

# Using Spatial, Temporal and Evidence-status Data to Improve Ballistic Imaging Performance

Yan Yang<sup>\*</sup>, Avi Koffman<sup>†</sup>, Gil Hocherman<sup>‡</sup>, Lawrence M. Wein<sup>§</sup>

November 8, 2011

---

<sup>\*</sup>Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305,  
yanyang@stanford.edu

<sup>†</sup>Division of Identification & Forensic Science, Israel Police Investigation Department, National Police  
Headquarters, Jerusalem 91906, Israel, avikfm@gmail.com

<sup>‡</sup>Division of Identification & Forensic Science, Israel Police Investigation Department, National Police  
Headquarters, Jerusalem 91906, Israel, ghocher@gmail.com

<sup>§</sup>Graduate School of Business, Stanford University, Stanford, CA 94305, lwein@stanford.edu

## Abstract

Ballistic imaging systems help solve crimes by comparing newly acquired images of cartridge casings or bullets to a database of images obtained from past crime scenes. We formulate and solve an optimization problem that bases its matching decisions (i.e., which database images should be forwarded to firearm examiners for verification) not only on the similarity scores between pairs of images, but also on extraneous data on the evidence status (i.e., whether the casings or bullets are recovered from a crime scene or elsewhere) of each new acquisition and the time and spatial location of each acquisition and each database entry. The objective is to maximize the detection probability (i.e., the probability that at least one true match in the database is detected given that at least one exists) subject to a constraint on the expected number of false positives. Using data on all cartridge casings matches detected in Israel during 2006-2008, we predict that the optimal use of extraneous information would increase the detection probability from 0.931 to 0.987 (i.e., a 81.4% reduction in the false negative rate), which is achieved by favoring (via setting lower similarity-score thresholds for generating a match) pairs of images that are closer together in space and time. An application of our model to the U.S. suggests that an optimal national search strategy, which favors local searches over inter-state searches, would increase the detection probability relative to current U.S. implementations, which are largely restricted to local (e.g., intra-state) searches.

The Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) developed the National Integrated Ballistic Information Network (NIBIN) to help state and local law enforcement agencies solve gun crimes [1]. NIBIN uses computerized imaging technology to maintain a database on cartridge casings and bullets that are either recovered from crime scenes (called “evidence”) or test-fired from weapons that are recovered by (or surrendered to) law enforcement officers (called “nonevidence”). NIBIN rapidly computes similarity scores between a newly acquired casing or bullet and the entries in the database, and – using software developed by a single vendor, Forensic Technology WAI, Inc. – generates a list of the (e.g., 10) most promising matches, which are subsequently analyzed by a forensic firearms examiner to obtain confirmed hits. In this manner, NIBIN can potentially identify a cold hit between a nonevidence acquisition and an earlier crime, or discover links between crimes, both of which generate new leads to assist in crime solving. This system, if used properly (i.e., data entered in a thorough and timely manner, and confirmed hits integrated with all other investigative information), can be very useful to local police departments (Appendix A of [2]). However, NIBIN is used very inconsistently by various U.S. municipalities, and rarely employed for nonlocal (e.g., interstate) searches [1].

Ballistic imaging technology – particularly for bullets – does not perform as well (e.g., as measured by Receiver Operating Characteristic curves) as some biometric technologies, such as fingerprints and DNA matching, that are also used for forensic purposes [2, 3]. The goal of this study is to examine whether ballistic imaging performance can be improved by combining it with other spatial, temporal and categorical data that are collected along with the ballistic image. More specifically, we introduce and optimize a threshold-based system (i.e., potential hits are defined by similarity scores above a certain threshold rather than by being ranked in the top 10) that allows the threshold used to compare a newly acquired casing or bullet with one in the database to depend on the time interval between the two

events (i.e., acquisition or crime), the spatial distance between the two events, and whether the new acquisition is evidence or nonevidence. The rationale behind this approach is that crime guns and their crimes cluster in space and time, and evidence acquisitions are more apt to have been involved in a previous crime than nonevidence acquisitions.

We use spatial, temporal and categorical data on all Israeli matches during 2006-2008, along with published performance curves for cartridge casings, to calibrate our model and assess the potential improvement in performance of our approach. We also use these data to crudely extrapolate to the U.S. setting, where we investigate the efficiency of performing non-local searches for both casings and bullets.

**Model.** Our model formulation holds for either cartridge casings or bullets. The evidence status of a new acquisition, or arrival, is denoted by the subscript  $i = 0$  for nonevidence recovered by law enforcement officers and  $i = 1$  for evidence obtained from a crime scene. The probability that an arrival is of evidence status  $i$  is  $q_i$  for  $i = 0, 1$ . Each new arrival is matched against a database consisting of evidence images; because nonevidence guns have been confiscated, nonevidence images are not added to the database for future matching.

Our model incorporates spatial and temporal information about pairs of images, which consist of an arrival and a database entry. The spatial proximity between a random arrival and a random database entry is described by a categorical variable denoted by the subscript  $j = 0, \dots, J$ . As described later, we use  $J = 1$  for Israel and  $J = 0, 1$  or  $2$  for the U.S., depending upon how NIBIN searches are performed. Let  $r_j$  be the probability that a random arrival and a random database entry have spatial proximity status  $j$  for  $j = 0, \dots, J$ .

