`. Last accessed: May 25, 2019. 78

Baker, S. and Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255. 30, 65

Baldwin, D. P., Bajic, S. J., Morris, M., and Zamzow, D. (2014). A study of false-positive and false-negative error rates in cartridge case comparisons. Technical report, AMES LAB IA. 64

Bell, S., Sah, S., Albright, T. D., Gates, S. J., Denton, M. B., and Casadevall, A. (2018). A call for more science in forensic science. *Proceedings of the National Academy of Sciences*, 115(18):4541–4544. 1

Bien, J. and Tibshirani, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, 106 495:1075–1084. 21

Branwen, G. (2012). Tor dnm-related arrests, 2011-2015. `https://www.gwern.net/DNM-arrests`, accessed 2019-05-10. kyle hall thehater90. 95

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32. 7, 76, 84

Brein, C. (2005). Segmentation of cartridge cases based on illumination and focus series. In Said, A. and Apostolopoulos, J. G., editors, *Image and Video Communications and Processing 2005*, volume 5685 of *Proceedings of the SPIE*, pages 228–238. 29

Broséus, J., Rhumorbarbe, D., Mireault, C., Ouellette, V., Crispino, F., and Décary-Hétu, D. (2016). Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective. *Forensic Science International*, 264:7 – 14. Special Issue on the 7th European Academy of Forensic Science Conference. 75, 80

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510. 40

Bureau of Alcohol, Tobacco, Firearms and Explosives (2018). Fact sheet - national integrated ballistic information network. 27

Buskirk, J. V., Bruno, R., Dobbins, T., Breen, C., Burns, L., Naicker, S., and Roxburgh, A. (2017). The recovery of online drug markets following law enforcement and other disruptions. *Drug and Alcohol Dependence*, 173:159 – 162. 75

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698. 38

Champod, C., Biedermann, A., Vuille, J., Willis, S., and De Kinder, J. (2016). ENFSI guideline for evaluative reporting in forensic science, a primer for legal practitioners. 180:189–193. 10

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated. xiii, 3, 14, 15, 77

Christin, N. (2013). Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 213–224, New York, NY, USA. ACM. 77, 78, 97

Christin, N. (2017). An EU-focused analysis of drug supply on the alphabay marketplace. EMCDDA report for contract CT.17.SAT.0063.1.0. Available at `http://www.emcdda.europa.eu/system/files/attachments/6622/AlphaBay-final-paper.pdf`. 82

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836. 40

Collaborative Testing Services (2015). Firearms examination test no. 15-526 summary report. 12

Copas, J. B. and Hilton, F. J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):287–320. 18

Daugman, J. (2004). How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):21–30. 29

De Kinder, J., Tulleners, F., and Thiebaut, H. (2004). Reference ballistic imaging database performance. *Forensic Science International*, 140(2–3):207 – 215. 27

D'Errico, J. (2004). inpaint_nans. MATLAB Central File Exchange, `https://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint-nans` (accessed April 28, 2017). 41

DHS S&T – CSD (2019). Information marketplace for policy and analysis of cyber-risk & trust (IMPACT). Retrieved May 25, 2019, from `https://impactcybertrust.org`. 78

Dolliver, D. S. and Kenney, J. L. (2016). Characteristics of drug vendors on the tor network: A cryptomarket comparison. *Victims & Offenders*, 11(4):600–620. 75

Dutch National Police (2017). `http://politiepcvh42eav.onion/hansafaq.html`, accessed 2017-08-20. 75

Everitt, B., Landau, S., and Leese, M. (2001). Cluster analysis, 4th edn: Arnold. *London, UK*. 21

Federal Bureau of Investigation (2018). Operation disarray: Shining a light on the dark web. `https://www.fbi.gov/news/stories/operation-disarray-040318`, accessed: 2019-05-10. 7, 74

Federal Bureau of Investigation (2019). Operation sabotor. `https://www.fbi.gov/news/stories/j-code-operation-sabotor-032619`, accessed: 2019-05-10. 74

