Bubble Trouble: Off-Line De-Anonymization of Bubble Forms

Joseph A. Calandrino, William Clarkson and Edward W. Felten
Department of Computer Science
Princeton University

Abstract

Fill-in-the-bubble forms are widely used for surveys, election ballots, and standardized tests. In these and other scenarios, use of the forms comes with an implicit assumption that individuals' bubble markings themselves are not identifying. This work challenges this assumption, demonstrating that fill-in-the-bubble forms could convey a respondent's identity even in the absence of explicit identifying information. We develop methods to capture the unique features of a marked bubble and use machine learning to isolate characteristics indicative of its creator. Using surveys from more than ninety individuals, we apply these techniques and successfully re-identify individuals from markings alone with over 50% accuracy. This bubble-based analysis can have either positive or negative implications depending on the application. Potential applications range from detection of cheating on standardized tests to attacks on the secrecy of election ballots. To protect against negative consequences, we discuss mitigation techniques to remove a bubble's identifying characteristics. We suggest additional tests using longitudinal data and larger datasets to further explore the potential of our approach in real-world applications.

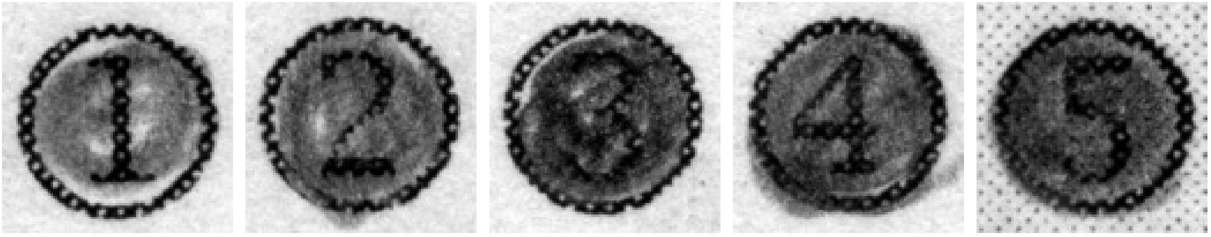
1 Introduction

Scantron-style fill-in-the-bubble forms are a popular means of obtaining human responses to multiple-choice questions. Whether conducting surveys, academic tests, or elections, these forms allow straightforward user completion and fast, accurate machine input. Although not every use of bubble forms demands anonymity, common perception suggests that bubble completion does not result in distinctive marks. We demonstrate that this assumption is false under certain scenarios, enabling use of these markings as a biometric. The ability to uncover identifying bubble marking patterns has far-reaching potential implications, from detecting cheating on standardized tests to threatening the anonymity of election ballots.

Bubble forms are widely used in scenarios where confirming or protecting the identity of respondents is critical. Over 137 million registered voters in the United States reside in precincts with optical scan voting machines [27], which traditionally use fill-in-the-bubble paper ballots. Voter privacy (and certain forms of fraud) relies on an inability to connect voters with these ballots. Surveys for research and other purposes use bubble forms to automate data collection. The anonymity of survey subjects not only affects subject honesty but also impacts requirements governing human subjects research [26]. Over 1.6 million members of the high school class of 2010 completed the SAT [8], one of many large-scale standardized tests using bubble sheets. Educators, testing services, and other stakeholders have incentives to detect cheating on these tests. The implications of our findings extend to any use of bubble forms for which the ability to "fingerprint" respondents may have consequences, positive or negative.

Our contributions. We develop techniques to extract distinctive patterns from markings on completed bubble forms. These patterns serve as a biometric for the form respondent. To account for the limited characteristics available from markings, we apply a novel combination of image processing and machine learning techniques to extract features and determine which are distinctive (see Section 2). These features can enable discovery of respondents' identities or of connections between completed bubbles.

To evaluate our results on real-world data, we use a corpus of over ninety answer sheets from an unrelated survey of high school students (see Section 3). We train on a subset of completed bubbles from each form, effectively extracting a biometric for the corresponding respondent. After training, we obtain a test set of addi-



(a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4 (e) Person 4 - Gray

Figure 1: Example marked bubbles. The background color is white in all examples except Figure 1(e), which is gray.

tional bubbles from each form and classify each test set. For certain parameters, our algorithms' top match is correct over 50% percent of the time, and the correct value falls in the top 3 matches 75% percent of the time. In addition, we test our ability to detect when someone other than the expected respondent completes a form, simultaneously achieving false positive and false negative rates below 10%. We conduct limited additional tests to confirm our results and explore details available from bubble markings.