The temporal proximity between an arrival and a random entry in a database is described by a continuous random variable with the subscript  $a$  for age, which is the time of acquisition of the arrival minus the time of acquisition of a random database entry (i.e., newly acquired evidence is added to the database immediately after it undergoes matching).

Let  $h(a)$  be the probability density function (PDF) of the random age.

In practice, ballistic imaging software involves an initial filtering step (e.g., by caliber, firing pin shape, and the number, twist rate and width of the rifling) followed by an investigation of multiple aspects of the cartridge casing or bullet [2]. For example, the matching of casings incorporates similarity scores for the breech face, firing pin and ejector mark, and two-dimensional bullet matching examines all possible rotations of bullets [2]. Although the mathematical modeling of multimodal matching is tractable (see [4, 5] for examples in biometrics), these detailed data are owned by the software vendor, which has published only aggregate performance curves. Since it is not possible to identify joint PDFs for similarity scores of the multiple aspects (e.g., breech face, firing pin, ejector mark) from aggregate performance curves, we assume in our model that an aggregate similarity score is generated as a result of comparing each arrival to each database entry, and let  $F(x)$  be the intra-gun cumulative distribution function (CDF) of the similarity score between two images emanating from the same gun, and  $G(y)$  be the inter-gun CDF of similarity scores between two images from different guns.

Before introducing our decision variables, we describe how the probability of a true match between an arrival and a database entry is affected by the evidence status and spatial and temporal proximity. One complexity in ballistic imaging that does not typically arise in biometric matching is that an arrival can match multiple entries in the evidence database. Let  $g_i(n)$  be the probability that an arrival of evidence status  $i$  has  $n$  true matches (during our statistical analysis, we differentiate between true matches and detected matches) in the database, for  $i = 0, 1$  and  $n = 0, 1, 2, \dots$ . Let a true match with arrival evidence status  $i$  (i.e., the arrival in this matching pair has evidence status  $i$ ) have probability  $p_{ij}$  of being in spatial proximity category  $j$ , for  $i = 0, 1$  and  $j = 0, \dots, J$ . Also, let  $h_m(a)$  be the PDF of the age of true matches between a random arrival and a random database entry.

As noted earlier, the matching software for ballistic images is rank-based with a candidate list of a fixed size, where the top (e.g., 10) matches generated by a new arrival are forwarded to a forensic firearms examiner for final verification. In our model, we use a threshold-based system, where similarity scores above the threshold are forwarded for human verification; this system generates a candidate list of variable size. In §4 of the Supporting Material, we show that the performance is very similar for these two systems, and later we discuss the implications of our threshold-based assumption.

Since the goal of this study is to exploit spatial, temporal and evidence status data to improve performance, we let the similarity score threshold depend on the evidence status of the arrival and the spatial proximity category and age of the pair being matched. We denote our decision variables by  $t_{ija}$ . Because  $i$  and  $j$  are categorical and the age  $a$  is a continuous quantity, we restrict  $t_{ija}$ 's dependence on  $a$  to take on a specific functional form. After testing linear, power and exponential functions, we settled on the exponential function,

$$t_{ija} = a_{ij}e^{b_{ij}a} + c_{ij}. \quad (1)$$

Hence, our optimization is over the  $6(J + 1)$  variables,  $\{a_{ij}, b_{ij}, c_{ij}, i = 0, 1, j = 0, \dots, J\}$ .

We are now in a position to formulate our optimization problem. The objective of our optimization problem is to maximize the probability that if at least one true match for an arrival exists in the database, then we detect at least one true match; we call this quantity the detection probability. Note that this probability is much more easily calculated than the expected number of true matches detected, which may be a more natural performance measure. However, because each database entry has already been through the matching process, it is often the case that if an arrival has more than one match, it is already known that these matching entries in the database are also linked to each other. Noting that the probability of at least one true match in the database for an arrival is  $\sum_{i=0}^1 q_i \sum_{n=1}^{\infty} g_i(n)$ , and that a true match involving an arrival of evidence status  $i$  and a pair of spatial proximity

category  $j$  goes undetected with probability  $\int_0^\infty F(t_{ija})h_m(a) da$ , it follows that our objective function is

$$\max_{a_{ij}, b_{ij}, c_{ij}} \frac{\sum_{i=0}^1 q_i \sum_{n=1}^\infty g_i(n) [1 - (\sum_{j=1}^J p_{ij} \int_0^\infty F(t_{ija})h_m(a) da)^n]}{\sum_{i=0}^1 q_i \sum_{n=1}^\infty g_i(n)}. \quad (2)$$

