CSAFE Presentations and Proceedings

Center for Statistics and Applications in Forensic Evidence

11-2018

Record Linkage and Matching Problems in Forensics

Xiao Hui Tan
Carnegie Mellon University

Recommended Citation

Tan, Xiao Hui, "Record Linkage and Matching Problems in Forensics" (2018). *CSAFE Presentations and Proceedings*. 10.
https://lib.dr.iastate.edu/csafe_conf/10

Record Linkage and Matching Problems in Forensics

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, illustrating this using examples from DNA and firearms identification. We propose some ways that record linkage can inform forensic matching. Finally, we develop methodology to match accounts on anonymous marketplaces. In forensic matching, the stakes are high and the consequences of false arrests or wrongful convictions are severe. The field would benefit from a more principled way of developing matching methods.

Comments

This is a manuscript of a proceeding published as Tai, Xiao Hui. "Record Linkage and Matching Problems in Forensics." In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 510-517. IEEE, 2018. Posted with permission of CSAFE.

Record Linkage and Matching Problems in Forensics

Xiao Hui Tai
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA, USA
xtai@andrew.cmu.edu

***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, illustrating this using examples from DNA and firearms identification. We propose some ways that record linkage can inform forensic matching. Finally, we develop methodology to match accounts on anonymous marketplaces. In forensic matching, the stakes are high and the consequences of false arrests or wrongful convictions are severe. The field would benefit from a more principled way of developing matching methods.**

***Index Terms*—record linkage, forensic matching, unstructured data**

I. Introduction

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 [1]. 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.

Forensic evidence refers to DNA, fingerprints, bullets and cartridge cases, shoeprints, digital and other 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, such that the evidence 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.[1] These are closely related however, since a one-to-many comparison could be treated as repeating a one-to-one comparison many times. In current practice for fingerprints and firearms, an automated database search might be conducted, followed by manual examination by a human examiner. Any final determination of a match that is used as court testimony is made by an examiner. This has led to recent public scrutiny due to the subjective nature and lack of scientific validity [2]. There has been a push towards automated methods, and these have been developed by various groups for database searches, objective identification, or both (e.g. [3], [4]). The focus of this paper is on automated forensic matching methods.

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. This correspondence has been noted by statisticians in the past, but to our knowledge this has not been formalized or exploited. [5] mention comparing an unidentified fingerprint with fingerprints of known individuals as a possible application of computer matching. [6] 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.

By thinking about forensics problems in the context of record linkage, we immediately have well-developed frameworks and tools at our disposal. We illustrate the correspondence using examples from DNA and firearms identification, and propose some ways that the literature from record linkage can inform forensic matching. Finally, we develop methodology for comparison of one type of digital evidence.

The main contributions of this paper are:

1) Demonstrating a link between forensic matching problems and record linkage, and establishing a framework to approach the former
2) Suggesting ways that record linkage can inform forensic matching

This research was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

[1] In forensics, a distinction between "same-source" and "specific-source" is also sometimes made, where the former refers to coming from a common, unknown source.

3) Proposing a method for matching seller accounts on anonymous marketplaces and presenting preliminary results.

The rest of the paper is organized as follows. Section II establishes the connection between record linkage and forensic matching problems, using DNA and firearms identification as illustrative examples. Section III explores some specific ways in which forensic matching might benefit from being thought of as a record linkage problem. Section IV describes methodology for matching seller accounts on anonymous marketplaces, and presents preliminary results, and Section V concludes the discussion.

II. Correspondence between Record Linkage and Forensic Matching

A standard framework for record linkage is in Figure 1. This framework is adapted from [1] and [7]. 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 clusters of records are produced from these comparisons, such that transitivity is preserved.

In terms of forensic matching, many different automated or semi-automated approaches have been developed by people working in different fields, in an ad-hoc manner. These might be for database searches or for one-to-one comparisons. Using the examples of DNA and firearms identification, we demonstrate how forensic matching problems fall into this framework, noting that this extends to automated forensic matching methods for other evidence types.

