Forensic Data Matching Problems

Xiao Hui Tai
Advisor: William F. Eddy
Committee: Nicolas Christin, Brian W. Junker, Joseph B. Kadane, Rebecca Nugent

November 15, 2017

Abstract

In forensics, evidence such as DNA, fingerprints, bullets and cartridge cases, shoeprints or digital evidence is often compared, to infer if they come from the same or different sources. This helps to generate leads through database searches, where information from different investigations can be combined, if pieces of evidence are judged to have come 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 can also be used as testimony in courts. We demonstrate how such matching problems fit into the record linkage framework commonly used in statistics and computer science, and explore how the latter can inform the former. In particular, we consider two forensic matching problems. The first is firearms identification, where cartridge cases are compared to infer if they were fired from the same gun. We have developed an open-source, fully automatic method to compare 2D optical images, and will extend this to 3D topographies. We propose computing "signatures" instead of using pixel-level information, so that database searches are feasible. Comparisons should produce some measure of the weight of evidence, such as a likelihood ratio, that can be reported in courts. Finally, we plan to compare our error rates to that of firearms examiners.

The second problem is matching accounts on anonymous marketplaces, to infer if they belong to the same seller. The goal is to generate accurate investigative leads, even if some attempt is made by sellers to obfuscate their identity. We first match accounts without relying solely on PGP keys, which can be thought of identifiers that may be unavailable, inaccurate, or spoofed. Ideally, the matching process should account for errors in these identifiers, which might include some types of deceptive use, such as users that create multiple identities. To improve on the current model we propose several alternatives, for example unsupervised methods or an active learning approach. The latter will also give us a set of "ground truth" labels that we can use for evaluating different approaches. Finally, we consider low-dimensional representations of the accounts, which mirrors the proposed work on cartridges. By using a common, well-established framework, we demonstrate how forensic matching problems can be tackled in general, in a more principled manner.

1 Introduction

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 (Peter Christen 2012). When dealing with duplicate entries in a single database, it is more commonly known as deduplication or duplicate detection. 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 databases, and so forth.

Forensic evidence refers to DNA, fingerprints, bullets and cartridge cases, shoeprints, digital evidence etc. left behind when a crime is committed. The underlying assumption is that the perpetrator of the crime, or tools he/she 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 samples are compared, to infer if they came from the same source. The comparison can be thought of from two perspectives: a database search used to generate investigative leads in a one-to-many comparison, versus an evaluation or "identification" in a one-to-one comparison, for example when evidence is compared with a sample taken from a suspect. These are closely related however, since a one-to-many comparison could be treated as repeating a one-to-one comparison many times. For many of the evidence types, automated methods for comparisons have been developed, either for database searches, objective identification, or both.

At first glance, record linkage and forensic matching might seem like separate problems, but we demonstrate a correspondence between the two – approaches traditionally used in forensic matching fit into the framework of statistical record linkage problems. By thinking about forensics problems in the context of record linkage, we immediately have well-developed frameworks and tools at our disposal.

We apply this sort of thinking to two forensic matching problems. The first is firearms identification. A gun is thought to leave unique marks on bullets and cartridge cases, and if these are retrieved from crime scenes, they can be compared to infer if they were fired from the same gun. Firearms identification has a long history, and automatic comparison methods have been developed by engineers and scientists both in industry and in academia. In the current system in the United States, a semi-automatic database search is performed to generate leads, while the final judgement of a match or non-match still relies on a human examiner. This subjective nature and lack of understanding of associated error rates has come under scrutiny in recent years (see e.g. President's Council of Advisors on Science and Technology 2016). In this part of the work, we focus on cartridge cases (as opposed to bullets). We describe how approaches that have been developed in an ad hoc manner fall under the record linkage framework, and develop open-source, fully automatic methods for matching, with the goal of performing both a database search and producing associated objective measures of uncertainty that can be taken to court.

The second problem is matching seller accounts on online anonymous marketplaces. These marketplaces provide anonymity protections to buyers and sellers, and are associated with illegal activities such as the sale of drugs and stolen personal data. This is a relatively recent phenomenon, with the first such marketplace being launched in February 2011 (Soska and Christin 2015). Due to the anonymous nature of such marketplaces, tracking down sellers is difficult and many cases are still being investigated. As far as we know, law enforcement seeks to connect seller accounts on the various marketplaces, but there does not currently appear to be an automatic way of doing so on a large scale, unlike in the other forensic disciplines. So far agents have relied primarily on "old-school" investigation methods, and in this part of the work, we apply record linkage techniques to match accounts automatically and on a large scale. We do so with the initial goal of generating investigative leads, as opposed to producing measures of uncertainty for court testimony.

In this thesis we demonstrate that it is possible to use a common framework and similar techniques to tackle these two problems, and forensic matching problems in general. The rest of the proposal is organized as follows. Section 2 explores the connection between record linkage and forensic matching problems. Section 3 and goes into detail about firearms identification, including background and literature, current and future work. Section 4 does the same for matching accounts on anonymous marketplaces. Finally, Section 5 summarizes our plans.

2 Record linkage and forensic matching

Record linkage involves the pairwise comparison of records to infer if they belong to the same or different real-world entities. A standard framework for matching is in Figure 1. This framework is adapted from Peter Christen (2012) and Ventura, Nugent, and Fuchs (2015).

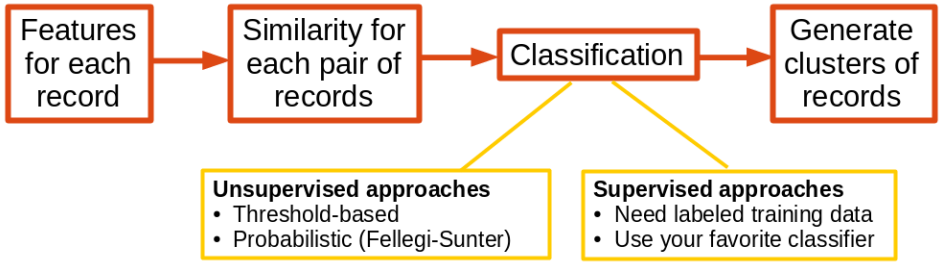

Figure 1: Standard framework for statistical record linkage problems. We begin with n records to be linked or deduplicated. Each of these records have some set of features. Next we generate pairwise comparisons from these n records, where each pairwise comparison consists of one or more similarity measures. Here an indexing scheme is sometimes used, where comparisons are generated, for example, only if the pair fulfills some criteria, so that not all $\frac{n(n-1)}{2}$ pairwise comparisons have to be computed. Next the pairwise comparisons are classified into matches and non-matches. In the final step we produce clusters of records from these comparisons, such that transitivity is preserved.

In forensic matching, pairwise comparisons are done for DNA evidence, fingerprints, bullets and cartridge cases, glass fragments, and so forth, to infer if they come from the same or different sources. Many different automated or semi-automated approaches have been developed for such purposes by people working in different fields, yet the correspondence between record linkage and forensic matching problems has been noted by statisticians since as early as 1990 (see e.g. Copas and Hilton 1990; Skinner 2007). The main conceptual difference is that in record linkage we are primarily concerned about disambiguating a data set, while in forensic matching problems we are interested in comparing a new sample against some database or existing sample. Practically however, for the latter we often start with some database which is used to develop classification methods. We can think of this as a record linkage problem where we disambiguate the database, and then use the same methods for comparisons of the new sample with the database.

The following are two examples of forensic matching problems on DNA and bullets, with descriptions of how they fit into the framework in Figure 1. A detailed discussion of work in cartridge cases is in Section 3.1.

DNA DNA profiling is often described as the gold standard of forensic matching methods, because there is a scientific basis for comparisons as well as computation of likelihood ratios. The US government maintains the Combined DNA Index System (CODIS), which is a database containing DNA profiles from crime scenes as well as known offenders. Each DNA profile consists of information from 13 (or more) locations on the DNA. By biological theory, each of these are independent. If all of them match exactly, a comparison is reported as a match, together with a likelihood ratio that is estimated as

$$LR = \frac{\mathbb{P}[\text{Evidence}|\text{Match}]}{\mathbb{P}[\text{Evidence}|\text{Non-match}]}. \tag{1}$$

The numerator is usually taken to be 1, and the denominator is estimated using known frequencies of DNA sequences in the relevant reference population. The likelihood ratio can be interpreted as the number of times more likely we are to observe the evidence if the profiles match than if they do not, and hence quantifies the weight of evidence. Using the framework in Figure 1, each DNA sample (record) is summarized using 13 features which correspond to the 13 locations on the DNA. The similarity for each pair of records is a vector of length 13, with each entry being a binary value for whether the feature is a match, and a threshold-based classification method might require having all 13 entries be 1. Mathematically, we can describe the similarity metric as the Hamming distance, $d_H(x, y) = \frac{1}{13}\sum_{i=1}^{13} I(x_i \neq y_i)$, and we use an unsupervised threshold-based classification approach, with a cutoff of 1. Since this is an exact matching scheme, transitivity is automatically preserved. Estimating the likelihood ratio is then an additional step, and the methodology used is essentially the same as in Fellegi and Sunter (1969), in the case where population frequencies are known.