As with most problems of this type, our goal is to maximize the detection probability subject to some type of constraint on the false positive rate. Because the rank-based approach forwards 10 candidates per arrival to a forensic examiner (although Israel forwards the top 30 candidates – 10 each from firing pin, breech face and ejector mark – we consider 10 in total, because there may be significant overlap in the three candidate lists), a natural constraint for our threshold-based approach would be to force the expected size of the candidate list (i.e., the expected number of similarity scores that exceed the threshold) per arrival to be no larger than 10. Because this quantity is very difficult to compute, we take an alternative approach and require the expected number of false positives per arrival to be no more than the expected number of false positives generated by a constant threshold system (i.e, the threshold does not vary with  $i$ ,  $j$  or  $a$ ) when the expected candidate list size is 10. In §5 of the Supporting Material, we show that this false positive constraint behaves nearly the same as if we used a constraint on the mean candidate list size. Let  $t_h$ , which is derived from data in our statistical analysis, be the threshold used in a constant threshold system that generates an expected candidate list size of 10. Canceling the database size,  $N$ , from both sides of the constraint, we obtain our false positive constraint,

$$\sum_{i=0}^1 \sum_{j=0}^J q_i r_j \int_0^\infty [1 - G(t_{ija})]h(a) da \leq 1 - G(t_h), \quad (3)$$

and our optimization problem is given by (1)-(3).

We solve (1)-(3) using a sequential quadratic programming algorithm (via the `fmincon` function in MATLAB [6]). Because the optimization problem does not possess the second-order properties required to guarantee that the algorithm converges to a global optimum,

we compared local optima resulting from various starting points in the large (12- or 18-dimensional) decision variable space to increase the likelihood that we are achieving a near-optimal solution.

**Statistical Analysis** In addition to the main scenario of Israeli cartridge casings, we crudely extrapolate our model to six scenarios in the U.S.: three for cartridge casings and three for bullets. The U.S. database associated with NIBIN is divided into 47 partitions that are grouped into 12 regions, and so the three scenarios for each type of ballistic image corresponds to the three possible geographical approaches to matching: each arrival undergoes only intra-partition searches, only intra-region searches, or national searches. We refer to these three approaches as partition, regional and national approaches.

For each of these seven scenarios, we have nine quantities to estimate, which naturally divide into four groups: (i) the similarity score CDFs  $F(x)$  and  $G(y)$ , (ii) the probabilities  $q_i$ ,  $r_j$  and  $p_{ij}$ , (iii) the PMF  $g_i(n)$  and the historical threshold  $t_h$ , and (iv) the age PDFs  $h(a)$  and  $h_m(a)$ . All parameter values appear in Table 1, and the detailed derivations of these values, along with graphs of the five probability distributions, for the various scenarios are given in the Supporting Material. Here we provide a broad overview of the parameter estimation procedure, beginning with the similarity score CDFs, followed by the other parameter values for Israeli casings, for the U.S. casings scenarios, and finally the three U.S. bullets scenarios.

We assume that the intra-gun and inter-gun similarity score CDFs,  $F(x)$  and  $G(y)$ , are lognormal for both casings and bullets, and estimate the values of the four parameters – denoted by  $\mu_F$ ,  $\sigma_F$ ,  $\mu_G$  and  $\sigma_G$  – in §1 of the Supporting Material. For cartridge casings, these parameter values are estimated using the lower left performance curve in Fig. 12 of [7], which already incorporates an initial filtering step (restricting to 9 mm Luger cartridge casings) and multiple measurements (the curve is generated using similarity scores for breech face, firing pin and ejector mark). For bullets, we estimate  $F(x)$  and  $G(y)$  from experimental

results using the new BulletTrax-3D technology. The inter-gun CDF  $G(y)$  is estimated from Fig. 2 of [8] while the intra-gun CDF  $F(x)$  is estimated from the lead bullet results in [9]. Since lead is softer than copper and brass, and leaves poorer quality images, the use of lead results serves as a conservative assumption.

In the Israeli cartridge casings data set, the arrivals data consist of all matches detected between January 1, 2006 and December 31, 2008, as well as the evidence status of each arrival during this time period. There were 7138 arrivals during 2006-2008, and 697 of these arrivals matched at least one entry in the database. An arrival can match multiple entries in the database, and there were a total of 1364 matching pairs (i.e., matches between an arrival and a database entry). The evidence database contains entries since 1980, and its average size during 2006-2008 was 11,350 entries, which covers the entire country. We know which of the 89 Israeli police precincts collected each arrival and each database entry. Because many pairs of precincts generated no matches during 2006-2008, we use only two spatial categories (i.e.,  $J = 1$ ): intra-location and inter-location. However, to increase the amount of intra-location matching and fully exploit the data, we first use a graph partitioning algorithm that consolidates the number of locations from 89 to 52 (see §2.1 of the Supporting Material for details); we refer to these 52 locations as stations. These data allow us to estimate the probabilities  $q_i$ ,  $r_j$  and  $p_{ij}$  in a straightforward manner (§2.1 of Supporting Material).

From the raw data pertaining to the matches and arrivals, we can construct the PMF  $g_i^*(m)$ , which is the probability of detecting  $m$  matches from an arrival that has evidence status  $i$ ; this PMF is not to be confused with  $g_i(n)$ , which is the PMF for true (i.e., detected plus undetected) matches. In §2.2 of the Supporting Material, we use the PMF  $g_i^*(m)$  to estimate the historical threshold  $t_h$  for a constant threshold policy that generates an average candidate list size of 10, in terms of the database size  $N$ , which is also included in Table 1.