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210. 16, 19, 22, 23

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395. 5, 57

George, W. (2004). The validation of the brasscatcher portion of the NIBIN/IBIS system part two: Fingerprinting firearms reality or fantasy. *AFTE Journal*, 36(4):289 thru 296. St. Louis County Police, Crime Laboratory, Firearm & Tool Mark Identification Section, St. Louis County, Missouri. 45

Geradts, Z. J., Bijhold, J., Hermsen, R., and Murtagh, F. (2001). Image matching algorithms for breech face marks and firing pins in a database of spent cartridge cases of firearms. *Forensic Science International*, 119(1):97–106. 28, 29, 30

Gerules, G., Bhatia, S. K., and Jackson, D. E. (2013). A survey of image processing techniques and statistics for ballistic specimens in forensic science. *Science & Justice*, 53(2):236 – 250. 29

Hare, E., Hofmann, H., and Carriquiry, A. (2016). Automatic Matching of Bullet Land Impressions. *ArXiv e-prints*. 20, 29

Hepler, A. B., Saunders, C. P., Davis, L. J., and Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, 219(1):129 – 140. 11

ITV (2016). Man jailed for 'dark web' drug dealing. `https://www.itv.com/news/central/2016-02-29/man-jailed-for-dark-web-drug-dealing/`, accessed 2019-05-10. richierich. 95

Iyer, H. K. and Lund, S. P. (2017). Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research (NIST JRES)-*, 122(Journal of Research (NIST JRES)-). 11

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270. 16, 83

Kamalakannan, S., Mann, C. J., Bingham, P. R., Karnowski, T. P., and Gleason, S. S. (2011). Automatic firearm class identification from cartridge cases. In *Image Processing: Machine Vision Applications IV*, volume 7877 of *Proceedings of the SPIE*, page 78770P. 29

Kruithof, K., Aldridge, J., Hétu, D. D., Sim, M., Dujso, E., and Hoorens, S. (2016). *Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands*. RAND Corporation, Santa Monica, CA. 75

Kumar, S., Cheng, J., Leskovec, J., and Subrahmanian, V. (2017). An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 857–866, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 76

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. 16, 83

Li, D. G. (2003). Image processing for the positive identification of forensic ballistics specimens. In *Sixth International Conference of Information Fusion, 2003. Proceedings of the*, volume 2, pages 1494–1498. 29

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. 85

Lightstone, L. (2010). The potential for and persistence of subclass characteristics on the breech faces of SW40VE Smith & Wesson Sigma pistols. *AFTE Journal*, 42(4):308–322. 25, 34, 45, 63

Lilien, R. (2017). Applied research and development of a three-dimensional topography system for imaging and analysis of striated and impressed tool marks for firearm identification using gelsight. *Forensic Science Seminar*, 7(2):43–53. 28, 30, 32, 61, 64

Möser, M., Soska, K., Heilman, E., Lee, K., Heffan, H., Srivastava, S., Hogan, K., Hennessey, J., Miller, A., Narayanan, A., and Christin, N. (2018). An empirical analysis of traceability in the monero blockchain. In *Proc. PETS*, volume 3, Barcelona, Spain. 77

Murphy, H. (2019). A leading cause for wrongful convictions: Experts overstating forensic results. `https://www.nytimes.com/2019/04/20/us/wrongful-convictions-forensic-results.html`, accessed: 2019-04-28. 1

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC. 1

National Research Council, Division on Engineering and Physical Sci, National Materials Advisory Board (2008). *Ballistic Imaging*. National Academies Press. 48

Ormsby, E. (2016). *Silk Road: insights from interviews with users and vendors*, chapter 6, page 65. The internet and drug markets (European Monitoring Centre for Drugs and Drug Addiction: Insights 21). Publications Office of the European Union, Luxembourg. 84

Ott, D., Thompson, R., and Song, J. (2017). Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests. *Forensic Science International*, 271:98 – 106. xiv, 28, 32, 45, 61, 62, 64, 71

Park, S. (2018). Learning algorithms for forensic science applications. 11

Popper, N. (2015). The tax sleuth who took down a drug lord. `https://www.nytimes.com/2015/12/27/business/dealbook/the-unsung-tax-agent-who-put-a-face-on-the-silk-road.html?mcubz=1`, accessed 2017-08-20. 74