Bullets When a bullet is fired, rifling, manufacturing defects, and impurities in the barrel create striation marks on the bullet, and this forms the basis for bullet identification. Unlike DNA matching, there is no well-understood scientific basis upon which comparisons can be made. Automated methods have been developed by various groups, including manufacturers and academia. These methods generally extract a profile or signature from the bullet lands (the surface between two bullet grooves), followed by some pre-processing to remove noise and so forth. This profile serves as the features for each record. Various similarity metrics have been developed, such as the correlation between aligned signatures, maximum number of consecutive matching striae and average Euclidean vertical distance between surface measurements of aligned signatures. For classification, both unsupervised threshold-based (e.g. Roberge and Beauchamp 2006) and supervised methods (e.g. Hare, Hofmann, and Carriquiry 2016) have been used.

As described earlier, using a record linkage framework allows us to take advantage of tools that have been developed, without having to reinvent the wheel. In the case of anonymous marketplaces, we directly apply associated techniques, while for cartridge cases and other areas where there are existing well-developed bodies of literature, we might think about improvements that can be made, for example:

- **Indexing** Indexing aims to reduce the quadratic complexity of the data matching process through the use of data structures to efficiently generate candidate record pairs that likely correspond to matches (Peter Christen 2012). Blocking is a straightforward approach to indexing, where records are grouped into blocks based on some similarity criteria, and pairwise comparisons are generated only for records in the same block. For forensic matching problems, as the sizes of databases grow, reducing computational complexity becomes increasingly important, particularly since these data are often images which are high-dimensional objects. Indexing techniques in the record linkage literature could be useful in this regard, and are explored in Sections 3.3 and 4.3.
- **Fellegi-Sunter, cutoffs, and weight of evidence** Fellegi and Sunter (1969) proposed a probabilistic framework for assigning matches, and this subsequently gained widespread popularity and is often considered the standard model for 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 $(a, b) \in M$:

$$\frac{\mathbb{P}[\gamma_{ab} = g|(a,b) \in M]}{\mathbb{P}[\gamma_{ab} = g|(a,b) \in U]}, \tag{2}$$

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. The parameters involved in the likelihood ratio are commonly estimated using the EM algorithm (Winkler 2000). Adapting such a framework for forensic matching has two distinct advantages. First we notice that Equation 2 is essentially a specific formulation of Equation 1. The likelihood ratio has become widely accepted as a measure for the strength of forensic evidence. Particularly in Europe, guidelines for forensic laboratories now recommend the reporting of likelihood ratios (Champod et al. 2016). However there is no consensus on how these likelihood ratios should be estimated. For example in DNA, a generative model is used to determine the probabilities of observing particular DNA sequences. As mentioned, this corresponds to the first method for calculating weights in Fellegi and Sunter (1969), given known population frequencies. In other forensic fields similar calculations are difficult because of the lack of knowledge of a scientific basis in which say, bullet striations are produced. Instead score-based likelihood ratios have been suggested, which look at distributions of similarity scores for known matching and non-matching pairs (see e.g. Hepler et al. 2012). Further investigation could reveal if methods from the record linkage literature could be relevant in a forensic context. The second advantage has to do with the cutoffs used which control for error rates (false positives and false negatives). This could be an improvement over threshold-based methods that have been suggested (e.g. Song 2013), which use somewhat arbitrary cutoffs without a proper quantification of error rates.

- **Generating clusters of records** In the last step in Figure 1, we generate clusters of records from pairwise similarities. For example, if A is similar to B and B is similar to C, then A, B and C might all be assigned the same cluster. If the focus of forensic matching is on producing leads through database searches, generating clusters within the database could be an easy way to generate additional leads: using the same example, if the new sample is similar to A, then we might also consider B and C to be leads.

3 Firearms identification

3.1 Background and literature

Firing a gun leaves marks on the bottom surface of the cartridge case. This mechanism is illustrated in Figure 2. The bottom surface of the cartridge is in contact with the breech block of the gun. 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 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 in Figure 2. Breechface marks 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).

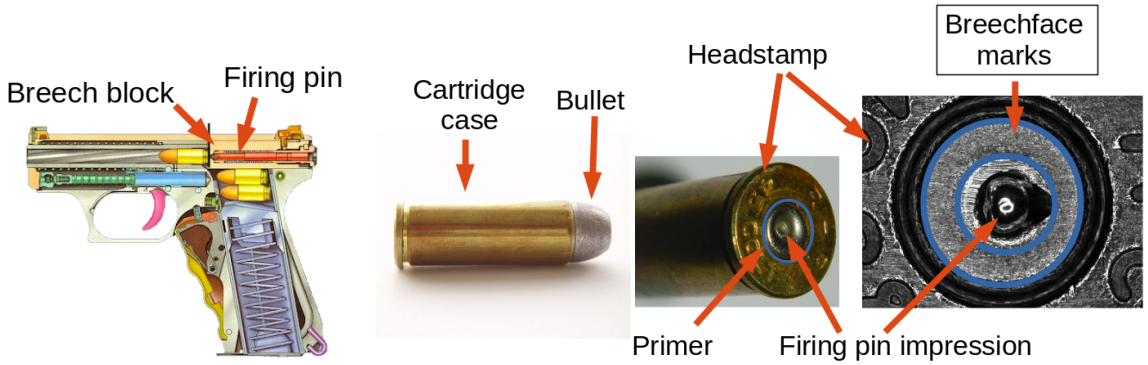

Figure 2: On the far left we have a gun that is about to be fired, showing the internal parts. Next we have a cartridge before firing, and the bottom surface of a cartridge case after firing. The firing pin impression is clearly visible as the "hole" in the center of the primer surface, while the breechface marks lie on the remaining primer surface. On the far right is an image of such a bottom surface, taken using a reflectance microscope. The breechface marks are in the gray donut shaped area in between the two marked rings, and are the focus of our analysis. (Source: https://commons.wikimedia.org/wiki/File:Rueckstoss-theorie.png?uselang=en, https://commons.wikimedia.org/wiki/File:45_Colt_-_1.jpg and https://commons.wikimedia.org/wiki/File:45er_3.jpg, retrieved 4/23/2017.)

Law enforcement officers routinely collect guns and cartridge cases from crime scenes, because of their potential usefulness in investigations. In current practice, cartridge cases are entered into a national database called the National Integrated Ballistics Information Network (NIBIN), through a computer-based platform called the Integrated Ballistics Identification System (IBIS), which was developed and is maintained by Ultra Electronics Forensic Technology (FTI). This platform captures an image of the "new" 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.

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) 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 estimated. They suggested two directions for the path forward. The first is to "continue to improve firearms analysis as a subjective method," and the second is to "convert firearms analysis from a subjective method to an objective method," through the use of automated methods and image-analysis algorithms.

There have been efforts by various groups, both commercial and academic, in line with this second recommendation. In the remainder of this section, we describe desirable features of methods, followed by specifics of commonly used techniques. We review related work in terms of these. Specifically, we cover methods by manufacturers (e.g. FTI, Cadre Forensics) and in academia (Roth et al. (2015);

Geradts et al. (2001); Thumwarin (2008); Riva and Champod (2014); NIST, e.g. Vorburger et al. (2007) and Song (2013)). As should be clear, different methods possess various subsets of the following desirable features, but none have all of them.

3.1.1 Desirable features of methods

Fully automatic In light of the criticism of subjectivity, methods should as much as possible operate without human input. We do acknowledge the possibility of algorithmic bias, but a fully automatic algorithm still has the advantage of producing replicable results. In the current NIBIN system, image acquisition involves manual corrections by firearms examiners. This subjective nature was highlighted in De Kinder, Tulleners, and Thiebaut (2004): breechface positioning (the outer circle in Figure 2) was manually corrected 18.5% of the time. In other work, all the methods except Riva and Champod (2014) similarly involve a manual selection of breechface marks. Riva and Champod (2014) requires users to first manually position a plane passing through the primer surface, after which the program computes a normal vector through the surface to select the breechface area automatically.