The most challenging part of our estimation procedure is to estimate the PMF  $g_i(n)$  of

true matches from the observed PMF  $g_i^*(m)$  of detected matches. In §2.2 of the Supporting Material, we estimate the PMF  $g_i(n)$  of true matches by stating a set of invertible linear equations that relate  $g_i(n)$ , the observed PMF  $g_i^*(m)$  of detected matches, and the unknown probabilities of detecting  $m$  matches given that  $n$  matches exist and the arrival has evidence status  $i$ . These probabilities are then estimated – as a function of  $n$ ,  $m$  and the false negative probability  $F(t_h)$  under the constant threshold system – by constructing a hidden Markov model whose hidden state transitions describe the detailed evolution of how the  $n$  true matches are assembled into detected groups as they sequentially arrive and are matched to previous arrivals.

Lastly, we assume that the age PDF of true matches is the same as the observed age PDF of detected matches in the Israeli database, and estimate  $h_m(a)$  by a lognormal after accounting for the lack of precise age data for 10% of the data (§2.3 of the Supporting Material). We only know the year in which each database entry was acquired, and we assume that the age equals December 31, 2010 minus the acquisition date. We fit these annual data to a piecewise cubic hermite interpolating polynomial to derive  $h(a)$  (§2.3 of the Supporting Material).

Turning to the U.S. casings parameters (§3 of the Supporting Material), the  $q_i$  probabilities are found from NIBIN data [1]. The  $r_j$  and  $p_{ij}$  probabilities depend on the geographic approach:  $J = 0$  for the partition approach,  $J = 1$  for the regional approach, and  $J = 2$  (where  $j = 0, 1, 2$  corresponds to intra-partition, inter-partition and intra-region, and inter-region, respectively) for the national approach. For lack of data in estimating  $r_j$ , we assume that each of the 47 partitions has the same arrival rate and the same number of database entries, and we make the same assumption for each of the 12 regions. Moreover, for lack of data in estimating  $p_{ij}$ , because the number of partitions in the U.S. is similar to the number of stations in Israel (47 vs. 52), we simply use the Israeli values in the U.S. regional approach;

in the U.S. national approach, because we have no data, we arbitrarily assume that half of inter-partition matches are intra-region and half are inter-region. Consequently, our results include a sensitivity analysis with respect to the  $p_{ij}$  values.

Although we have the evidence database size for casings in NIBIN [1], we use a somewhat different approach to estimate  $g_i(n)$  and  $t_h$  because we do not have data on  $g_i^*(m)$ , as we did for Israel. Hence, we approach this problem in reverse by first estimating  $g_i(n)$ . We assume that  $g_i(n)$  under the U.S. national approach is the same as the Israeli  $g_i(n)$ . Then we solve jointly for  $t_h$  and  $g_i^*(m)$  using the hidden Markov model. A similar approach is used for the partition and regional approaches except that we first use a binomial model to randomly thin  $g_i(n)$  to account for the fact that some of the true matches will occur outside of the area being searched.

For the U.S. age distribution, we assume that  $h_m(a)$ , the age PDF of true matches, is the same in the U.S. as it is in Israel. To estimate  $h(a)$  from the annual arrival rates to the U.S., we sample randomly from a uniform distribution within each year based on the annual Israeli data to compute the ages, and then fit the sampled ages to a lognormal distribution.

Finally, for the three U.S. bullets scenarios (§3 of the Supporting Material), we know the evidence database size  $N$  for bullets and the fraction of arrivals of each evidence status, which yield different values of  $q_i$  and  $t_h$ . All other parameter values (except for  $F(x)$  and  $G(y)$ ) are the same as in the three U.S. casings scenarios.

**Results.** Under the constant threshold policy (i.e., which employs the threshold  $t_h$ ) for cartridge casings in Israel, the probability that at least one true match for an arrival is detected, given that at least one true match exists, is 0.931. This detection probability increases to 0.987 under the optimal policy derived from equations (1)-(3), which represents a 81.4% reduction (from 0.069 to 0.013) in the false negative rate. The optimal thresholds from (1) are given in the last row of Table 3 of the Supporting Material, and are higher for

inter-station matches, nonevidence arrivals and older ages (Fig. 1). By optimizing each of the three types of information in isolation and in pairs (Table 3 in the Supporting Material), we find that optimizing age offers slightly more improvement than optimizing spatial information, while optimizing evidence status provides very little improvement. In addition, the impact of optimizing age and spatial information is subadditive.

For all three geographic approaches in the U.S. (Table 2), the detection probability is the probability that at least one true match for an arrival is detected, given that at least one true nationwide match exists. Because the three constant threshold policies and the optimal solutions under the partition and regional approaches are all feasible solutions to problem (1)-(3) for the national approach, the optimal solution to the national approach will always be optimal over all scenarios considered. For cartridge casings, the optimal policy generates only a modest improvement in detection probability over the constant threshold policy (the false negative probability is reduced by 20.3%) in the partition approach because only age and evidence status can be optimized. Note that the detection probability of 0.873 under the constant threshold policy for the partition approach is higher than the fraction of true matches that are intra-partition (which is  $\sum_{i=0}^1 q_i p_{i0} = 0.844$ ) because some arrivals generate multiple matches in the database. Although the constant threshold policy has a detection probability of only 0.817 in the national approach, the optimal policy for cartridge casings in the national approach achieves a detection probability of 0.970, i.e., a 76.3% reduction in the false negative probability relative to the constant threshold policy under the partition approach, which can be viewed as the status quo policy for the U.S. This high detection probability is achieved by increasing the thresholds for inter-partition, intra-region scores, and using a very high threshold for inter-region scores (Fig. 11 in the Supporting Material).