Popper, N. (2017). Opioid dealers embrace the dark web to send deadly drugs by mail. `https://www.nytimes.com/2017/06/10/business/dealbook/opioid-dark-web-drug-overdose.html`, accessed: 2017-08-20. 74

President's Council of Advisors on Science and Technology (2016). *Report to the President on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Executive Office of the President, Washington, D.C. 1, 27, 64

Riva, F. and Champod, C. (2014). Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases. *Journal of Forensic Sciences*, 59(3):637–647. 11, 28, 29, 30, 31, 32, 33, 44

Roberge, D. and Beauchamp, A. (2006). The use of BulletTRAX-3D in a study of consecutively manufactured barrels. *AFTE Journal*, 38(2):166. 20

Roth, J., Carriveau, A., Liu, X., and Jain, A. K. (2015). Learning-based ballistic breech face impression image matching. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Arlington, VA. `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7358774&isnumber=7358743` (accessed April 28, 2017). 28, 29, 30, 31, 32, 36, 40, 54, 64

RStudio, Inc (2013). *Easy web applications in R.* URL: `http://www.rstudio.com/shiny/`. 65, 99

Skinner, C. J. (2007). The probability of identification: applying ideas from forensic statistics to disclosure risk assessment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):195–212. 18

Song, J. (2013). Proposed "NIST ballistics identification system (NBIS)" based on 3D topography measurements on correlation cells. *AFTE Journal*, 45(2):184–194. 6, 22, 28, 31, 61, 64

Song, J. (2015). Proposed "Congruent Matching Cells(CMC)" method for ballistic identification and error rate estimation. *AFTE Journal*, 47(3):177–185. 28, 30, 33

Song, J., Chu, W., Vorburger, T. V., Thompson, R., Renegar, T. B., Zheng, A., Yen, J., Silver, R., and Ols, M. (2012). Development of ballistics identification–from image comparison to topography measurement in surface metrology. *Measurement Science and Technology*, 23(5):054010. 34

Song, J., Vorburger, T. V., Chu, W., Yen, J., Soons, J. A., Ott, D. B., and Zhang, N. F. (2018). Estimating error rates for firearm evidence identifications in forensic science. *Forensic Science International*, 284:15 – 32. 28, 32, 56, 61

Soska, K. and Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 33–48, Berkeley, CA, USA. USENIX Association. 6, 75, 77, 78, 80, 82, 83, 97

Tai, X. H. (2018). Record linkage and matching problems in forensics. In *2018 IEEE International Conference on Data Mining Workshops, ICDM Workshops, Singapore, Singapore, November 17-20, 2018*, pages 510–517. 2

Tai, X. H. and Eddy, W. F. (2018). A fully automatic method for comparing cartridge case images,. *Journal of Forensic Sciences*, 63(2):440–448. xiv, 2, 36, 44, 54, 55

Tai, X. H., Soska, K., and Christin, N. (2019). Adversarial matching of dark net market vendor accounts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, New York, NY, USA. ACM. xi, 2, 78, 79

Thumwarin, P. (2008). An automatic system for firearm identification. In *2008 International Symposium on Communications and Information Technologies*, pages 100–103. 28, 29, 31

Tong, M., Song, J., Chu, W., and Thompson, R. M. (2014). Fired cartridge case identification using optical images and the Congruent Matching Cells (CMC) method. *Journal of Research of the National Institute of Standards and Technology*, 119:575–582. 54, 64

Tunali, E., Leloglu, U., and Sakarya, U. (2009). Method for automatic region segmentation on cartridge case base and selection of the best mark region for cartridge case comparison. WO Patent App. PCT/IB2009/051,609. 29

Ultra Electronics Forensic Technology (2018). https://www.ultra-forensictechnology.com/wp-content/uploads/2018/05/IBIS-TRAX-HD3D_EN.pdf, accessed 2019-03-02. 28

United States District Court, District of New Jersey (2014). United states district court district of new jersey. https://antilop.cc/sr/files/Tormail_Sean_Roberson_Complaint.pdf, accessed 2019-05-10. goldcard slacker. 95

United States District Court, District of North Dakota (2018). Indictment. https://www.justice.gov/opa/press-release/file/1058211/download, accessed 2019-05-10. goldmountain. 95