One conceptual difference is that in record linkage we might be more concerned about disambiguating a data set to get a final deduplicated 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 matching 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. Another feature is that in forensic matching problems, we are often dealing with much more complex, unstructured data, such as multivariate data, images and text.

A. DNA

DNA profiling 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 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, Hispanic, 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

$$LR = \frac{\mathbb{P}[\text{Evidence, e.g. 13 STRs match}|\text{Same source}]}{\mathbb{P}[\text{Evidence}|\text{Different source}]}. \quad (1)$$

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 (for example the race of the suspect). The likelihood ratio can be interpreted as the number of times more likely we are 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 framework in Figure 1, each DNA sample (record) is summarized using 13 features which correspond to the 13 STRs. The similarity for each pair of records is a vector of length 13, with each entry being a binary value for whether the number of repeats at the STR location is the same. A threshold-based classification method is then used, requiring that all 13 entries be 1 in order to be classified as a match. 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 an unsupervised threshold-based classification approach is used, with a cutoff of 1 to be classified as a match. Since this is an exact matching scheme, transitivity is automatically preserved.

Estimating the likelihood ratio is then an additional step. The methodology used is essentially the same as in the Fellegi-Sunter model, which is often considered the standard model for record linkage, in the case where population frequencies of the features are known [8]. In this case, the distribution of number of repeats at each location is estimated for each race from standard population databases, and is treated as known. Using the language from [8], in the numerator of Equation 1 are m-probabilities and in the denominator u-probabilities. The Fellegi-Sunter model and its connection to likelihood ratios in forensic matching is discussed in greater detail in Section III.

B. Firearms identification

A gun is thought to leave identifiable marks on bullets and cartridge cases, 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. The mechanism in which marks are imparted is illustrated in Figure 2. When a gun is fired, the

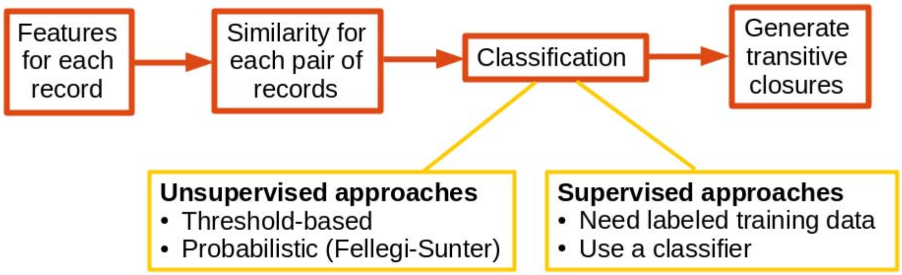

Features for each record
Similarity for each pair of records
Classification
Generate transitive closures
Unsupervised approaches
• Threshold-based
• Probabilistic (Fellegi-Sunter)
Supervised approaches
• Need labeled training data
• Use a classifier

Fig. 1. Standard framework for statistical record linkage problems.

cartridge is hit by the firing pin of the gun, 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. Rifling, manufacturing defects, and impurities in the barrel create striation marks on the bullet. As for the cartridge case, at least two kinds of marks are created: the firing pin impression is caused by the firing pin hitting the cartridge, and breechface marks are impressed by the bottom surface of the cartridge pressing against the breech block of the gun. 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. [9]).

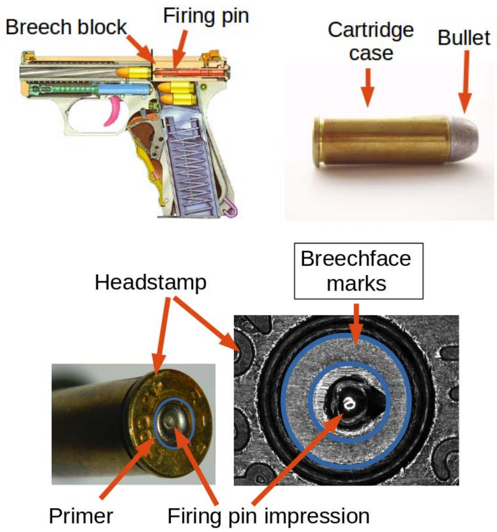


Fig. 2. On the top we have a gun that is about to be fired, showing the internal parts, and a cartridge before firing. On the bottom left is the bottom surface of a cartridge case after firing. The firing pin impression is clearly visible as the "hole" in the center of the primer surface, while the breechface marks lie on the remaining primer surface. On the bottom right is an image of such a bottom surface, taken using a reflectance microscope. (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 8/5/2018.)