We perform a sensitivity analysis on the  $p_{ij}$  values in the U.S. (Table 4 in the Supporting

Material), which were chosen somewhat arbitrarily due to lack of data. We increase the  $p_{00}$  and  $p_{10}$  values in the regional approach by multiplying both  $1 - p_{00}$  and  $1 - p_{10}$  by either  $1/3$  or  $2/3$  (so that the fraction of true matches that are intra-partition increases from 0.844 to 0.896 and 0.948), and we consider two values of the percentage of inter-partition matches that are inter-region ( $\frac{p_{i2}}{1-p_{i0}}$  for  $i = 0, 1$ ): the base-case value of 0.5 and also 0.2. Of the two parameters, the fraction of true matches that are intra-partition has a bigger impact, although the improvement achieved by the optimal solution in the national approach still reduces the false negative rate relative to the constant threshold policy under the partition approach by 72.6% and 66.5% when the fraction of matches that are intra-partition is 0.896 and 0.948, respectively. The performance gap between the optimal solution under the national approach and the constant threshold policy under the partition approach widens slightly when the fraction of inter-partition matches that are inter-region drops from 0.5 to 0.2.

For the U.S. bullet scenarios (Table 2), the optimal policy offers almost no improvement under the partition approach. Bullets achieves a higher detection probability than casings under the constant threshold policy in the partition and regional strategies because the database is smaller for bullets than for casings. Even though the detection probability of the constant threshold policy degrades to 0.360 under the national approach, the optimal policy under the national approach achieves a detection probability of 0.962, compared to 0.900 for the constant threshold policy under the partition approach.

**Discussion.** Our main result stems from the analysis of the Israeli casings scenario: exploiting information – particularly spatiotemporal information – that is extraneous to the ballistic imaging process can improve the performance of ballistic imaging systems. While the increase in detection probability from 0.931 to 0.987 is modest due to the high base-level detection probability, this improvement is impressive when viewed as a 81.4% reduction in

the false negative rate. This result confirms that crime guns and their crimes do indeed cluster in space and time, and this information can be exploited to solve more crimes. Although Israel performs nationwide searches, we can also analyze the counterfactual scenario in which Israel only performs intra-precinct searches. Even if all intra-precinct matches are detected, the detection probability is only 0.729, which – when compared to 0.931 – reveals the benefit of performing nationwide searches in Israel.

As discussed in detail below, the results for the six U.S. scenarios are more speculative than the Israeli results. Nonetheless, the U.S. results reveal the tradeoff under the constant threshold policy between improved coverage and deteriorating matching performance as the system expands from a partition approach to a regional approach and on to a national approach: the partition approach necessarily misses all inter-partition matches but performs better on intra-partition matches due to the small database that it searches, while the national approach computes a similarity score for all possible matches but its matching is less accurate due to the large size of the national database. However, the optimal solution to (1)-(3) allows us to largely bypass this tradeoff: by setting higher threshold levels for more distant (e.g., inter-regional) searches (Fig. 11 in the Supporting Material), we can get full coverage and suffer only a small degradation in matching accuracy.

As noted earlier, in areas of the U.S. where NIBIN is consistently used, the intra-partition approach is typically employed; indeed, NIBIN is designed to discourage national searches, which need to be requested separately for each region [2]. While our numerical results for U.S. cartridge casings may not be accurate, they offer sufficient improvement over the status quo (from 0.873 to 0.970 for casings and from 0.900 to 0.962 for bullets, and robust results in the sensitivity analysis for casings) to warrant a research effort that estimates the parameter values in Table 1 from data that are not in the open literature and recomputes the detection probabilities in Table 2. As noted in [2], progress in this area is problematic

because the sole vendor, Forensic Technology WAI, Inc., has much of the necessary data, and hence the National Institute of Standards and Technology, which works on certain technical aspects of ballistic imaging [10], may be in the best position to perform or enable future research. Finally, if it is deemed that an optimal regional or national approach is worthy of implementation, other organizational hurdles need to be overcome to incentivize local law enforcement departments to perform inter-partition searches [2].

Although we highlighted the spatial issues in our U.S. analysis, it is worth noting that the false negative rate under the partition approach is reduced by  $\approx 20\%$  by solely exploiting the temporal clustering of crimes committed by crime guns, under the assumption that the temporal clustering in the U.S. is the same as it is in Israel (Figs. 7 and 8 in the Supporting Material). However, for the NIBIN system to achieve this reduction, it would need to add the new images to the NIBIN database in a timely manner, which is not being done now [1]. Indeed, the temporal clustering suggests that NIBIN performance might improve if new images were entered into the NIBIN database in a Last-In First-Out (LIFO) manner rather than in First-in First-Out (FIFO) order.