United States District Court, District of Oregon (2013). Indictment. https://www.gwern.net/docs/sr/2013-12-10-hammertime-indictment.pdf, accessed 2019-05-10. drbechen. 95

United States District Court, Eastern District of California (2016). Affidavit of matthew larsen. https://www.justice.gov/usao-edca/file/836576/download, accessed 2017-08-20. caliconnect. 7, 74, 75, 95

United States District Court, Eastern District of California (2017). Criminal complaint. http://ia601509.us.archive.org/10/items/gov.uscourts.caed.320736/gov.uscourts.caed.320736.11.0.pdf, accessed 2017-08-20. PureFireMeds –¿ HumboldtFarms. 74, 95

United States District Court, Eastern District of New York (2016). Affidavit in support of removal to the eastern district of california. https://regmedia.co.uk/2016/08/12/almashwali_arrest.pdf, accessed 2017-08-20. dark51. 7, 75, 95

United States District Court, Eastern District of New York (2017). Trafficker of fentanyl, heroin, and methamphetamine on dark web marketplace alphabay pleads guilty to drug distribution charge. https://www.justice.gov/usao-edca/pr/trafficker-fentanyl-heroin-and-methamphetamine-dark-web-marketplace-alphabay-pleads, accessed 2019-05-10. blime-sub. 95

United States District Court, Middle District of Florida (2014). Plea agreement. https://www.courtlistener.com/recap/gov.uscourts.flmd.297482.3.0.pdf, accessed 2019-05-10. undergroundsyndicate. 95

United States District Court, Middle District of Florida (2016). Indictment. https://www.dropbox.com/s/tz9z5jly649l7lk/flmd-047016068221.pdf, accessed 2019-05-10. darkexpresso. 95

United States District Court, Western District of Pennsylvania (2017). Affidavit in support of complaint. https://www.scribd.com/document/353858458/United-States-of-America-v-HENRY-KOFFIE-a-k-a-NarcoBoss-Criminal-Complaint, accessed 2019-05-10. narcoboss. 95

van Wegberg, R., Tajalizadehkhoob, S., Soska, K., Akyazi, U., Hernandez Ganan, C., Klievink, B., Christin, N., and van Eeten, M. (2018). Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *Proc. USENIX Security*, Baltimore, MD. 77

Ventura, S. L., Nugent, R., and Fuchs, E. R. (2015). Seeing the non-stars: (some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, 44(9):1672 – 1701. The New Data Frontier. 20

Vorburger, T., Yen, J., Bachrach, B., Renegar, T., Filliben, J., Ma, L., Rhee, H., Zheng, A., Song, J., Riley, M., Foreman, C., and Ballou, S. (2007). Surface topography analysis for a feasibility assessment of a National Ballistics Imaging Database. Technical Report NISTIR 7362, National Institute of Standards and Technology, Gaithersburg, MD. 29, 30, 32, 36, 40, 45, 54, 56, 61, 64, 65

Wang, X., Peng, P., Wang, C., and Wang, G. (2018). You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, pages 431–442, New York, NY, USA. ACM. 75

Weller, T. J., Zheng, A., Thompson, R., and Tulleners, F. (2012). Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides. *Journal of Forensic Sciences*, 57(4):912–917. 61, 64

Winkler, W. E. (2000). Using the em algorithm for weight computation in the felligi-sunter model of record linkage. 17

Zeifman, L. E. (2014). *A New Parametric Model for the Point Spread Function (PSF) and Its Application to Hubble Space Telescope Data [dissertation]*. Carnegie Mellon University, Pittsburgh, PA. 39

Zhou, J., Xin, L.-p., Rong, G., and Zhang, D. (2001). Algorithm of automatic cartridge identification. *Optical Engineering*, 40(12):2860–2865. 29