Open source Another shortcoming of the current NIBIN system is that the algorithms used in the database search step are proprietary, and proper testing is difficult. Earlier studies have pointed out high miss rates (De Kinder, Tulleners, and Thiebaut 2004). This means that when examiners testify as to the "certainty" of identifications, it is impossible to tell if there were in fact other cartridge cases that were even more similar, that were missed by the search. As for other work, to our knowledge, none of the methods are open source.

Produce a measure of the weight of evidence The importance of this was described in Section 2. The current system relies on examiners to make conclusions, and does not produce any measure of weight of evidence. NIST and Riva and Champod (2014) have proposed score-based likelihood ratios (also described in Section 2), which use the distribution of similarity scores as opposed to the features, for the "Evidence" portion of Equation 1. NIST uses a parametric distribution applicable to all guns (Song 2015), while Riva and Champod (2014) proposed using case-specific distributions for each questioned cartridge case and sample being compared to. These are to be determined empirically by conducting test fires for each comparison.

Scalable As previously described, there are two perspectives for comparisons, a database search involving a large number of images (there are millions in NIBIN), and evaluation or identification which is currently conducted by firearms examiners. Developing separate methods for each of these tasks has some advantages: the search can be optimized for speed, while the identification focuses on accuracy and meaningful similarity scores and measures of uncertainty. This is the approach taken by FTI (Ultra Electronics Forensic Technology, n.d.) and Geradts et al. (2001), who focus on database search, as well as NIST (Song 2015) and Riva and Champod (2014), who focus on identification. As for work by other groups, it is unclear if the distinction was made. Generally, there seems to be little focus on methods suitable for identification that are scalable and therefore usable for database searches.

Suitable for both 2D and 3D 2D optical (grayscale) images are taken using reflectance microscopes, and have been used in NIBIN systems for almost two decades. 3D topographies use confocal microscopes which measure surface contours directly. NIST has advocated the latter because of their insensitivity to lighting conditions (Song et al. 2012). There remains interest in 2D images, however, because of the large amount of data that has been collected over the years, cost of 3D

equipment, as well as a lack of validated methods for 3D data. Hence it is important for methods to be developed for both 2D optical images and 3D topographies. FTI and NIST have worked on both; Roth, Geradts, and Thumwarin use 2D images, while Cadre and Riva use 3D topographies exclusively.

3.1.2 Specific techniques

Next we discuss the steps common to the methods. Generally, they are as follows, and it is clear from this presentation they fit into the framework in Figure 1.

Pre-processing A comprehensive review of imaging systems and processing techniques is in Gerules, Bhatia, and Jackson (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, Leloglu, and Sakarya 2009; Kamalakannan et al. 2011). Pixels that are judged to be outliers are often removed and interpolated (e.g. Vorburger et al. 2007; Roth et al. 2015). Filters are used to smooth the image, and highlight individual as opposed to class characteristics (e.g. 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), as well as wavelet transforms (Geradts et al. 2001).

Signatures or features Most methods use the pixel values after pre-processing directly, but some groups generate "signatures" or features that are used instead. FTI's search algorithm uses "small" and "big signatures" (National Research Council, Division on Engineering and Physical Sci, National Materials Advisory Board 2008). Cadre extracts "corner-like textured regions... which correspond to the same types of ridges, ... that a trained firearms examiner would identify." (Cadre Forensics, n.d.) Roth et al. (2015) generate local cells at various locations and of different sizes. Thumwarin (2008) computes absolute values of Fourier coefficients. Geradts et al. (2001) uses the Kanade Lucas Tomasi (KLT) equation, used for image tracking problems, which extracts features which correspond to prominent marks, such as strong edges. Their positions form the signature for each image.

Similarity metrics For methods that use pixel-level information, the following are examples of similarity metrics. These are usually accompanied by some means of alignment, e.g. maximizing over possible rotations and translations.

- CCF_{max}: The CCF_{max} or maximum cross-correlation function, 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). The cross-correlation between two images I_1 and I_2 (or more generally, 2-dimensional matrices), is computed as

$$CCF_{I_1,I_2}(k,l) = \frac{\sum_{i,j} I_1(i,j)I_2(i+k,j+l)}{\sqrt{\sum_{i,j} I_1(i,j)^2}\sqrt{\sum_{i,j} I_2(i,j)^2}}, \tag{3}$$

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

- Metrics from difference image: First compute the difference between corresponding points in aligned images, $D(i,j) = I_1(i,j) - I_2(i,j)$, for each i, j. Then 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.

For methods that involve features, Cadre uses metrics such as the number of matched features and the average difference in feature appearance. Roth et al. (2015) use a function of the correlation between each local cell. Thumwarin (2008) uses Euclidean distance between the feature vectors in a clustering method. Geradts et al. (2001)'s KLT method uses the number of points that are matched.

Classification Both unsupervised and supervised methods have been used. 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, and 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 Current work

We have developed an open-source, fully automated method for comparison of 2D optical images of breechface marks. We briefly describe our work in terms of the framework in Figure 1: we use 803 images as the records. The features for each record are the individual pixel values in each image. We compute all $803 * 802$ or 644,006 pairwise comparisons; we do not divide by two because for each comparison, one image needs to be registered (rotated and translated) to the other, and this operation is not necessarily commutative. We then compute a single similarity score for each pair of records in an unsupervised approach.

3.2.1 Data

Although a national database of firearms data exists (NIBIN), these data are not publicly available. Instead, we use data from NIST's Ballistics Toolmark Research Database (https://tsapps.nist.gov/NRBTD), an open-access research database of bullet and cartridge case toolmark data. The database contains images originating from studies conducted by various groups in the firearm and toolmark community. These were imaged at NIST using a Leica FS M 2D reflectance microscope. The magnification was 2X with a lateral resolution of $2.53\mu m$, producing 2592×1944 pixel, 256-grayscale PNG images. We analyze all the data sets available on 2/15/2017. There are a total of 803 images. Gun manufacturers include Glock, Hi-Point, Ruger, Sig Sauer, and Smith & Wesson, and ammunition brands include CCI, Federal, PMC, Remington, Speer, Wolf and Winchester. The various studies include those involving consecutively manufactured pistol slides (the pistol slide contains the breech block which is responsible for breechface marks, and it was hypothesized that slides manufactured consecutively would make cartridge cases harder to differentiate), a large number of firings (termed persistence studies because they investigate the persistence of marks), as well as different makes and models of guns and ammunition. A full description of the data, broken down by study, is included in Appendix A. Apart from 2D images, 3D topographies are also available.

3.2.2 Methodology

The general framework is as described in Figure 1 and Section 3.2, but because of the nature of the data, various pre-processing steps are needed in order to generate usable features for each record. Generating similarities for each pair of records also becomes a non-trivial task. In Tai and Eddy (in press), we propose a fully automated method for performing all of these steps. These are listed below. We first pre-process each image (Steps 1-4), and generate the similarity score in Step 5. Step 6 computes an empirical "random match probability," which relates to the denominator of the likelihood ratio in Equation 1. These steps build on methodology published in Vorburger et al. (2007) and implemented in Roth et al. (2015). The main improvements are in steps 1, 3 and 6.

1. Automatically select breechface marks by removing the non-primer areas and the firing pin impression
2. Level image; to adjust for non-uniform lighting caused by the surface being tilted on a plane
3. Remove circular symmetry; to adjust for non-uniform lighting caused by the surface having differences in depth that are circular in nature
4. Outlier removal and filtering to highlight certain features
5. Maximize correlation by translations and rotations
6. Compute the probability of obtaining a higher score by chance given a known database.

This method is implemented in an R software package at https://github.com/xhtai/cartridges, where each step is described in detail, and the implementations are available. We describe Steps 5 and 6 in more detail below.