The model in equations (1)-(3) can be applied in several other ways. The spatial categories are quite general and could be used to exploit spatial patterns in the illegal gun market in the U.S. [11]. The proposed Reference Ballistic Image Database (RBID), which would maintain a national database from firings of newly manufactured and imported guns, could be accommodated in our model by introducing a third type of evidence status (i.e.,  $i = 2$  would correspond to new guns). Although a national RBID was deemed to be impractical due to its large database size [2], this issue could be revisited using our approach, which would allow very high thresholds for new guns. Moreover, if RBID incorporated point-of-sale data, then our approach could use lower thresholds for the miniscule fraction of retailers that sell the majority of crime guns in the U.S. [12]. Note that since ballistic

imaging is a search process that is followed by human verification, the retailers who sell many crime guns would be unaffected by the increased false positive rate associated with their lower thresholds.

There are several limitations of our analysis of Israeli cartridge casings. The biggest shortcoming is that the actual problem deals with similarity scores that are based on multi-modal (breech face, firing pin, ejector mark) measurements that are possibly correlated and possibly repeated (Israel acquires two samples from each evidence and nonevidence gun that is recovered), and that come from a variety of gun and ammunition types. Due to the lack of raw similarity score data, we use a performance curve (Fig. 12 of [7]) for combined breech face, firing pin and ejector mark scores for 9 mm caliber guns. Hence, the detection probability of 0.931 under the constant threshold policy is not necessarily an accurate prediction of Israel’s current performance, although our predicted relative improvement of the optimal policy over the current threshold policy is likely to be robust.

A second limitation is that we use a threshold-based approach rather than the rank-based approach that is in current use. If similarity scores were independent of gun and ammunition type, then a threshold-based approach would perform at least as well as a rank-based approach. However, if similarity scores vary by gun and ammunition type (which is likely to be the case), then our optimal policy may not work well. There are two ways to adapt our ideas to this setting: (i) have the threshold  $t_{ijk}$  also depend on the gun and ammunition type, or (ii) change the optimization problem to a rank-based system, where the decision variables are changed from  $t_{ijk}$  to multiplicative scaling factors  $\theta_{ijk}$  (i.e., a similarity score  $s$  would be transformed to  $\theta_{ijk}s$ ), which need not vary by gun or ammunition type. The former approach is feasible (e.g., it has been used for fingerprints with different image qualities [13]) but tedious, and the latter approach is preferable.

The final limitation of the Israeli analysis is the implicit assumption that the prob-

abilities  $q_i$ ,  $r_j$  and  $p_{ij}$  do not vary over time. During 2006-2008 in Israel, these quantities were reasonably stable. Nonetheless, there are three concerns. The first is if criminals adapt their behavior as a result of the ballistic imaging system. In Boston, criminals were found not to increase their use of revolvers, which do not eject cartridge casings, as a result of the implementation of a casing imaging system [2], and so this concern may be unfounded, particularly given the system's lack of transparency from the criminal's viewpoint. The second concern is changes in the mobility patterns of crime guns, which could occur for a variety of reasons. A third concern is changes in police procedures (e.g., spatial reallocation of law enforcement resources). The latter two concerns can be partially mitigated by periodically updating the estimates of these probabilities.

In addition to these limitations, the U.S. scenarios suffer from several other serious shortcomings, which stem from the paucity of U.S. data. First, on the positive side, we have U.S. data on the database size  $N$  and the  $q_i$  probabilities, and the  $r_j$  probabilities are likely to be representative of a typical partition. Furthermore, although we assume (for lack of data) that the PMF  $g_i(n)$  and the age PDFs  $h(a)$  and  $h_m(a)$  for the U.S. are similar to the Israeli values, we suspect that our qualitative results are not very sensitive to the form of these probability distributions.

However, there are two questionable assumptions made regarding the  $p_{ij}$  probabilities. First, because the number of U.S. partitions (47) is approximately equal to the number of stations (52) in Israel, we assume that the U.S.  $p_{ij}$  values for the regional approach are equal to the known values from Israel; i.e., we assume that there is the same amount of spatial clustering of crime guns and their crimes in both countries. However, Israel has a much smaller land area than the U.S., and we may be underestimating the amount of U.S. crime that is intra-partition. We are not aware of any U.S. data that directly measures  $p_{ij}$ , e.g., the fraction of true matches between an arrival and a database entry that are

intra-partition, intra-regional and inter-regional. However,  $\approx 67\%$  of traced crime guns are recovered and purchased in the same state [11, 12, 14, 15], suggesting that at least two-thirds of true matches are intra-partition. Of the remaining 33% that originated out-of-state, it is unknown whether or not these guns were also used in crimes in their originating state. In this context, our base-case assumption that the fraction of matches that are intra-partition is 0.844 is consistent with approximately half of the guns that originate in other states being involved in crimes in their originating states. Although it is known that  $\approx 88\%$  of recovered crime guns changed hands [16] (including sales in the second-hand market and theft [15]), apparently many criminals prefer to obtain new guns so that they cannot be charged with crimes that they did not commit [17].

The second questionable assumption surrounding the  $p_{ij}$  probabilities is that the fraction of inter-partition crimes that are inter-region is arbitrarily set at 0.5. Sensitivity analysis regarding both of these  $p_{ij}$  assumptions, which allows considerably less mobility of crime guns than in the base case (Table 4 of the Supporting Material) shows that our results are quite robust: e.g., the optimal policy reduces the false negative rate by 68.0% relative to the status quo if the fraction of matches that are intra-partition is 0.948.