Depending on the application, these techniques can have positive or negative repercussions (see Section 4). Analysis of answer sheets for standardized tests could provide evidence of cheating by test-takers, proctors, or other parties. Similarly, scrutiny of optical-scan ballots could uncover evidence of ballot-box stuffing and other forms of election fraud. With further improvements in accuracy, the methods developed could even enable new forms of authentication. Unfortunately, the techniques could also undermine the secret ballot and anonymous surveys. For example, some jurisdictions publish scanned images of ballots following elections, and employers could match these ballots against bubble-form employment applications. Bubble markings serve as a biometric even on forms and surveys otherwise containing no identifying information. We discuss methods for minimizing the negative impact of this work while exploiting its positive uses (see Section 5).

Because our test data is somewhat limited, we discuss the value of future additional tests (see Section 7). For example, longitudinal data would allow us to better understand the stability of an individual's distinguishing features over time, and stability is critical for most uses discussed in the previous paragraph.

2 Learning Distinctive Features

Filling in a bubble is a narrow, straightforward task. Consequently, the space for inadvertent variation is relatively constrained. The major characteristics of a filled-in bubble are consistent across the image population—most are relatively circular and dark in similar locations with slight imperfections—resulting in a largely homogeneous set. See Figure 1. This creates a challenge in capturing the unique qualities of each bubble and extrapolating a respondent's identity from them.

We assume that all respondents start from the same original state—an empty bubble with a number inscribed corresponding to the answer choice (e.g., choices 1-5 in Figure 1). When respondents fill in a bubble, opportunities for variation include the pressure applied to the drawing instrument, the drawing motions employed, and the care demonstrated in uniformly darkening the entire bubble. In this work, we consider applications for which it would be infeasible to monitor the exact position, pressure, and velocity of pencil motions throughout the coloring process.[1] In other contexts, such as signature verification, these details can be useful. This information would only strengthen our results and would be helpful to consider if performing bubble-based authentication, as discussed in Section 4.

2.1 Generating a Bubble Feature Vector

Image recognition techniques often use feature vectors to concisely represent the important characteristics of an image. As applied to bubbles, a feature vector should capture the unique ways that a mark differs from a perfectly completed bubble, focusing on characteristics that tend to distinguish respondents. Because completed bubbles tend to be relatively homogeneous in shape, many common metrics do not work well here. To measure the unique qualities, we generate a feature vector that blends several approaches from the image recognition literature. Specifically, we use PCA, shape descriptors, and a custom bubble color distribution to generate a feature vector for each image.

[1] Clarkson et al. [7] use multiple scans to infer the 3D surface texture of paper, which may suggest details like pressure. We assume multiple scans to be infeasible for our applications.

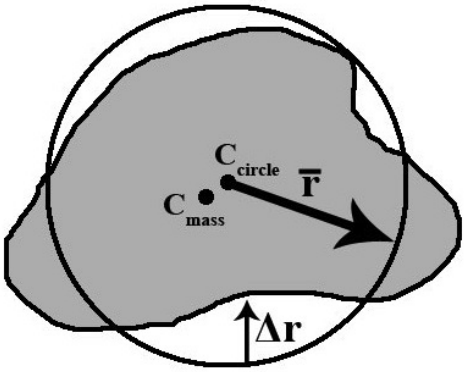


Figure 2: An example bubble marking with an approximating circle. The circle minimizes the sum of the squared deviation from the radius. We calculate the circle's center and mean radius, the marking's variance from the radius, and the marking's center of mass.

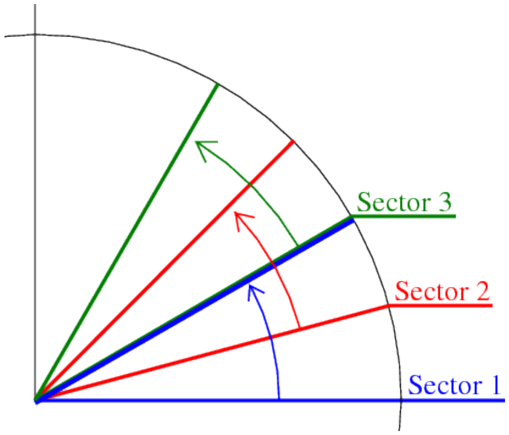


Figure 3: Each dot is split into twenty-four 15° slices. Adjacent slices are combined to form a sector, spanning 30°. The first few sectors are depicted here.