Similarity Score The similarity score that we use is the CCF_{max}, as described in Section 3.1. Let I_1 and I_2 be $m \times m$ square images (we zero-pad the images so that they are square and of the same size). We then consider lags of $-\lfloor \frac{m}{2} \rfloor \leq k, l \leq \lfloor \frac{m}{2} \rfloor$, and store the largest cross-correlation value, i.e.

$$\max_{-\lfloor \frac{m}{2} \rfloor \leq k,l \leq \lfloor \frac{m}{2} \rfloor} CCF_{I_1,I_2}(k,l). \tag{4}$$

Now, to search over different rotation angles, following Roth et al. (2015), we rotate one of the two images at angles 2.5° apart $(-177.5°, -175°, ..., 180°)$, to find the rotation angle θ' corresponding to the maximum correlation. Around θ', we consider rotation angles .5° apart, to obtain the maximum correlation between the two images, or CCF_{max}, over integer-valued translations and many rotation angles.

Weight of Evidence Here we focus on the denominator of the likelihood ratio (leaving the numerator for future work) and use a score-based approach. This involves the distribution of CCF_{max} values for non-matches, and we are interested in the probability of obtaining a higher score if the pair is a non-match, in other words by chance. Assuming all non-match CCF_{max} values are drawn from the same distribution, then given such a distribution, computing the right tail proportion would give us the required probability. Without such an assumption, and without access to a theoretical distribution of CCF_{max} values, we instead use a known database to make an empirical calculation. This can be applied to all guns, and is based on the assumption that we do not know the identity of either of the two images involved in the pairwise comparison.

In our case, with 803 images, we take one to be the new image, and the remainder form the known database. We consider the 802*801 CCF_{max} values, and take the subset corresponding to non-matching pairs. These can be thought of as a sample from the unknown theoretical distribution. The right tail probability of this empirical distribution would then give us the required probability. An illustration is in Figure 3.

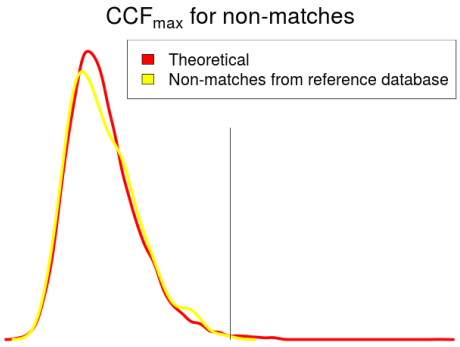

Figure 3: Illustrations of the unknown theoretical distribution of CCF_{max} values for true non-matches, and the empirical distribution constructed from values computed using a known database. The vertical line represents a new CCF_{max} value, and the proportion of the empirical distribution to the right of this vertical line is the probability of observing a higher value.

3.2.3 Results

We examine the similarity scores generated by the 644,006 pairwise comparisons. We can split these into matches and non-matches, and compare their distributions. This helps us evaluate if there is a threshold above which we are able to conclude a match. The results are in Figure 4. We find that there is no cutoff which separates the similarity scores for matches versus non-matches, meaning that a threshold-based approach would not be particularly successful, when we consider all of the 803 images.

In Figure 4, we also plot the scores by data set, to see if there are differences in performance related to the types of images that are included; this is also the "standard" presentation and allows comparison with other work. We find that the results are comparable to what have been published by other groups (e.g. Vorburger et al. 2007; Roth et al. 2015). Some observations that we make are that data sets involving consecutively manufactured slides have very good performance, while those involving multiple copies of the same gun firing different ammunition tend to have poor performance. Whether matched pairs produce high or low similarity scores seems to depend heavily on the specific firearm and ammunition combination, with pairs involving the same firearm and ammunition, fired in close succession having a better chance at producing high similarity scores. This however is not guaranteed for all firearms and ammunition types. Results are also affected by shot-to-shot variability.

3.3 Goals and future plans

The goal of this portion of the work is to devise an approach for comparing breechface images of cartridge cases that possesses all the desirable features as described in Section 3.1.1. To repeat,

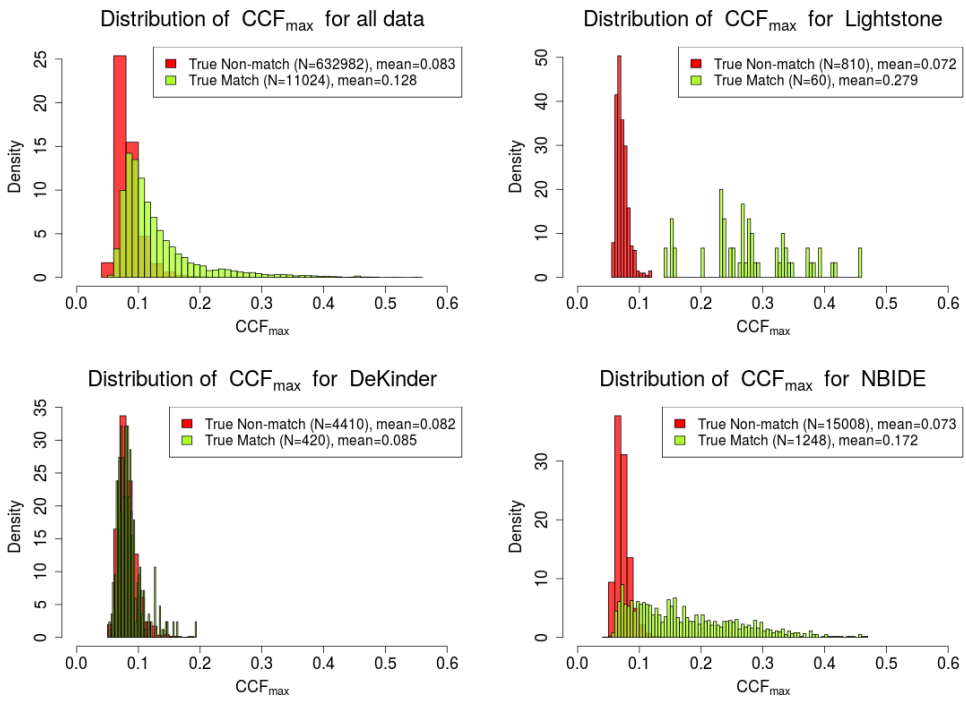

Figure 4: Distribution of CCF_{max} for matches and non-matches. The Laura Lightstone study involves consecutively manufactured pistol slides, firing the same ammunition. It achieves perfect or close to perfect separation between the match and non-match distributions. DeKinder involves multiple copies of the same make and model of gun, and many different brands of ammunition. There is very poor separation between the match and non-match distributions. The NBIDE study involves different makes and models of guns, and multiple brands of ammunition.

the method should be fully automatic, open source, produce a measure of the weight of evidence, be scalable so that large database searches can be performed, and be suitable for both 2D and 3D. Shortcomings of our current work are that the method is adapted for 2D images and is not computationally efficient. In terms of accuracy, there are matching pairs from some studies with particularly low scores, and these should be investigated further.

3.3.1 3D topographies and pre-processing

3D topographies differ from the 2D images in that the two measure different properties of the same physical object (surface contours and reflectance respectively). In terms of the images produced, the lateral resolution or physical size of each pixel is different ($3.125\mu m$ and $2.53\mu m$ respectively, for NIST's data); also, they were imaged separately so the orientation of the cartridge case could be different in the corresponding 2D and 3D images. We plan to first do all comparisons between the 3D images only, and as a next step, to consider comparisons of 2D versus 3D data. The latter is less straightforward, and would involve methods for interpolation to make the pixel sizes the same, rotations to make the orientations the same, and possible scaling such that pixel values and depth values can be compared.

As for pre-processing, in our current work, the automatic removal of the firing pin had poor

performance on some images, particularly those produced by Glock pistols, due to its unique rectangular shape. This should be looked into, and it would be interesting to see if using 3D topographies results in improvements. More work should also be done to investigate the effects and appropriateness of the various pre-processing steps, as well as any impact due to the order in which these steps are done. We propose to explore this using the VANOVA approach detailed in McNamee and Eddy (2001). (It also remains to be seen if better pre-processing and using 3D topographies will increase similarity scores for matching pairs.)

3.3.2 Producing signatures for large database searches

Here we propose producing summary signatures instead of using pixel-level information. We define a signature to be a low-dimensional representation of the image. An appropriate signature should result in a significant reduction in storage requirements, and support the efficient computation of useful distance metrics. By this, we mean that using such distance metrics for classification should not lead to substantially poorer results than when using pixel-level information.

One proposed option for creating signatures is to use wavelets. This originates from the literature on iris recognition (Daugman 2004), where systems have been developed and commercialized. The basic idea is to decompose the image using a wavelet basis. We then consider only coefficients from certain levels which are likely to possess information relating to individual characteristics. Further, instead of using the coefficients directly, we propose to convert them into binary features, for example by coding positive coefficients as 1 and negative ones as 0. In this manner, we reduce a high-dimensional image object into a much smaller binary vector. Pairs of binary vectors can then be compared using the Hamming distance, which is a very fast computation.