Finally, the three-dimensional bullet CDFs  $F(x)$  and  $G(y)$  are estimated from a fairly small experiment and may not be representative of actual performance. More generally, ballistic imaging technology is in a state of flux, with the vendor recently introducing three-dimensional ballistic imaging matching systems [18] for which very little published performance data exist. In addition, ballistic imaging systems may perform better in controlled experiments than in the field.

In conclusion, we develop a data-driven approach to improve the performance of ballistic imaging systems, and predict that it could reduce the false negative rate for casings by over 80% in Israel, using matching data from 2006-2008. This improvement is achieved by

requiring a very close match for pairs of images that are distant in space and/or time. An application of the model to the U.S. suggests that our approach – by favoring local matches – can largely bypass the tradeoff between increased coverage and decreased accuracy that plagues the typical implementation in the U.S. system. While our U.S. results are speculative, they are sufficiently intriguing (the optimal national system increases the detection probability from 0.873, which is achieved by the status quo approach, to 0.970, and in a sensitivity analysis the optimal approach always reduces the false negative rate by at least 68.0%) to suggest that the U.S. Department of Justice and/or the National Institute of Standards and Technology gather the necessary data to refine our U.S. analysis and determine whether an optimized national approach should be implemented.

## References

- [1] Office of the Inspector General, U.S. Department of Justice (2005) *The Bureau of Alcohol, Tobacco, Firearms and Explosives' National Integrated Ballistic Information Network program*. Audit Report 05-30, Washington, D.C.
- [2] Cork, DL, Rolph JE, Meieran ES, Petrie CV, eds. (2008) *Ballistic imaging*. National Academies Press, Washington, D.C.
- [3] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council (2009) *Strengthening forensic science in the United States: a path forward*. National Academies Press, Washington, D.C.
- [4] Prabhakar S, Jain AK (2002) Decision-level fusion in fingerprint verification. *Pattern Recognition* 35:861-874.

- [5] Baveja M, Wein, LM (2009) An effective two-finger, two-stage biometric strategy for the US-VISIT Program. *Operations Research* 57:1068-1081.
- [6] MATLAB - the language of technical computing. Accessed at <http://www.mathworks.com/products/matlab/index.html> on September 5, 2011.
- [7] Beauchamp A, Roberge D (2005) Model of the behavior of the IBIS correlation scores in a large database of cartridge scores. Unpublished manuscript. Accessed at [www.forensictechnology.com/Default.aspx?app=LeadgenDownload&shortpath=docs/LargeDatabaseFinal.pdf](http://www.forensictechnology.com/Default.aspx?app=LeadgenDownload&shortpath=docs/LargeDatabaseFinal.pdf) on September 2, 2011.
- [8] Roberge D, Beauchamp A (2006) The use of BulletTRAX-3D in a study of consecutively manufactured barrels. *AFTE Journal* 38:166-172.
- [9] Binck TB (2008) Comparing the performance of IBIS and BulletTRAX-3D technology using bullets fired through 10 consecutively rifled barrels. *J. Forensic Science* 53:677-682.
- [10] Vorburger TV, Yen JH, Bachrach B, Renegar TB, Filliben JJ, Ma L *et al.* (2007) Surface topography analysis for a feasibility assessment of a national ballistics imaging database. National Institute of Standards and Technology Internal Report 7362, Gaithersburg, MD.
- [11] Wintemute GJ, Romero MP, Wright MA, Grassel KM (2004) The life cycle of crime guns: a description based on guns recovered from young people in California. *Annals Emergency Medicine* 43:733-742.
- [12] Wintemute GJ, Braga AA (2011) Opportunities for state-level action to reduce firearm violence: proceeding from the evidence. *Am. J. Public Health* 101:e1-e3.

- [13] Wein LM, Baveja M (2005) Using fingerprint image quality to improve the identification performance of the U.S. Visitor and Immigrant Status Indicator Technology Program. *PNAS* 102:7772-7775.
- [14] Bureau of Alcohol, Tobacco and Firearms (1999) Crime gun trace reports (2000). Department of the Treasury, Washington, D.C.
- [15] Pierce GL, Braga AA, Hyatt Jr. RR, Koper CS (2004) Characteristics and dynamics of illegal firearms markets: implications for a supply-side enforcement strategy. *Justice Quarterly* 21:391-422.
- [16] Bureau of Alcohol, Tobacco and Firearms (2000) Crime gun trace reports (2002). Department of the Treasury, Washington, D.C.
- [17] Cook PJ, Braga AA. Comprehensive firearms tracing: strategic and investigative uses of new data on firearms markets. *Arizona Law Review* 43:277-309.
- [18] IBIS TRAX-3D. Accessed at <http://www.forensictechnology.com/IBISTRAX/> on September 15, 2011.