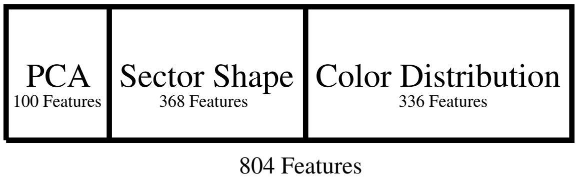


Figure 4: Feature vector components and their contributions to the final feature vector length.

Principal Component Analysis (PCA) is one common technique for generating a feature set to represent an image [16]. At a high level, PCA reduces the dimensionality of an image, generating a concise set of features that are statistically independent from one another. PCA begins with a sample set of representative images to generate a set of eigenvectors. In most of our experiments, the representative set was comprised of 368 images and contained at least one image for each (respondent, answer choice) pair. Each representative image is normalized and treated as a column in a matrix. PCA extracts a set of eigenvectors from this matrix, forming a basis. We retain the 100 eigenvectors with the highest weight. These eigenvectors account for approximately 90% of the information contained in the representative images.

To generate the PCA segment of our feature vector, a normalized input image (treated as a column vector) is projected onto the basis defined by the 100 strongest eigenvectors. The feature vector is the image's coordinates in this vector space—i.e., the weights on the eigenvectors. Because PCA is such a general technique, it may fail to capture certain context-specific geometric characteristics when working exclusively with marked bubbles.

To compensate for the limitations of PCA, we capture shape details of each bubble using a set of geometric descriptors and capture color variations using a custom metric. Peura et al. [24] describe a diverse a set of geometric descriptors that measure statistics about various shapes. This set includes a shape's center of mass, the center and radius of a circle approximating its shape, and variance of the shape from the approximating circle's radius (see Figure 2). The approximating circle minimizes the sum of squared radius deviations. We apply the specified descriptors to capture properties of a marked bubble's boundary. Instead of generating these descriptors for the full marked bubble alone, we also generate the center of mass, mean radius, and radial variance for "sectors" of the marked bubble. To form these sectors, we first evenly divide each dot into twenty-four 15° "slices." Sectors are the 24 overlapping pairs of adjacent slices (see Figure 3). Together, these geometric descriptors add 368 features.

Finally, we developed and use a simple custom metric to represent color details. We divide a dot into sectors as in the previous paragraph. For each sector, we create a histogram of the grayscale values for the sector consisting of fifteen buckets. We throw away the darkest bucket, as these pixels often represent the black ink of the circle border and answer choice numbering. Color distribution therefore adds an additional 14 features for each sector, or a total of 336 additional features.

The resulting feature vector consists of 804 features that describe shape and color details for a dot and each of its constituent sectors (see Figure 4). See Section 3.3, where we evaluate the benefits of this combination of features. Given feature vectors, we can apply machine learning techniques to infer distinguishing details and differentiate between individuals.

2.2 Identifying Distinguishing Features

Once a set of feature vectors are generated for the relevant dots, we use machine learning to identify and utilize the important features. Our analysis tools make heavy use of Weka, a popular Java-based machine learning workbench that provides a variety of pre-implemented learning methods [12]. In all experiments, we used Weka version 3.6.3.

We apply Weka's implementation of the Sequential Minimal Optimization (SMO) supervised learning algorithm to infer distinctive features of respondents and classify images. SMO is an efficient method for training support vector machines [25]. Weka can accept a training dataset as input, use the training set and learning algorithm to create a model, and evaluate the model on a test set. In classifying individual data points, Weka internally generates a distribution over possible classes, choosing the class with the highest weight. For us, this distribution is useful in ranking the respondents believed to be responsible for a dot. We built glue code to collect and process both internal and exposed Weka data efficiently.

3 Evaluation

To evaluate our methods, we obtained a corpus of 154 surveys distributed to high school students for research unrelated to our study. Although each survey is ten pages, the first page contained direct identifying information and was removed prior to our access. Each of the nine available pages contains approximately ten questions, and each question has five possible answers, selected by completing round bubbles numbered 1-5 (as shown in Figure 1).

From the corpus of surveys, we removed any completed in pen to avoid training on writing utensil or pen color.[2] Because answer choices are numbered, some risk exists of training on answer choice rather than marking patterns—e.g., respondent X tends to select bubbles with "4" in the background. For readability, survey questions alternate between a white background and a gray background. To avoid training bias, we included only surveys containing at least five choices for each answer 1-4 on a white background (except where stated otherwise), leaving us with 92 surveys.

For the 92 surveys meeting our criteria, we scanned the documents using an Epson v700 Scanner at 1200 DPI. We developed tools to automatically identify, extract, and label marked bubbles by question answered and choice selected. After running these tools on the scanned images, we manually inspected the resulting images to ensure accurate extraction and labeling.