Firearms identification 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. In current practice in the United States, evidence is entered into a national database called the National Integrated Ballistics Information Network (NIBIN). A computer-based 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. Other automatic comparison methods have been developed by engineers and scientists both in industry and in academia. We describe how these approaches fall under the record linkage framework.

Bullets Automated methods methods 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. [10]) and supervised methods (e.g. [11]) have been used. For example, [11] used a total of seven similarity measures, and a random forest classifier, reporting no classification errors on a test data set.

Cartridge cases Comparison of cartridge cases follows a similar process. Some groups use the entire image as features for each record (image) (see e.g. [4], [12]–[14]), while other groups extract features from these images (e.g. [15] extracts "corner-like textured regions... which correspond to the same types of ridges, ... that a trained firearms examiner would identify.") For each pair, similarity measures include the correlation between aligned images, average Euclidean distance, and the number of matched features. For 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. As an example, [13] use a supervised boosting-based classification method.

III. How Record Linkage can Inform Forensic Matching

Using a record linkage framework allows us to take advantage of tools that have been developed, without having to reinvent the wheel. In Section IV we demonstrate how associated techniques can be directly applied to develop methods for new forensic matching problems. In forensic domains such as

DNA, fingerprints and firearms identification where there are existing well-developed bodies of literature, we might think about improvements that can be made. The following is a list of some potential ideas.

1) **Fellegi-Sunter, cutoffs, and weight of evidence.** [8] proposed a probabilistic framework for assigning matches, and this subsequently gained widespread popularity. 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 [16].
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 Europe has moved decisively in this direction, and guidelines for forensic laboratories now recommend the reporting of likelihood ratios for all forensic disciples [17]. However there is no consensus on how these likelihood ratios should be estimated. For DNA, a generative model is used to determine the probabilities of observing particular DNA profiles, as described in Section II. 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. [18]). Further investigation could reveal if other methods from the record linkage literature that are used to estimate likelihood ratios, 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 would be an improvement over threshold-based methods that have been suggested in the forensic literature, which use somewhat arbitrary cutoffs without a proper quantification of error rates. For example, in cartridge case comparisons, [19] suggests using a cutoff of six matching cells (regions on the cartridge case) to determine matches, with the rationale being that in casework for bullets, examiners use six consecutive matching striae as a cutoff.
2) **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 [1]. 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.
3) **Generating clusters of records.** In the last step in Figure 1, clusters of records are generated 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.
4) **Deduplicating existing databases.** The last two suggestions both hint 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, and it is unclear if this is currently being done. The record linkage literature can definitely lend itself towards this effort.

This of course is not in any way 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.

IV. Matching accounts on anonymous marketplaces

In this section we demonstrate the usefulness of adapting record linkage methods to a new forensic matching problem: matching seller accounts on online anonymous marketplaces. We describe our proposed methodology and share preliminary results.

Anonymous marketplaces have certain features that provide anonymity protections to buyers and sellers. They run on the dark web using browsers like Tor, use cryptocurrencies for payment, and tools such as PGP for encryption. Such marketplaces are hence most commonly 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 [20]. Due to the anonymous nature of such marketplaces, tracking down sellers is difficult and many cases are still being investigated.

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. 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, unlike in the other forensic disciplines. So far agents have relied primarily on "old-school" investigative methods, for example, using online forums or manual comparisons of items sold [21], searching for PGP keys on Grams [22] or using captured login credentials to seize accounts on other marketplaces [23]. In this part of the paper, we apply record linkage techniques to match accounts automatically and on a large scale.