Apart from wavelets, a variety of methods have been used in database search and duplicate detection (e.g. Ghosh et al. 2007; Gionis, Indyk, and Motwani 1999) and for indexing in record linkage (e.g. P. Christen 2012), with overlap between the fields. Generally, these use a dimensionality reduction method (e.g. Fourier-based methods, Principal Components Analysis, FastMap), possibly in combination with an indexing method such as k-d trees or K-means clustering. All these approaches can be explored and possibly compared.

Another consideration is image registration, in particular rotations. Our current approach involves rotating the image through 360° and computing cross-correlations after each rotation. This is a very time-consuming process. A wavelet transform followed by computation of a Hamming distance could be faster than computing a cross-correlation, but other possibilities could also be looked into.

3.3.3 Likelihood ratios and comparison with examiners

Since one goal is to produce measures of the weight of evidence that can be used as court testimony, to supplement, or in place of firearms examiners, it is important to produce measures of the weight of evidence, as well as to compare error rates with those of examiners. Our current work has focused on the denominator of the likelihood ratio, and we would like to investigate both the denominator and the numerator, using the proposed signature-based approach. An important issue here is what reference population to use for constructing match and non-match distributions. Also, as described in Section 2, the literature in estimating likelihood ratios in record linkage could be useful. For example, it might be possible to adapt ideas such as using conditional independence assumptions

to estimate probabilities of observing particular similarity vectors for matching and non-matching pairs.

With regards to comparison with examiners, the PCAST report (President's Council of Advisors on Science and Technology 2016) mentions one study (Baldwin et al. (2014)) which was found to be acceptable, for estimating examiner error rates. This study reported both false positive and false negative rates for 218 participating examiners. As part of a collaboration with Iowa State University and the Defense Forensic Science Center, we have the opportunity to image a subset of these cartridge cases. The plan is to start with around 1,000 cartridge cases, and after imaging we plan to compare the results of our methods to those reported in that study. The 3D microscope used at ISU uses a slightly different technology than that at NIST, so we would have to ensure that all pre-processing and comparison methodologies work on these images.

4 Matching accounts on anonymous marketplaces

4.1 Background and literature

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 PGP to preserve anonymity. As a result these marketplaces are used primarily for sales 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 dozens of new marketplaces appeared, and this trend has continued in the subsequent years (Soska and Christin 2015).

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 make it very difficult to link accounts to real-world identities. Authorities have increased their efforts in cracking down on sellers and operators of anonymous marketplaces, and there have been some successes; a coordinated effort by global law enforcement resulted in the takedown of two large marketplaces in July 2017, and there has also been a recent string of arrests of top marketplace vendors (e.g. United States District Court, Eastern District of California 2016; United States District Court, Eastern District of California 2017).

Sellers can operate accounts on different marketplaces, and also multiple accounts on the same marketplace. In earlier sections we 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 the perhaps the most well-known marketplace-related arrest, the creator of Silk Road was known by a pseudonym, Dread Pirate Roberts. Authorities linked this 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 is that the combined sales on accounts operated by the same individual could factor into prosecutorial

decisions. To this end, arrests are frequently accompanied by a report of the various handles used by sellers, as well as the number of transactions or sales revenues.

As a side note, linking seller accounts is also an interesting problem from a non-forensics perspective. For example, this helps in estimating the size of the market. Information on unique sellers can be used for further analysis, such as estimating seller volumes and behavior, tying seller accounts to particular countries to track sales, and so forth. Characteristics of matched pairs can also shed light on the motivations for using multiple accounts.

As far as we know, law enforcement currently relies on "old-school," more manual investigative methods to link seller accounts, for example, using online forums or manual comparisons of items sold (see e.g. United States District Court, Eastern District of New York 2016), searching for PGP keys on Grams (United States District Court, Eastern District of California 2016) or using captured login credentials to seize accounts on other marketplaces (e.g. Dutch National Police 2017). Unlike in other forensic disciplines, there does not currently seem to be an automatic way of matching accounts or doing database searches on a large scale. Although to our knowledge a national database of seller account information (such as CODIS for DNA or NIBIN for firearms) does not exist, marketplaces are publicly available and can be scraped to generate the required information for such an automated technique to succeed. Matching accounts in an automatic way could allow the generation of a large number of leads much more quickly. As law enforcement increases their efforts targeting marketplace-related activity, we might expect such work to become more important.

There are some unique challenges associated with matching seller accounts. Users could create different accounts for different purposes, such as to sell different products, or targeting different markets. Different users could appear to be very similar if they are somehow related, such as being friends, or being part of the same distribution network. Sellers could copy each others' listing information such as item descriptions, or write in a similar style. If accounts belong to teams instead of individuals, behavior could be even more unpredictable. Sellers could also intentionally use accounts for deceptive purposes, for example sockpuppets or Sybils typically refer to multiple identities created by the same person: Accounts could belong to scammers, who do not want their accounts to be associated with one another. Sellers could also want anonymity due to participation in illegal activities, thus creating different personas. Another example of deceptive behavior is impersonators who pretend to be other users, for example well-known users with good reputations. This might help a user to appear credible and drum up sales. These types of "adversarial" behavior could complicate the matching process. A final challenge is that accounts could change behaviors, personas or ownership over time.

An interesting feature of these marketplaces is the availability of PGP keys, which could serve a partial unique identifier. Briefly, to receive a PGP-encrypted message, a seller would generate a PGP key-pair, consisting of a public key and a private key. The public key is listed as part of the seller's profile or item listings, and the sender would encrypt his message using this public key. Only the person in possession of the private key will be able to decrypt the message. These public keys are unique and could be used to link seller accounts. 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. Also, using PGP keys to label if pairs of accounts belong to the same seller could result in erroneous labels. These errors could be unintentional on the seller's part, for example the same seller using multiple PGP keys, or intentional, such as in an effort to assume different identities. Other examples of errors are different individuals working as a team and sharing a PGP key-pair, or a seller advertising another's PGP public key, in order to impersonate them. In summary, PGP

keys can be thought of as identifiers that could be unavailable, inaccurate or spoofed.

There have been some efforts by researchers to link user accounts on anonymous marketplaces. All these are deterministic methods (they use exact matching schemes). Soska and Christin (2015) used PGP keys, aliases and information from the Grams (a marketplace search engine) seller directory. Broséus et al. (2016) used PGP keys and aliases, as well as manual comparison of profile information. 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. In terms of specifically identifying deceptive use, Baravalle, Lopez, and Lee (2016) identified sockpuppets by "looking at vendors consistently using exactly the same images for their products," concluding that this use was "fairly limited," with no more than 2 sockpuppets per vendor. Apart from this, there seems to be little additional research, however deceptive use has been more widely studied in other contexts in the computer science literature, particularly in online forums and on social media (see e.g. Kumar et al. 2017; Tsikerdekis and Zeadally 2014).

4.2 Current work

In this part of the work, we have come up with a matching scheme that uses available PGP public keys to generate match labels, together with similary measures based on other available information, such as profile pages and item descriptions. We train a supervised model that is subsequently run on pairs with missing PGP keys.

4.2.1 Data

As described in Section 4.1, there is no national database of seller account information, but marketplaces are publicly available and information can be gathered by scraping associated web pages. Here the data that we are using are from Soska and Christin (2015)'s data collection effort, which consists of pages scraped from various marketplaces, including Agora, Evolution, Hydra, Pandora and Silk Road 2. These were collected from 2013 to 2015, and include seller pages, item listings as well as feedbacks. Information from seller pages includes account IDs or handles, PGP keys and profile descriptions. Item listings include item titles, descriptions, prices, and the shipping origin and destination. Feedback refers to reviews left by buyers, and is a proxy for the number of sales, since feedback is often mandatory on such marketplaces. Feedback information includes the approximate date the feedback was left, the item it corresponds to, and the buyer's comment. These pages were scraped regularly during the collection interval (2013-2015), so for a particular user, there would be multiple captures of their profile page and item listings, each with an associated timestamp.

In an effort to deal with the timing issues that were described earlier, in particular accounts changing behaviors, personas or ownership, we only consider seller behavior in a fixed chunk of time. Here we choose a subset of seller accounts having feedback (sales) any time from May-Aug 2014. We then use their profile page with the closest timestamp to Aug 31, 2014. All item listings with sales between May and August 2014 are used, together with the feedback received during this period. For each of these items, description pages with the closest timestamp to Aug 31, 2014 are used. Selection

of this particular time period also addresses a separate issue, that there is inherent instability in marketplaces, with marketplaces shutting down and new ones opening; this chosen period is one of relative stability.