Due to the criteria that we imposed on the surveys, each survey considered has at least twenty marked bubbles on a white background, with five bubbles for the "1" answer, five for the "2" answer, five for the "3" answer, and five for the "4" answer.[3] For each experiment, we selected our training and test sets randomly from this set of twenty bubbles, ensuring that sets have equal numbers of "1," "2," "3," and "4" bubbles for each respondent and trying to balance the number of bubbles for each answer choice when possible.

In all experiments, a random subset of the training set was selected and used to generate eigenvectors for PCA. We required that this subset contain at least one example from each respondent for each of the four relevant answer choices but placed no additional constraints on selection. For each dot in the training and test sets, we generated a feature vector using PCA, geometric descriptors, and color distribution, as described in Section 2.1.

We conducted two primary experiments and a number of complementary experiments. The first major test explores our ability to re-identify a respondent from a test set of eight marks given a training set of twelve marks per respondent. The second evaluates our ability to detect when someone other than the official respondent completes a bubble form. To investigate the potential of bubble markings and confirm our results, we conducted seven additional experiments. We repeated each experiment ten times and report the average of these runs.

Recall from Section 2.2 that we can rank the respondents based on how strongly we believe each one to be responsible for a dot. For example, the respondent that created a dot could be the first choice or fiftieth choice of our algorithms. A number of our graphs effectively plot a cumulative distribution showing the percent of test cases for which the true corresponding respondent falls at or above a certain rank—e.g., for 75% of respondents in the test set, the respondent's true identity is in the top three guesses.

3.1 Respondent Re-Identification

This experiment measured the ability to re-identify individuals from their bubble marking patterns. For this test, we trained our model using twelve sample bubbles per respondent, including three bubbles for each answer choice 1-4. Our test set for each respondent contained the remaining two bubbles for each answer choice, for a total of eight test bubbles. We applied the trained model

[2] We note that respondents failing to use pencil or to complete the survey anecdotally tended not to be cautious about filling in the bubbles completely. Therefore, these respondents may be more distinguishable than those whose surveys were included in our experiment.

[3] To keep a relatively large number of surveys, we did not consider the number of "5" answers and do not use these answers in our analysis.

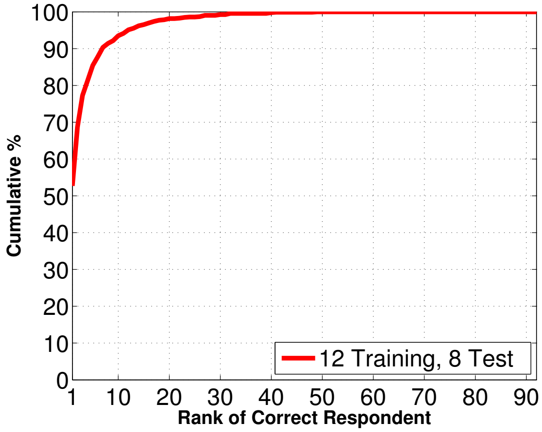


Figure 5: Respondent re-identification with 12 training bubbles and 8 test bubbles per respondent.

to each of the 92 respondents' test sets and determined whether the predicted identity was correct.

To use multiple marks per respondent in the test set, we classify the marks individually, yielding a distribution over the respondents for each mark in the set. After obtaining the distribution for each test bubble in a group, we combine this data by averaging the values for each respondent. Our algorithms then order the respondents from highest to lowest average confidence, with highest confidence corresponding to the top choice.

On average, our algorithm's first guess identified the correct respondent with 51.1% accuracy. The correct respondent fell in the top three guesses 75.0% of the time and in the top ten guesses 92.4% of the time. See Figure 5, which shows the percentage of test bubbles for which the correct respondent fell at or above each possible rank. This initial result suggests that individuals complete bubbles in a highly distinguishing manner, allowing re-identification with surprisingly high accuracy.

3.2 Detecting Unauthorized Respondents

One possible application of this technique is to detect when someone other than the authorized respondent creates a set of bubbles. For example, another person might take a test or survey in place of an authorized respondent. We examined our ability to detect these cases by measuring how often our algorithm would correctly detect a fraudulent respondent who has claimed to be another respondent. We trained our model using twelve training samples from each respondent and examined the output of our model when presented with eight test bubbles. The distribution of these sets is the same as in Section 3.1.