There have been some efforts by researchers to link user accounts on anonymous marketplaces. Most of these are deterministic methods (they use exact matching schemes). [20] used PGP keys, aliases and information from the Grams (a marketplace search engine) seller directory. [24] used PGP keys and aliases, as well as manual comparison of profile information. [25] similarly used exact matching on PGP keys, aliases and profile descriptions. [26] found an 8% overlap between aliases between two marketplaces, but made no further attempt to infer if they belonged to the same entities. [27] 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. Most recently, [28] use images taken from item listings, and train a deep neural network for detecting multiple accounts on the same and different marketplaces.

A. Data

To our knowledge 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 [20]'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 feedback.[2] Information from seller pages includes account IDs or handles, PGP keys and profile descriptions. Item listings include item titles, descriptions, prices, and the shipping country (origin and destination). Feedback refers to reviews left by buyers, and can be used as 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 could be multiple captures of their profile page and item listings, each with an associated timestamp.

To develop our model, we chose 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.

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 feedback. PGP keys are extracted from profile and item descriptions.

B. Methodology

Similarity Metrics From the information available for each record, as described in Section IV-A, we generate the following similarity metrics for each pairwise comparison.

From profiles:

1) Edit distance between the IDs
2) Same or different marketplace
3) Jaccard similarity between Bag-of-Words representation of profile descriptions
4) Absolute difference between diversity coefficient[3]

From item listings:

1) Jaccard similarity between Bag-of-Words representation of item titles and descriptions
2) Inventory-related [29] Jaccard similarities: consider unique categories, (Category, dosage) pairs, (Category, unit) pairs, and (Category, dosage, unit) tuples
3) Absolute difference between number of tokens in the Bag-of-Words representation of item descriptions

From feedback (sales):

1) Absolute difference between number of days active (defined as the period from May-Aug 2014 between which sales are recorded)
2) Absolute difference between number of listings with feedback (sales)
3) Absolute difference between the number of feedback, and number of feedback normalized by days active and marketplace total

Classification An interesting feature of these marketplaces is the availability of PGP keys. 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 potentially be used to link seller accounts, but usage of PGP encryption is not mandatory, and keys do not necessarily correspond to unique sellers.

In order to use PGP keys to generate labels, we do the following. We consider PGP public keys that are posted on

[2] Anonymized data are publicly available at https://arima.cylab.cmu.edu/markets/cybercrime.php.

[3] Briefly, this measures the diversity in terms of categories of items sold. See [20] for more details.

profiles or item descriptions of each seller account. If a pair of accounts share at least one PGP key, we label them as a match, and if not, we label them as a non-match. This relies on the assumption that two accounts with the same key belong to the same seller, and accounts with different keys belong to different sellers. This may not always be true, for example, the same seller might use multiple PGP keys, one for each account. On the other hand, a seller might advertise another seller's PGP public key in an attempt at impersonation, so a pair of accounts with the same key might in fact belong to different sellers. As a result, some pairs would be mislabeled, but in the absence of ground-truth data for whether pairs of accounts belong to the same seller, we use these labels generated by PGP keys to train our model.

Out of the 3,512 accounts to be matched, 2,820 (about 80%) have at least one known PGP key. Of all the pairwise comparisons, the match statuses generated in the manner described are: 3,974,069 non-match, 721 match, 2,190,526 unavailable. We train a random forest classifier with 100 trees, using 10-fold cross-validation on the comparisons with known PGP-match status, in order to generate predictions for the PGP-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 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 PGP labels 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 seller. 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 seller account in the same cluster matches with every other account in cluster, and the second only requires that each account in the cluster matches with one other account in the cluster.

C. Results

Preliminary results of the classification step on the labeled data are in Figure 3. 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 .88. The random forest generates variable importances based on how much the prediction error increases when out-of-bag data for that variable is permuted [30], and the most important variables in this case tend to be textual features. In decreasing order of importance, these are ID distance, Jaccard similarity between descriptions, item titles and profiles.