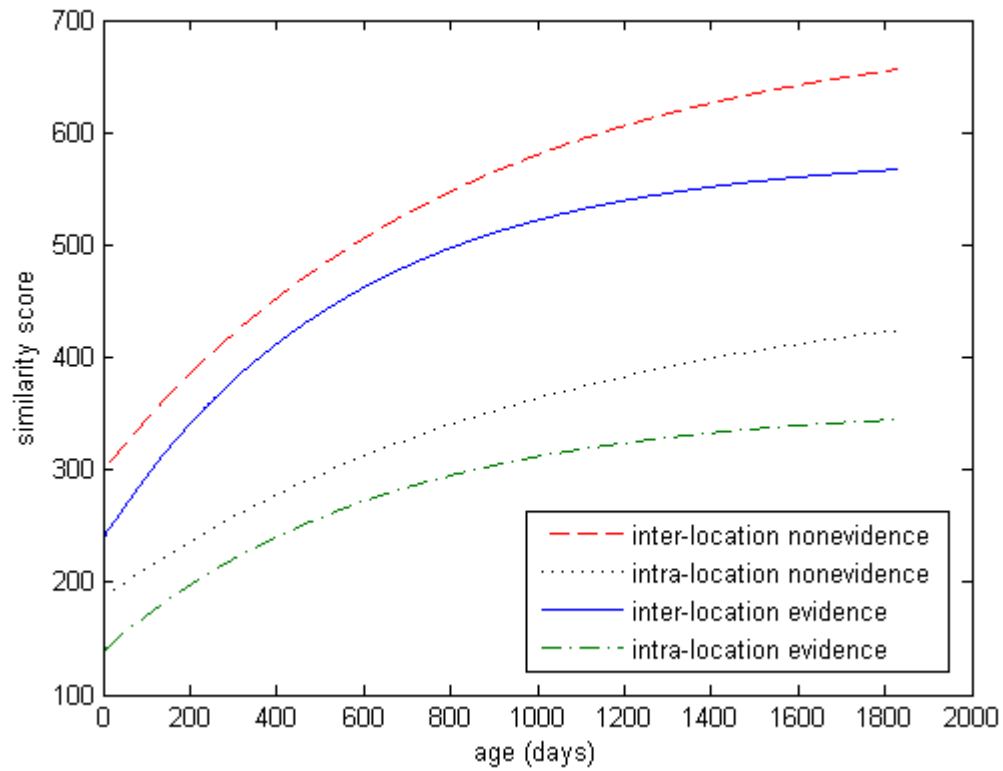


Figure 1: Optimal thresholds for Israeli cartridge casings.

Parameters	Definition	Israel	U.S. Partition	U.S. Regional	U.S. National
$\mu_F, \sigma_F$	Intra-gun score PDF parameters	Lognormal $\mu_F = 6.79, \sigma_F = 0.59$ (casings);	$\mu_F = 0.59$ (casings);	$\mu_F = 6.34, \sigma_F = 0.07$ (bullets)	
$\mu_G, \sigma_G$	Inter-gun score PDF parameters	Lognormal $\mu_G = 4.52, \sigma_G = 0.50$ (casings);	$\mu_G = 0.50$ (casings);	$\mu_G = 5.68, \sigma_G = 0.20$ (bullets)	
$q_i$	Arrival evidence status probability	$q_0 = 0.504, q_1 = 0.496$	$q_0 = 0.72, q_1 = 0.28$ (casings);	$q_0 = 0.81, q_1 = 0.19$ (bullets)	
$r_j$	Spatial proximity probability for arrival-database pair	$r_0 = 0.101,$ $r_1 = 0.899$	$r_0 = 1$	$r_0 = \frac{1}{12}$ $r_1 = \frac{11}{12}$	$r_0 = \frac{1}{47}$ $r_1 = \frac{1}{12} - \frac{1}{47}, r_2 = \frac{11}{12}$
$p_{ij}$	Spatial proximity probability for matching pair of evidence $i$	$p_{00} = 0.832$ $p_{10} = 0.875$	$p_{00} = p_{10} = 1$	$p_{00} = 0.832$ $p_{10} = 0.875$	$p_{00} = 0.832, p_{01} = 0.084$ $p_{10} = 0.815, p_{11} = 0.0625$
$N$	Evidence database size (casings)	11, 350	3304	12, 940	155, 282
$\bar{N}$	Evidence database size (bullets)	$N/A$	917	3591	43, 098
$t_h$	Historical threshold (casings)	442.6	364.9	450.6	627.6
$t_h$	Historical threshold (bullets)	$N/A$	465.8	512.9	594.0
$g_t(n)$	PMF for true matches		Fig. 6 and Fig. 9 in Supporting Material		
$h_m(a)$	Age PDF of matches		Lognormal $\mu_a = 4.90, \sigma_a = 1.69$		
$h(a)$	Age PDF of database records		Lognormal $\mu_a = 7.59, \sigma_a = 0.98$		

Table 1: Parameters values for the seven scenarios.

Policy	U.S. Casings Scenarios			U.S. Bullets Scenarios		
	Partition	Regional	National	Partition	Regional	National
Constant Threshold	0.873	0.881	0.817	0.900	0.897	0.360
Optimal Thresholds	0.899	0.946	0.970	0.902	0.938	0.962

Table 2: Detection probabilities for the six U.S. scenarios under both the constant threshold policy and the optimal policy derived from (1)-(3).