Appendix

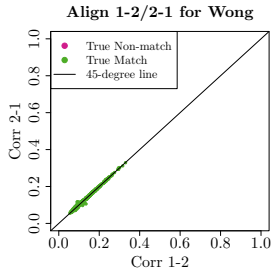
Appendix A

Aligning I_1 to I_2 versus I_2 to I_1

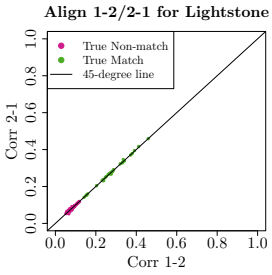
A.1 2D

A.2 3D

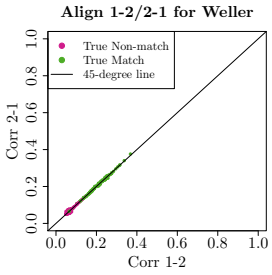
Figure A.1: Similarity scores derived from aligning image 1 to image 2, versus image 2 to image 1, for matches and non-matches by study in 2D.



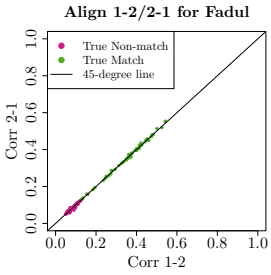
(a) Cary Wong (Persistence)



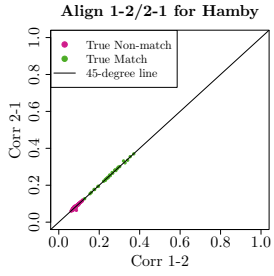
(b) Lightstone (consecutively manufactured)



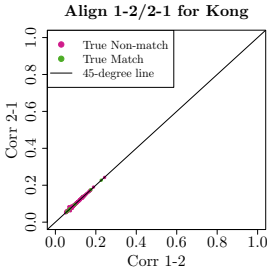
(c) Weller (consecutively manufactured)



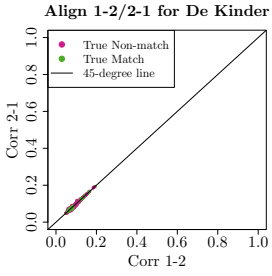
(d) Fadul (consecutively manufactured)



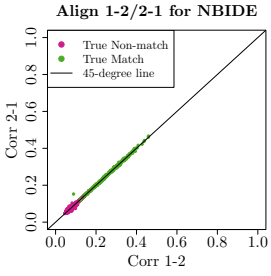
(e) Hamby (consecutively manufactured)



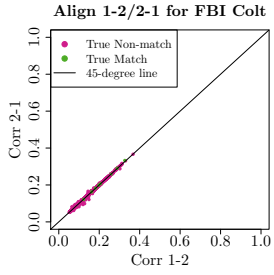
(f) Kong



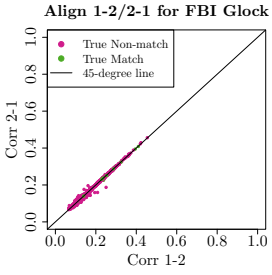
(g) De Kinder



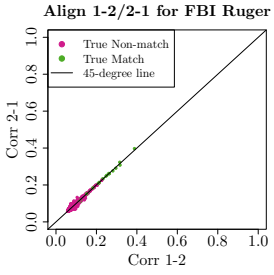
(h) NBIDE



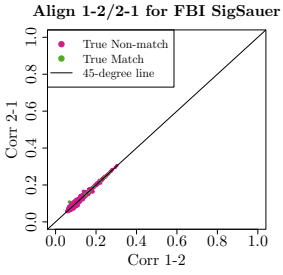
(i) FBI Colt



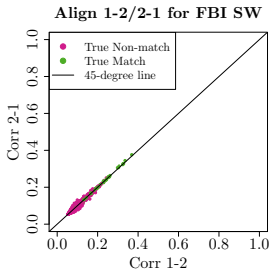
(j) FBI Glock



(k) FBI Ruger

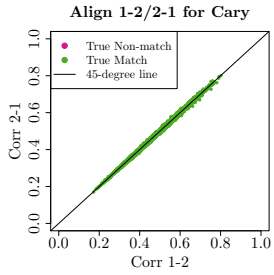


(l) FBI Sig Sauer

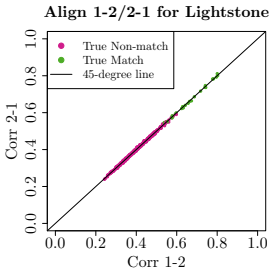


(m) FBI Smith & Wesson

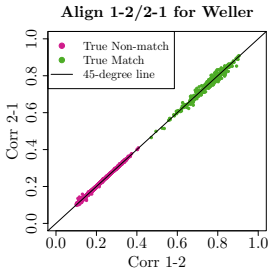
Figure A.2: Similarity scores derived from aligning image 1 to image 2, versus image 2 to image 1, for matches and non-matches by study in 3D.