The first method of generating clusters (complete linkage) results in 2,530 clusters, and the cluster sizes, or number of accounts per seller, are tabulated in Table I.

TABLE I
NUMBER OF ACCOUNTS PER SELLER USING COMPLETE LINKAGE.

Accounts per seller (cluster)	1	2	3	4	5
Number of sellers	1824	479	185	35	7

The second method (single linkage) results in 2,487 clusters, with larger cluster sizes. These are in Table II.

TABLE II
NUMBER OF ACCOUNTS PER SELLER USING SINGLE LINKAGE.

Accounts per seller (cluster)	1	2	3	4	5	6	7	8
Number of sellers	1790	445	193	48	8	1	1	1

An example of the largest cluster of accounts is in Table III. Although we are unable to verify with absolute certainty if these accounts indeed belong to the same seller, manual examination as well as a Reddit post[4] corroborate this finding. In this example we also note that a simpler method, such as solely using PGP keys or aliases would be inadequate in identifying all accounts belonging to this seller.

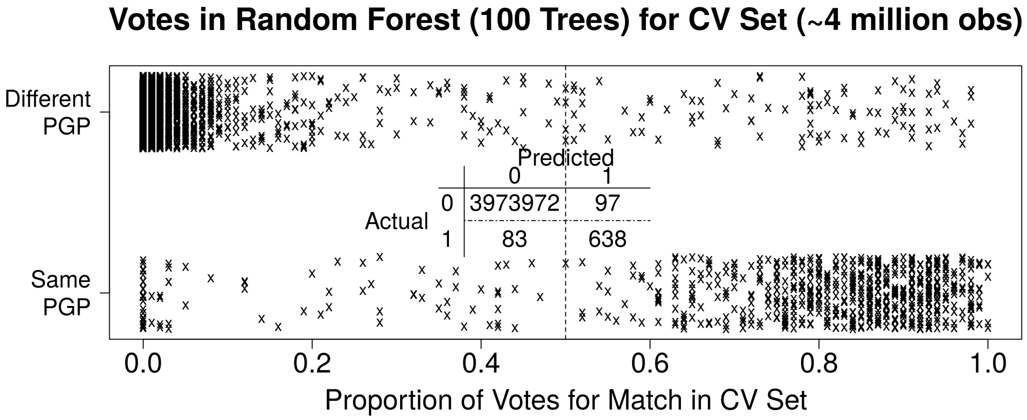
TABLE III
LARGEST CLUSTER OF ACCOUNTS FOUND.

Alias	Marketplace	PGP
scaptain	Pandora	
Scaptain	Silk Road 2	B
tambourineman	Agora	A
Tambourineman	Silk Road 2	A
tambourinemans	Pandora	
tomorrowman	Agora	B
tomorrowman	Hydra	
tomorrowman	Pandora	

D. Discussion

One tricky aspect of the data is the absence of readily available ground-truth labels for whether pairs of accounts belong to the same seller or not. So far we have used PGP keys to generate training labels, but we note that these could be erroneous. This has implications on model and parameter selection. Some experimentation was done in terms of optimizing the random forest parameters, such as the number of trees and the number of variables tried at each split. For example, increasing the number of variables tried at each split tended to increase the number of predicted matches, while decreasing the number of trees led to more errors. Due to the large size of the data set, increasing the number of trees beyond 100 was somewhat infeasible because of long computation times and high memory requirements. We also tried other classifiers such as boosting and logistic regression, with similar

[4] https://www.reddit.com/r/SilkRoad/comments/2a5g2d/best_heroin_vendor_that_ships_to_canada/



		Predicted	
		0	1
Actual	0	3973972	97
	1	83	638

Fig. 3. Proportion of cross-validation votes from the random forest on PGP-labeled data. The confusion matrix is also included, where the numbers are counts corresponding to the number of crosses in the corresponding quadrant. The top-right and bottom-left quadrants are when the model predictions do not agree with PGP labels.

or slightly 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.