For these tests, we set a threshold for the lowest rank accepted as the respondent. For example, suppose that

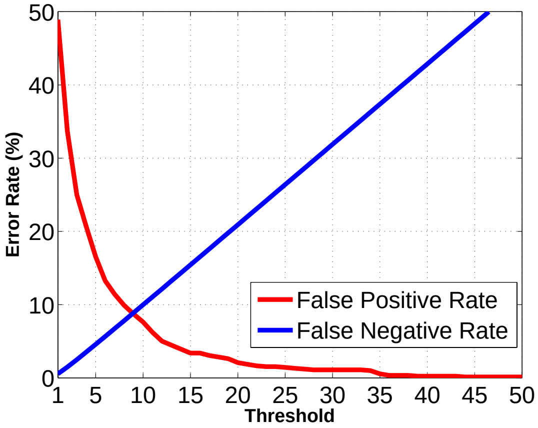


Figure 6: False positive and false negative rates when detecting unauthorized respondents.

the threshold is 12. To determine whether a given set of test bubbles would be accepted for a given respondent, we apply our trained model to the test set. If the respondent's identity appears in any of the top 12 (of 92) positions in the ranked list of respondents, that test set would be accepted for the respondent. For each respondent, we apply the trained model both to the respondent's own test bubbles and to the 91 other respondents' test bubbles.

We used two metrics to assess the performance of our algorithms in this scenario. The first, false positive rate, measures the probability that a given respondent would be rejected (labeled a cheater) for bubbles that the respondent actually completed. The second metric, false negative rate, measures the probability that bubbles completed by any of the 91 other respondents would be accepted as the true respondent's. We varied the threshold from 1 to 92 for our tests. We expected the relationship between threshold and false negative rate to be roughly linear: increasing the threshold by 1 increases the probability that a respondent randomly falls above the threshold for another respondent's test set by roughly 1/92.[4]

Our results are presented in Figure 6. As we increase the threshold, the false positive rate drops precipitously while the false negative rate increases roughly linearly. If we increase the threshold to 8, then a fraudulent respondent has a 7.8% chance of avoiding detection (by being classified as the true respondent), while the true respondent has a 9.9% chance of being mislabeled a cheater. These error rates intersect with a threshold approximately equal to 9, where the false positive and false negative rates are 8.8%.

[4] This is not exact because the order of these rankings is not entirely random. After all, we seek to rank a respondent as highly as possible for the respondent's own test set.

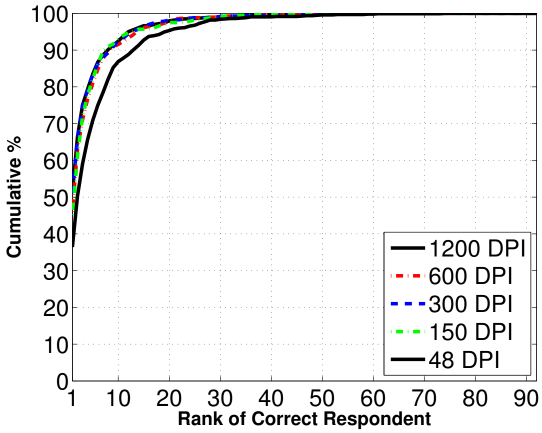


Figure 7: Respondent re-identification accuracy using lower-resolution images. Note that the 1200, 600, 300, and 150 DPI lines almost entirely overlap.

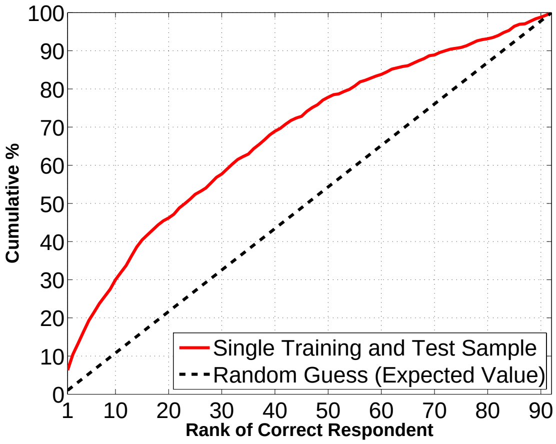


Figure 8: One marked bubble per respondent in each of the training and test sets. The expected value from random guessing is provided as reference.

3.3 Additional Experiments

To study the information conveyed by bubble markings and support our results, we performed seven complementary experiments. In the first, we evaluate the effect that scanner resolution has on re-identification accuracy. Next, we considered our ability to re-identify a respondent from a single test mark given a training set containing a single training mark from each respondent. Because bubble forms typically contain multiple markings, this experiment is somewhat artificial, but it hints at the information available from a single dot. The third and fourth supplemental experiments explored the benefits of increasing the training and test set sizes respectively while holding the other set to a single bubble. In the fifth test, we examined the tradeoff between training and test set sizes. The final two experiments validated our results using additional gray bubbles from the sample surveys and demonstrated the benefits of our feature set over PCA alone. As with the primary experiments, we repeated each experiment ten times.