This results in 3,512 seller accounts, selling 40,995 different items with 422,044 sales (pieces of feedback) being selected for our analysis. From these, we extract IDs, profile descriptions, item listing titles and descriptions, and number of feedback. PGP keys are extracted from profile and item descriptions.

4.2.2 Methodology

The general framework is as described in Figure 1 and Section 3.2. Some of the details are as follows.

Similarity Metrics From the information available for each record, as described in Section 4.2.1, we create the following similarity metrics for each pairwise comparison.

- Edit distance between the IDs
- Same or different marketplace
- Jaccard similarity between tokens in the Bag-of-Words representation of:
  - Profile descriptions
  - Titles of item listings
  - Descriptions of item listings
- Inventory-related (Leontiadis, Moore, and Christin 2013) Jaccard similarities: for item listings, consider unique
  - Categories
  - (Category, dosage) pairs
  - (Category, unit) pairs
  - (Category, dosage, unit) tuples
- Absolute difference between:
  - Number of listings with feedback (sales)
  - Diversity coefficient (see Soska and Christin 2015)
  - Number of days active (defined as the period from May-Aug 2014 between which sales are recorded)
  - Number of feedback
  - Number of feedback normalized by days active and marketplace total
  - Number of tokens in the Bag-of-Words representation of item descriptions

Classification Out of the 3,512 accounts to be matched, 2,820 (about 80%) have at least one known PGP key. We use these available keys to generate match/non-match labels (if a pair of records share at least one PGP key, we label them as a match). Of all the pairwise comparisons, the match statuses generated in this way are: 3,974,078 non-match, 712 match, 2,190,526 unavailable.

We then train a random forest classifier with 100 trees, using 10-fold cross-validation on the comparisons with known match status, in order to generate predicted match and non-match labels for the labeled data. We record the number of votes for each pair. As for the unlabeled data (pairs with at least one missing PGP key), we train the same model on all the labeled data and predict on these, similarly recording the number of votes for each pair.

Generate Clusters of Accounts With the number of votes in favor of a match (same PGP) label for each pair, we use hierarchical clustering to generate clusters of accounts that are predicted to

belong to the same seller. Note that for the labeled data, we use the predicted number of votes generated as described, instead of using the labels generated by PGP keys directly, since these could be erroneous. Now, we take the dissimilarity measure to be $1 - x$, where x is the proportion of votes in favor of a match label in the random forest. We then use two schemes to generate clusters of accounts that belong to the same entity. The first is complete linkage with a cutoff of .5, and the second is single linkage with the same cutoff.

To describe the above in words, we take .5 to be the cutoff above which we assign a match label to a pair. Then the first scheme necessitates that each member in the same cluster matches with every other member in cluster, and the second only requires that each member in the cluster matches with one other member in the cluster. These are illustrated in Figure 5.

Figure 5: Toy example to illustrate the two schemes to generate clusters of accounts. Each node represents an account and an edge means that the pair is a predicted match. In the first scheme, complete linkage is used, so the indicated edge is removed. In the second scheme, two edges are added instead.

4.2.3 Results

The results of the classification step on the labeled data are in Figure 6. The corresponding confusion matrix is also included; it is generated by classifying pairs with at least 50% of the votes as matches. The F1-score is .87. The first method of generating clusters results in 2,549 clusters, while the second results in 2,502 larger clusters. The largest cluster sizes are 5 and 8 respectively. At this point it is difficult to say which scheme is more accurate.

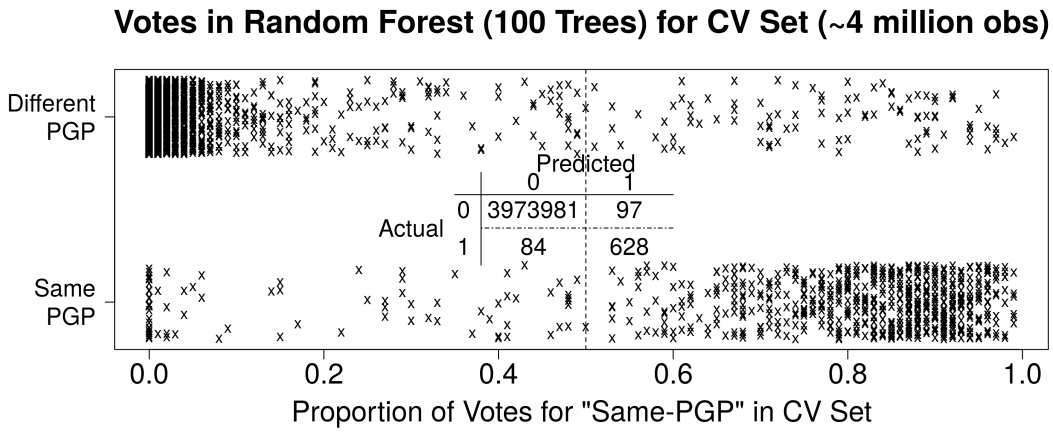

Figure 6: Proportion of cross-validation votes for random forest on labeled data. The confusion matrix is also included, where the numbers are counts corresponding to the number of crosses in the corresponding quadrant.

4.3 Goals and future plans

To reiterate, the goal of this portion of the work is to devise methods to match seller accounts, to generate accurate investigative leads, even if some attempt is made by sellers to obfuscate their identity. We would also like methods to be scalable. As far as we know, matching is currently used for investigative purposes, and evidence of a specific match is not used in courts in the same way as DNA or cartridge cases. It is unclear at this point how such evidence might be used or reported in courts in the future, but our current focus will not be on the quantification of weight of evidence. So far, we have used probabilistic methods that do not solely rely on PGP keys, but that use these to generate match/non-match labels, where available, to train a classification model. We noted that these keys may not always generate accurate labels. In proposed work, the focus will be on methods to deal with errors in the generated labels, as well as deriving a low-dimensional representation of the accounts to address the issue of scalability.

4.3.1 Improvements to model and other features

There are several straightforward ways in which the current model can be improved. There are other feature comparisons that can be done that might improve results. The parameters in the random forest model could be tuned, and we could also try other classifiers. Different methods for generating clusters of accounts can also be investigated. With regards to the features, we would like to look into other features based on seller behavior that are less easily spoofed, for example, the times where accounts are active or when sales are made, as opposed to what users put on their profiles. There is also much more work that can be done in terms of generating textual features. For example, instead of just using tokens from the Bag-of-Words representations, we could use tf-idf values. Other possibilities include looking at spelling errors, use of symbols, and so forth.

4.3.2 Discovering and fixing label errors

Next, we consider methods to discover the errors in labels generated by PGP keys. Although the random forest classifier is to some extent robust to label noise, we would like to take more concrete steps to deal with this. We could consider straightforward cutoffs, for example, we could select all pairs with say a $>.8$ Jaccard similarity of profile descriptions, and switch any non-match labels (different PGP keys) to match labels. More principled approaches for dealing with label noise are suggested in Frenay and Verleysen (2014). For example, ensemble-based methods can be used to remove instances that are often misclassified. Such automated methods have some limitations. First, some errors in labels would be difficult to pick out because they seem like expected behavior. Second, it is often difficult to distinguish mislabeled examples from true exceptions.

We also propose an active learning approach, which involves a manual review of a subset of labels. This would address the second limitation stated above. In this process, we might also be able to identify some types of deceptive use. Finally, this will help to develop "ground truth" labels, using which various techniques can be compared. Of course, the effectiveness of such a method relies on the accuracy of the manual labeler. One possible way to get true ground truth labels is to look at criminal complaints that include information on various accounts operated by an individual, but this information is few and far between. Also, it may be incomplete.

Sarawagi and Bhamidipaty (2002) suggest selecting pairs where the classifier's prediction differs

from a human matcher, as well as pairs in which the classifier is unsure. Here we modify this slightly, and propose initially selecting the following.

- Pairs which the initial model gets wrong
- Pairs with high similarity on any feature but different PGP or a non-match prediction: this supplements the earlier set, by adding those with high similarity on any one feature but not necessarily having a large number of votes. It also enables pairs with missing PGP keys to be included.
- Pairs in which the initial model is uncertain
- Pairs that are added or removed in the last (clustering) step
- Other random pairs.