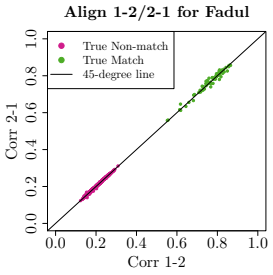
(a) Cary Wong (Persistence)



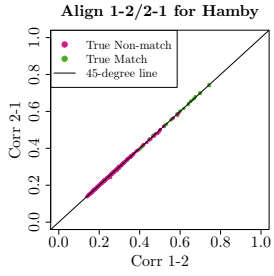
(b) Lightstone (consecutively manufactured)



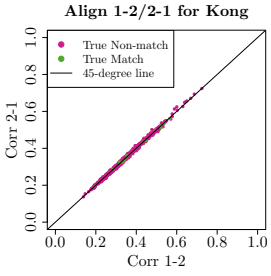
(c) Weller (consecutively manufactured)



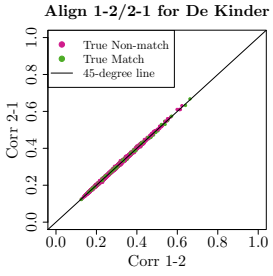
(d) Fadul (consecutively manufactured)



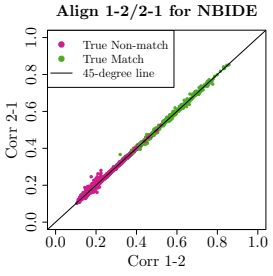
(e) Hamby (consecutively manufactured)



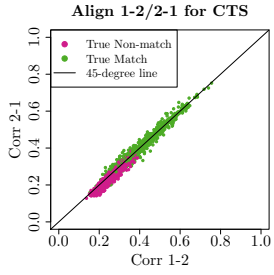
(f) Kong



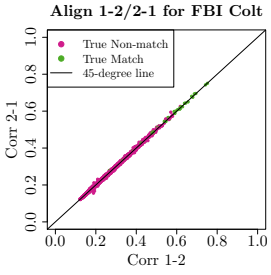
(g) De Kinder



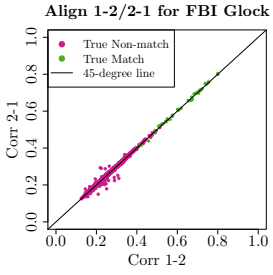
(h) NBIDE



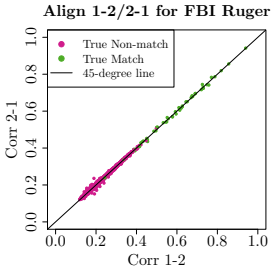
(i) CTS



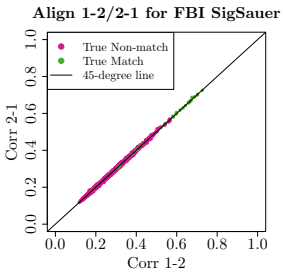
(j) FBI Colt



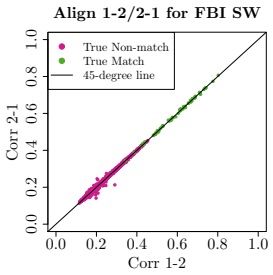
(k) FBI Glock



(l) FBI Ruger



(m) FBI Sig Sauer



(n) FBI Smith & Wesson

Appendix B

Shiny Interface for Manual Labeling of Marketplaces Accounts

Figure B.1: Shiny interface for manual labeling. Information about pairs of accounts is displayed, including their IDs, marketplaces, the computed similarity measures, as well as additional information such as their full item listings and history of profile descriptions.