Effect of resolution on accuracy. In practice, high-resolution scans of bubble forms may not be available, but access to lower resolution scans may be feasible. To determine the impact of resolution on re-identification accuracy, we down-sampled each ballot from the original 1200 DPI to 600, 300, 150, and 48 DPI. We then repeated the re-identification experiment of Section 3.1 on bubbles at each resolution.

Figure 7 shows that decreasing the image resolution has little impact on performance for resolutions above 150 DPI. At 150 DPI, the accuracy of our algorithm's first guess decreases to 45.1% from the 51.1% accuracy observed at 1200 DPI. Accuracy remains relatively strong even at 48 DPI, with the first guess correct 36.4% of the time and the correct respondent falling in the top ten guesses 86.8% of the time. While down-sampling may not perfectly replicate scanning at a lower resolution, these results suggest that strong accuracy remains feasible even at resolutions for which printed text is difficult to read.

Single bubble re-identification. This experiment measured the ability to re-identify an individual using a single marked bubble in the test set and a single example per respondent in the training set. This is a worst-case scenario, as bubble forms typically contain multiple markings. We extracted two bubbles from each survey and trained a model using the first bubble.[5] We then applied the trained model to each of the 92 second bubbles and determined whether the predicted identity was correct. Under these constrained circumstances, an accuracy rate above that of random guessing (approximately 1%) would suggest that marked bubbles embed distinguishing features.

On average, our algorithm's first guess identified the correct respondent with 5.3% accuracy, five times better than the expected value for random guessing. See Figure 8, which shows the percentage of test bubbles for which the correct respondent fell at or above each possible rank. The correct respondent was in the top ten guesses 31.4% of the time. This result suggests that individuals can inadvertently convey information about their

[5] Note: In this experiment, we removed the restriction that the set of images used to generate eigenvectors for PCA contains an example from each column.

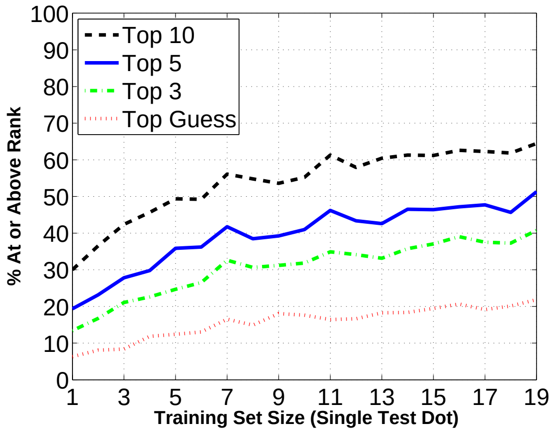


Figure 9: Increasing the training set size from 1 to 19 dots per respondent.

identities from even a single completed bubble.

Increasing training set size. In practice, respondents rarely fill out a single bubble on a form, and no two marked bubbles will be exactly the same. By training on multiple bubbles, we can isolate patterns that are consistent and distinguishing for a respondent from ones that are largely random. This experiment sought to verify this intuition by confirming that an increase in the number of training samples per respondent increases accuracy. We held our test set at a single bubble for each respondent and varied the training set size from 1 to 19 bubbles per respondent (recall that we have twenty total bubbles per respondent).

Figure 9 shows the impact various training set sizes had on whether the correct respondent was the top guess or fell in the top 3, 5, or 10 guesses. Given nineteen training dots and a single test dot, our first guess was correct 21.8% of the time. The graph demonstrates that a greater number of training examples tends to result in more accurate predictions, even with a single-dot test set. For the nineteen training dots case, the correct respondent was in the top 3 guesses 40.8% of the time and the top 10 guesses 64.5% of the time.

Increasing test set size. This experiment is similar to the previous experiment, but we instead held the training set at a single bubble per respondent and varied the test set size from 1 to 19 bubbles per respondent. Intuitively, increasing the number of examples per respondent in the test set helps ensure that our algorithms guess based on consistent features—even if the training set is a single noisy bubble.