An important direction of future work will involve methods to deal with these potentially erroneous labels. We could first flag record pairs for manual review. These might include pairs where the model predictions and PGP-labels do not match, pairs in which the model is unsure, or clusters of large sizes, and so forth. The advantage is that a human can use information that may not be captured by the model, for example from feedback messages, or from other sources such as online forums. These manually generated labels can then be used in a number of ways, depending on the number of pairs labeled. Supervised models can be trained using the manually labeled pairs, and in this case both models and parameters can be optimized fully since we know that the labels are accurate. A semi-supervised approach can also be used: the model can be retrained with the manual labels, together with some or all of the PGP labels. A final option could be to use unsupervised methods, and leave the manually labeled pairs out to use as a test set.

From a record linkage perspective, there is an unusual aspect of these data, and that is the potential "adversarial" behavior of sellers. Sellers might intentionally use accounts for deceptive purposes, for example sockpuppets or Sybils typically refer to multiple identities created by the same person. These could be used by 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. Such behavior is not typically observed in other record linkage applications, such as matching Census records.

V. Conclusion

In this paper, we demonstrated the correspondence between forensic matching and record linkage problems, and suggested some ways in which record linkage might inform forensic matching. We applied this thinking to a new forensic domain, developing methodology to match seller accounts on anonymous marketplaces. By using a common, well-established framework, we show how forensic matching problems can be approached in general, in a more principled manner.

With recent public criticism about current forensic practice and its reliance on subjective opinions of examiners, there has been a push towards objective, automated methods. We have begun to see increased efforts on this front, and drawing links to established record linkage methodology could become even more useful as new methods are developed.

As a reviewer highlighted, there are some peculiarities of the forensic field that are worth keeping in mind in developing new methods. First is the potential adversarial nature of the information. We described this in Section IV with the marketplaces data, but this could apply more broadly in other forensic disciplines. Second is the danger of false positives and the potential impact on human life. The consequences of false arrests or wrongful convictions are severe, and this could have implications on classification methods; as an example it might be beneficial to attach a much higher cost to false positives compared to false negatives.

Acknowledgements

I am grateful to my advisor William F. Eddy for his unending support and openness to new ideas. I would like to thank Nicolas Christin for very generously giving me access to the marketplace data, and welcoming me to his research group. Many interesting ideas and improvements came out of our weekly meetings; thanks in particular also to Kyle

Soska for numerous helpful discussions. Finally, thanks to the anonymous reviewers for their insightful comments.

References

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

[2] President's Council of Advisors on Science and Technology, *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, Sep. 2016.

[3] P. Thumwarin, "An automatic system for firearm identification," in *2008 International Symposium on Communications and Information Technologies*, Oct 2008, pp. 100–103.

[4] F. Riva and C. Champod, "Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases," *Journal of Forensic Sciences*, vol. 59, no. 3, pp. 637–647, 2014.

[5] J. B. Copas and F. J. Hilton, "Record linkage: Statistical models for matching computer records," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 153, no. 3, pp. 287–320, 1990. [Online]. Available: http://www.jstor.org/stable/2982975

[6] C. J. Skinner, "The probability of identification: applying ideas from forensic statistics to disclosure risk assessment," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 1, pp. 195–212, 2007. [Online]. Available: http://dx.doi.org/10.1111/j.1467-985X.2006.00457.x

[7] S. L. Ventura, R. Nugent, and E. R. Fuchs, "Seeing the non-stars: (some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records," *Research Policy*, vol. 44, no. 9, pp. 1672 – 1701, 2015, the New Data Frontier. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048733314002406

[8] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049

[9] L. Lightstone, "The potential for and persistence of subclass characteristics on the breech faces of SW40VE Smith & Wesson Sigma pistols," *AFTE Journal*, vol. 42, no. 4, pp. 308–322, 2010.

[10] D. Roberge and A. Beauchamp, "The use of BulletTRAX-3D in a study of consecutively manufactured barrels," *AFTE Journal*, vol. 38, no. 2, p. 166, 2006.