We note that seller behavior could be very complex, and this proposed approach would not be able to idenfity all label errors, as some could seem like expected behavior. It would also be difficult to detect all types of deceptive use. The table in Appendix B summarizes some types of seller behavior, highlighting some difficult cases.

4.3.3 Unsupervised approaches and low-dimensional representation

Finally, we propose using unsupervised methods that do not rely on the PGP keys to generate labels. Whether pairs have the same PGP key might be used as a feature instead. Also, in order to improve scalability, we propose deriving a low-dimensional representation of each account. This mirrors the approach proposed for cartridge cases in Section 3.3.2. Due to the large amount of textual information associated with each account, one possibility is to do this using word embeddings, for example doc2vec, which is based on word2vec. Then some suitable similarity measure can be used (e.g. L1 or L2 distance, cosine similarity) for classification. Further, as described earlier, indexing methods can also be used in order to reduce computation time.

5 Summary of plans

Future work is detailed in Sections 3.3 and 4.3. The following list is a summary. For firearms:

- **Adapting to 3D and pre-processing.** We will adapt methods to 3D, and if time permits, develop methods to compare 2D to 3D data. We hope to improve the automatic selection of breechface marks, and investigate the effects, appropriateness, and order of the various pre-processing steps.
- **Producing signatures for large database searches.** We propose using wavelets and possibly other methods to generate signatures, and will look into indexing methods and similarity measures that might produce accuracies as close to that using pixel-level information.
- **Likelihood ratios.** We will explore methods for estimating likelihood ratios to quantify the weight of evidence.
- **Comparison with examiners.** We will compare our error rates with that of human examiners.

For marketplaces:

- **Generate additional features.** We would like to improve on the current set of features, including adding features related to seller behavior that may be more difficult to spoof.

- **Methods for discovering errors in labels.** We plan to adapt the active learning approach detailed in Sarawagi and Bhamidipaty (2002), which selects pairs for manual labeling using certain specified criteria. Other methods may also be considered.
- **Unsupervised methods.** We plan to use unsupervised methods which do not rely on labels generated by PGP keys, but might use those labels as a feature instead.
- **Low-dimensional representations.** We will explore word embeddings or other methods to derive a low-dimensional representation of the data, as well as indexing methods for scalability.

5.1 Tentative timeline

A tentative timeline based on the above tasks is as follows. As for priorities, the items within each list (cartridges and marketplaces) are ranked from highest to lowest priority.

Task	Spring '18	Fall '18	Spring '19
3D and pre-processing	X		
Signatures	X	X	
Likelihood ratios		X	X
Comparison with examiners			X
Additional features	X	X	
Discovering label errors	X	X	
Unsupervised approach		X	X
Low-dimensional representation		X	X
Writing		X	X

References

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): 336–40.

Baldwin, David P, Stanley J Bajic, Max Morris, and Daniel Zamzow. 2014. "A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons." AMES LAB IA.

Baravalle, A., M. S. Lopez, and S. W. Lee. 2016. "Mining the Dark Web: Drugs and Fake Ids." In *2016 Ieee 16th International Conference on Data Mining Workshops (Icdmw)*, 350–56. doi:10.1109/ICDMW.2016.0056.

Brein, C. 2005. "Segmentation of cartridge cases based on illumination and focus series." In *Image and Video Communications and Processing 2005*, edited by A. Said and J. G. Apostolopoulos, 5685:228–38. doi:10.1117/12.585763.

Broséus, J., D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, and D. Décary-Hétu. 2016. "Studying Illicit Drug Trafficking on Darknet Markets: Structure and Organisation from a Canadian Perspective." *Forensic Science International* 264: 7–14. doi:http://dx.doi.org/10.1016/j.forsciint.2016.02.045.

Buskirk, Joe Van, Raimondo Bruno, Timothy Dobbins, Courtney Breen, Lucinda Burns, Sundresan Naicker, and Amanda Roxburgh. 2017. "The Recovery of Online Drug Markets Following Law Enforcement and Other Disruptions." *Drug and Alcohol Dependence* 173: 159–62. doi:https://doi.org/10.1016/j.drugalcdep.2017.01.004.

Cadre Forensics. n.d. http://www.cadreforensics.com/technology.html, accessed 2017-08-20.

Champod, Christophe, Alex Biedermann, Joelle Vuille, Sheila Willis, and Jan De Kinder. 2016. "ENFSI Guideline for Evaluative Reporting in Forensic Science, a Primer for Legal Practitioners" 180 (January): 189–93.

Christen, P. 2012. "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication." *IEEE Transactions on Knowledge and Data Engineering* 24 (9): 1537–55. doi:10.1109/TKDE.2011.127.

Christen, Peter. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated.

Copas, J. B., and F. J. Hilton. 1990. "Record Linkage: Statistical Models for Matching Computer Records." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 153 (3). [Wiley, Royal Statistical Society]: 287–320. http://www.jstor.org/stable/2982975.

Daugman, J. 2004. "How Iris Recognition Works." *IEEE Transactions on Circuits and Systems for Video Technology* 14 (1): 21–30. doi:10.1109/TCSVT.2003.818350.

De Kinder, Jan, Frederic Tulleners, and Hugues Thiebaut. 2004. "Reference Ballistic Imaging Database Performance." *Forensic Science International* 140 (2–3): 207–15. doi:http://dx.doi.org/10.1016/j.forsciint.2003.12.002.

Dolliver, Diana S., and Jennifer L. Kenney. 2016. "Characteristics of Drug Vendors on the Tor Network: A Cryptomarket Comparison." *Victims & Offenders* 11 (4): 600–620.

doi:10.1080/15564886.2016.1173158.

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

Fellegi, Ivan P., and Alan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328): 1183–1210. doi:10.1080/01621459.1969.10501049.

Frenay, B., and M. Verleysen. 2014. "Classification in the Presence of Label Noise: A Survey." *IEEE Transactions on Neural Networks and Learning Systems* 25 (5): 845–69. doi:10.1109/TNNLS.2013.2292894.

Geradts, Zeno J, Jurrien Bijhold, Rob Hermsen, and Fionn Murtagh. 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). Elsevier: 97–106.

Gerules, George, Sanjiv K. Bhatia, and Daniel E. Jackson. 2013. "A Survey of Image Processing Techniques and Statistics for Ballistic Specimens in Forensic Science." *Science & Justice* 53 (2): 236–50. doi:http://dx.doi.org/10.1016/j.scijus.2012.07.002.

Ghosh, Pratim, E. Drelie Gelasca, K.R. Ramakrisnan, and B.S. Manjunath. 2007. *Duplicate Image Detection in Large Scale Databases*. Edited by Kolkata Indian Statistical Institute. Book Chapter in Platinum Jubilee Volume. https://vision.ece.ucsb.edu/sites/vision.ece.ucsb.edu/files/publications/pratim_2007_book.pdf.

Gionis, Aristides, Piotr Indyk, and Rajeev Motwani. 1999. "Similarity Search in High Dimensions via Hashing." In *Proceedings of the 25th International Conference on Very Large Data Bases*, 518–29. VLDB '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645925.671516.

Hare, E., H. Hofmann, and A. Carriquiry. 2016. "Automatic Matching of Bullet Land Impressions." *ArXiv E-Prints*, January.

Hepler, Amanda B., Christopher P. Saunders, Linda J. Davis, and JoAnn Buscaglia. 2012. "Score-Based Likelihood Ratios for Handwriting Evidence." *Forensic Science International* 219 (1): 129–40. doi:http://dx.doi.org/10.1016/j.forsciint.2011.12.009.

Kamalakannan, S., C. J. Mann, P. R. Bingham, T. P. Karnowski, and S. S. Gleason. 2011. "Automatic firearm class identification from cartridge cases." In *Image Processing: Machine Vision Applications Iv*, 7877:78770P. doi:10.1117/12.872414.

Kruithof, Kristy, Judith Aldridge, David Décary Hétu, Megan Sim, Elma Dujso, and Stijn Hoorens. 2016. *Internet-Facilitated Drugs Trade: An Analysis of the Size, Scope and the Role of the Netherlands*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR1607.html.

Kumar, Srijan, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. 2017. "An Army of Me: Sockpuppets in Online Discussion Communities." In *Proceedings of the 26th International Conference on World Wide Web*, 857–66. WWW '17. Republic; Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/3038912.3052677.