Figure 10 shows the impact of various test set sizes on whether the correct respondent was the top guess or fell in the top 3, 5, or 10 guesses. We see more gradual improvements when increasing the test set size than observed when increasing training set size in the previous test. From one to nineteen test bubbles per respondent, the accuracy of our top 3 and 5 guesses increases relatively linearly with test set size, yielding maximum improvements of 4.3% and 7.6% respectively. For the top-guess case, accuracy increases with test set size from 5.3% at one bubble per respondent to 8.1% at eight bubbles then roughly plateaus. Similarly, the top 10 guesses case plateaus near ten bubbles and has a maximum improvement of 8.0%. Starting from equivalent sizes, the marginal returns from increasing the training set size generally exceed those seen as test set size increases. Next, we explore the tradeoff between both set sizes given a fixed total of twenty bubbles per respondent.

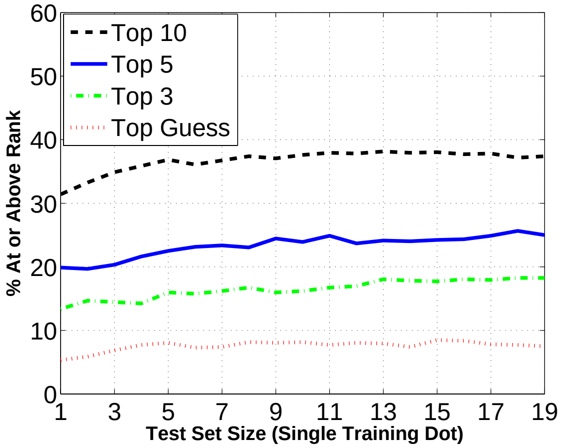


Figure 10: Increasing the test set size from 1 to 19 dots per respondent.

Training-test set size tradeoff. Because we have a constraint of twenty bubbles per sample respondent, the combined total size of our training and test sets per respondent is limited to twenty. This experiment examined the tradeoff between the sizes of these sets. For each value of x from 1 to 19, we set the size of the training set per respondent to x and the test set size to $20 - x$. In some scenarios, a person analyzing bubbles would have far larger training and test sets than in this experiment. Fortunately, having more bubbles would not harm performance: an analyst could always choose a subsample of the bubbles if it did. Therefore, our results provide a lower bound for these scenarios.

Figure 11 shows how varying training/test set sizes affected whether the correct respondent was the top guess or fell in the top 3, 5, or 10 guesses. As the graph demonstrates, the optimal tradeoff was achieved with roughly

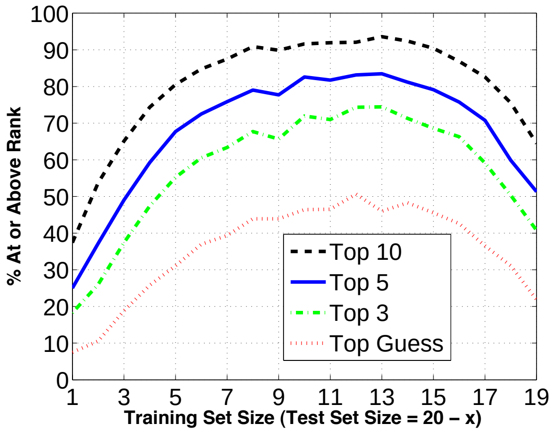


Figure 11: Trade-off between training and test set sizes.

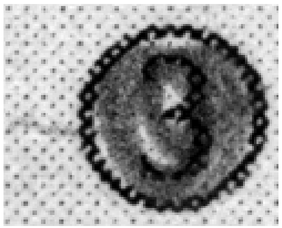


Figure 12: This respondent tends to have a circular pattern with a flourish stroke at the end. The gray background makes the flourish stroke harder to detect.

twelve bubbles per respondent in the training set and eight bubbles per respondent in the test set.

Validation with gray bubbles. To further validate our methods, we tested the accuracy of our algorithms with a set of bubbles that we previously excluded: bubbles with gray backgrounds. These bubbles pose a significant challenge as the paper has both a grayish hue and a regular pattern of darker spots. This not only makes it harder to distinguish between gray pencil lead and the paper background but also limits differences in color distribution between users. See Figure 12.

As before, we selected surveys by locating ones with five completed (gray) bubbles for each answer choice, 1-4, yielding 97 surveys. We use twelve bubbles per respondent in the training set and eight bubbles in the test set, and we apply the same algorithms and parameters for this test as the test in Section 3.1 on a white background.