[11] E. Hare, H. Hofmann, and A. Carriquiry, "Automatic Matching of Bullet Land Impressions," *ArXiv e-prints*, Jan. 2016.

[12] T. Vorburger, J. Yen, B. Bachrach, T. Renegar, J. Filliben, L. Ma, H. Rhee, A. Zheng, J. Song, M. Riley, C. Foreman, and S. Ballou, "Surface topography analysis for a feasibility assessment of a National Ballistics Imaging Database," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NISTIR 7362, 2007.

[13] J. Roth, A. Carriveau, X. Liu, and A. K. Jain, "Learning-based ballistic breech face impression image matching," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Arlington, VA, Sept 2015, pp. 1–8, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7358774&isnumber=7358743 (accessed April 28, 2017).

[14] X. H. Tai and W. F. Eddy, "A fully automatic method for comparing cartridge case images,," *Journal of Forensic Sciences*, vol. 63, no. 2, pp. 440–448, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.13577

[15] Cadre Forensics, http://www.cadreforensics.com/technology.html, accessed 2017-08-20.

[16] W. E. Winkler, "Using the em algorithm for weight computation in the felligi-sunter model of record linkage," 2000.

[17] C. Champod, A. Biedermann, J. Vuille, S. Willis, and J. De Kinder, "ENFSI guideline for evaluative reporting in forensic science, a primer for legal practitioners," vol. 180, pp. 189–193, 01 2016.

[18] A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence," *Forensic Science International*, vol. 219, no. 1, pp. 129 – 140, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0379073811006013

[19] J. Song, "Proposed "NIST ballistics identification system (NBIS)" based on 3D topography measurements on correlation cells," *AFTE Journal*, vol. 45, no. 2, pp. 184–194, 2013.

[20] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *Proceedings of the 24th USENIX Conference on Security Symposium*, ser. SEC'15. Berkeley, CA, USA: USENIX Association, 2015, pp. 33–48. [Online]. Available: http://dl.acm.org/citation.cfm?id=2831143.2831146

[21] United States District Court, Eastern District of New York, "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, 2016, dark51.

[22] United States District Court, Eastern District of California, "Affidavit of matthew larsen," https://www.justice.gov/usao-edca/file/836576/download, accessed 2017-08-20, 2016, caliconnect.

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

[24] J. Broséus, D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, and D. Décary-Hétu, "Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective," *Forensic Science International*, vol. 264, pp. 7 – 14, 2016, special Issue on the 7th European Academy of Forensic Science Conference. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0379073816300676

[25] K. Kruithof, J. Aldridge, D. D. Hétu, M. Sim, E. Dujso, and S. Hoorens, *Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands*. Santa Monica, CA: RAND Corporation, 2016. [Online]. Available: https://www.rand.org/pubs/research_reports/RR1607.html

[26] D. S. Dolliver and J. L. Kenney, "Characteristics of drug vendors on the tor network: A cryptomarket comparison," *Victims & Offenders*, vol. 11, no. 4, pp. 600–620, 2016. [Online]. Available: http://dx.doi.org/10.1080/15564886.2016.1173158

[27] J. V. Buskirk, R. Bruno, T. Dobbins, C. Breen, L. Burns, S. Naicker, and A. Roxburgh, "The recovery of online drug markets following law enforcement and other disruptions," *Drug and Alcohol Dependence*, vol. 173, pp. 159 – 162, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0376871617300741

[28] X. Wang, P. Peng, C. Wang, and G. Wang, "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*, ser. ASIACCS '18. New York, NY, USA: ACM, 2018, pp. 431–442. [Online]. Available: http://doi.acm.org/10.1145/3196494.3196529

[29] N. Leontiadis, T. Moore, and N. Christin, "Pick your poison: Pricing and inventories at unlicensed online pharmacies," in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, ser. EC '13. New York, NY, USA: ACM, 2013, pp. 621–638. [Online]. Available: http://doi.acm.org/10.1145/2482540.2482610

[30] A. Liaw, M. Wiener *et al.*, *Classification and regression by randomForest*, Std. 3, 2002.