Leontiadis, Nektarios, Tyler Moore, and Nicolas Christin. 2013. "Pick Your Poison: Pricing and Inventories at Unlicensed Online Pharmacies." In *Proceedings of the Fourteenth Acm Conference on Electronic Commerce*, 621–38. EC '13. New York, NY, USA: ACM. doi:10.1145/2482540.2482610.

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*, 2:1494–8. doi:10.1109/ICIF.2003.177417.

Lightstone, Laura. 2010. "The Potential for and Persistence of Subclass Characteristics on the Breech Faces of Sw40ve Smith & Wesson Sigma Pistols." *AFTE Journal* 42 (4): 308–22.

McNamee, Rebecca L., and William F. Eddy. 2001. "Visual Analysis of Variance: A Tool for Quantitative Assessment of fMRI Data Processing and Analysis." *Magnetic Resonance in Medicine* 46 (6). John Wiley & Sons, Inc.: 1202–8. doi:10.1002/mrm.1317.

National Research Council, Division on Engineering and Physical Sci, National Materials Advisory Board. 2008. *Ballistic Imaging*. Edited by Daniel L. Cork, John E. Rolph, Eugene S. Meieran, and Carol V. Petrie. National Academies Press. http://www.ebook.de/de/product/7390977/national_research_council_division_on_engineering_and_physical_sci_national_materials_advisory_board_ballistic_imaging.html.

Popper, Nathaniel. 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.

———. 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.

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*. Washington, D.C.: Executive Office of the President.

Riva, Fabiano, and Christophe Champod. 2014. "Automatic Comparison and Evaluation of Impressions Left by a Firearm on Fired Cartridge Cases." *Journal of Forensic Sciences* 59 (3): 637–47. doi:10.1111/1556-4029.12382.

Roberge, Danny, and Alain Beauchamp. 2006. "The Use of BulletTRAX-3d in a Study of Consecutively Manufactured Barrels." *AFTE Journal* 38 (2). ASSOCIATION OF FIREARMS AND TOOL MARF EXAMINERS: 166.

Roth, J., A. Carriveau, X. Liu, and A. K. Jain. 2015. "Learning-Based Ballistic Breech Face Impression Image Matching." In *2015 Ieee 7th International Conference on Biometrics Theory, Applications and Systems (Btas)*, 1–8. Arlington, VA. doi:10.1109/BTAS.2015.7358774.

Sarawagi, Sunita, and Anuradha Bhamidipaty. 2002. "Interactive Deduplication Using Active Learning." In *Proceedings of the Eighth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 269–78. KDD '02. New York, NY, USA: ACM. doi:10.1145/775047.775087.

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). Blackwell Publishing Ltd: 195–212. doi:10.1111/j.1467-985X.2006.00457.x.

Song, J, W Chu, T V Vorburger, R Thompson, T B Renegar, A Zheng, J Yen, R Silver, and M Ols. 2012. "Development of Ballistics Identification–from Image Comparison to Topography Measurement in Surface Metrology." *Measurement Science and Technology* 23 (5): 054010.

Song, John. 2013. "Proposed 'NIST Ballistics Identification System (NBIS)' Based on 3d Topography

Measurements on Correlation Cells." *AFTE Journal* 45 (2): 184–94.

———. 2015. "Proposed 'Congruent Matching Cells(CMC)' Method for Ballistic Identification and Error Rate Estimation." *AFTE Journal* 47 (3): 177–85.

Soska, Kyle, and Nicolas Christin. 2015. "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem." In *Proceedings of the 24th Usenix Conference on Security Symposium*, 33–48. SEC'15. Berkeley, CA, USA: USENIX Association. http://dl.acm.org/citation.cfm?id=2831143.2831146.

Tai, Xiao Hui, and William F. Eddy. in press. "A Fully Automatic Method for Comparing Cartridge Case Images." *Journal of Forensic Sciences*, n/a–n/a. doi:10.1111/1556-4029.13577.

Thumwarin, P. 2008. "An Automatic System for Firearm Identification." In *2008 International Symposium on Communications and Information Technologies*, 100–103. doi:10.1109/ISCIT.2008.4700162.

Tsikerdekis, Michail, and Sherali Zeadally. 2014. "Online Deception in Social Media." *Commun. ACM* 57 (9). New York, NY, USA: ACM: 72–80. doi:10.1145/2629612.

Tunali, E., U.M. Leloglu, and U. Sakarya. 2009. "Method for Automatic Region Segmentation on Cartridge Case Base and Selection of the Best Mark Region for Cartridge Case Comparison." Google Patents. http://www.google.com.na/patents/WO2009130651A1?cl=en.

Ultra Electronics Forensic Technology. n.d. https://cdn2.hubspot.net/hubfs/71705/IBIS%20TRAX-HD3D%20EN%20Brochure%20Nov%2029%202016%20web.pdf, accessed 2017-08-31.

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.

———. 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.

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.

Ventura, Samuel L., Rebecca Nugent, and Erica R.H. Fuchs. 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. doi:http://dx.doi.org/10.1016/j.respol.2014.12.010.

Vorburger, T., J. Yen, B. Bachrach, T. Renegar, J. Filliben, L. Ma, H. Rhee, et al. 2007. "Surface Topography Analysis for a Feasibility Assessment of a National Ballistics Imaging Database." NISTIR 7362. Gaithersburg, MD: National Institute of Standards; Technology.

Winkler, William E. 2000. "Using the Em Algorithm for Weight Computation in the Felligi-Sunter Model of Record Linkage." In.

Zhou, Jie, Le-ping Xin, Gang Rong, and David Zhang. 2001. "Algorithm of Automatic Cartridge Identification." *Optical Engineering* 40 (12): 2860–5. doi:10.1117/1.1417497.

A Data in NIST Database

Table A1: Summary of data available in NIST's Ballistics Toolmark Research Database on 2/15/2017. Metadata available for download provide additional information such as study details, the type of firing pin, material of the primer, etc. There are a total of 803 images (16 images in the NBIDE study are repeated in the FBI: Ruger and FBI: S&W data sets).

Study	Images	Firearm	Number of firearms	Slides per firearm	Cartridge	Test fires per firearm/slide
Laura Lightstone	30	S&W 40VE	1	10	PMC	3
Todd Weller	50	Ruger P95DC	1	10	Winchester	5
Thomas Fadul	40	Ruger P95PR15	1	10	Federal	3-5
Hamby	30	Hi-Point C9	1	10	Remington	3
Kong	36	S&W 10-10	12	1	Fiocchi	3
Cary Wong	91	Ruger P89	1	1	Winchester	91
De Kinder	70	Sig Sauer P226	10	1	Remington	2
					CCI	1
					Wolf	1
					Winchester	1
					Speer	1
					Federal	1
FBI: Ruger	100	Various Rugers	50	1	Remington	2
FBI: S&W	138	Various S&Ws	69	1	Remington	2
FBI: Glock	90	Various Glocks	45	1	Remington	2
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

B Types of Seller Behavior

Table B1: Examples of types of seller behavior. We note whether associated pairs will be selected for manual review in our proposed active learning approach, if labels generated by PGP keys are erroneous, and whether such cases would be considered deceptive behavior. This is not an exhaustive list, and further investigation might reveal additional types of behavior. The main takeaway is that seller behavior could be very complex. Pairs with very similar characteristics (e.g. rows 3 and 4) could belong to the same or different sellers, so even after correcting erroneous labels, any model that we train is still likely to be imperfect.

Example	Selected for review	Erroneous label	Deceptive
Same seller with distinct accounts, e.g. selling different products, targeting different markets, one is less active/dormant	Yes if same PGP No if different/no PGP	No Yes (not caught)	No
Different sellers appear similar because they are related, e.g. being friends, being part of the same distribution network	Maybe if same PGP Yes if different/no PGP	Yes (not always caught) No	No
Different sellers copying each others' listing information or writing style	Yes (assume using different/no PGP)	No	No
Same seller with the same account information but using multiple PGP keys	Yes	Yes	No
Teams of sellers; could exhibit a wide variety of behaviors	Maybe	Maybe	No
Sockpuppet failing to create distinct enough identities	Yes (assume using different/no PGP)	Yes	Yes
Sockpuppet creating distinct identities	Yes if same PGP No if different/no PGP	No Yes (not caught)	Yes
Impersonators copying all information	No if same/no PGP Yes if different PGP	Yes (not caught) No	Yes
Impersonators copying some information	Maybe if same PGP Yes if different/no PGP	Yes (not always caught) No	Yes