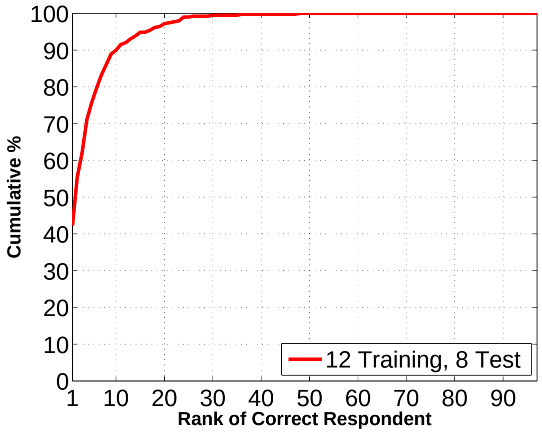


Figure 13: Using the unmodified algorithm with the same configuration as in Figure 5 on dots with gray backgrounds, we see only a mild decrease in accuracy.

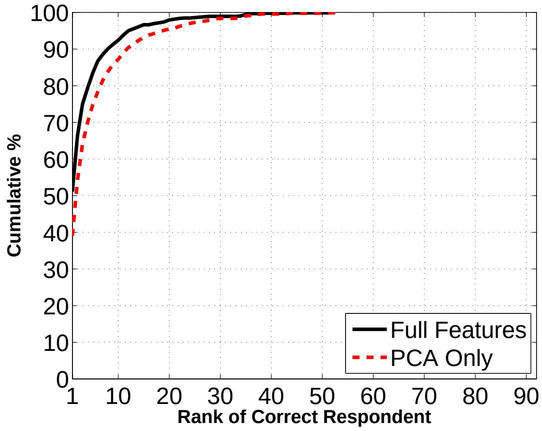
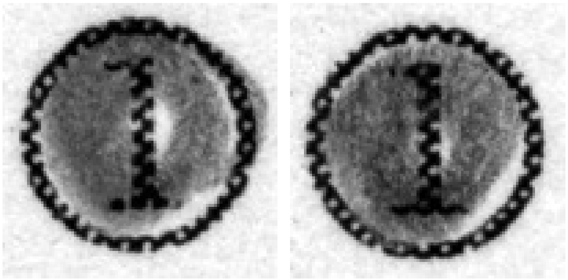


Figure 14: Performance with various combinations of features.

Figure 13 shows the percentage of test cases for which the correct respondent fell at or above each possible rank. Our first guess is correct 42.3% of the time, with the correct respondent falling in the top 3, 5, and 10 guesses 62.1%, 75.8%, and 90.0% of the time respectively. While slightly weaker than the results on a white background for reasons specified above, this experiment suggests that our strong results are not simply a byproduct of our initial dataset.

Feature vector options. As discussed in Section 2.1, our feature vectors combine PCA data, shape descriptors, and a custom color distribution to compensate for the limited data available from bubble markings. We tested the performance of our algorithms for equivalent parameters with PCA alone and with all three features combined. This test ran under the same setup as Figure 5 in Section 3.1.

For both PCA and the full feature set, Figure 14 shows the percentage of test cases for which the correct respondent fell at or above each possible rank. The additional features improve the accuracy of our algorithm's first



(a) Person A

(b) Person B

Figure 15: Bubbles from respondents often mistaken for each other. Both respondents use superficially similar techniques, leaving unmarked space in similar locations.

guess from 39.0% to 51.1% and the accuracy of the top ten guesses from 87.2% to 92.4%.

3.4 Discussion

Although our accuracy exceeds 50% for respondent re-identification, the restrictive nature of marking a bubble limits the distinguishability between users. We briefly consider a challenging case here.

Figure 15 shows marked bubbles from two respondents that our algorithm often mistakes for one another. Both individuals appear to use similar techniques to complete a bubble: a circular motion that seldom deviates from the circle boundary, leaving white-space both in the center and at similar locations near the border. Unless the minor differences between these bubbles are consistently demonstrated by the corresponding respondents, differentiating between these cases could prove quite difficult. The task of completing a bubble is constrained enough that close cases are nearly inevitable. In spite of these challenges, however, re-identification and detection of unauthorized respondents are feasible in practice.

4 Impact

This work has both positive and negative implications depending on the context and application. While we limit our discussion to standardized tests, elections, surveys, and authentication, the ability to derive distinctive bubble completion patterns for individuals may have consequences beyond those examined here. In Section 7, we discuss additional tests that would allow us to better assess the impact in several of these scenarios. In particular, most of these cases assume that an individual's distinguishing features remain relatively stable over time, and tests on longitudinal data are necessary to evaluate this assumption.

4.1 Standardized Tests

Scores on standardized tests may affect academic progress, job prospects, educator advancement, and school funding, among other possibilities. These high stakes provide an incentive for numerous parties to cheat and for numerous other parties to ensure the validity of the results. In certain cheating scenarios, another party answers questions on behalf of an official test-taker. For example, a surrogate could perform the entire test, or a proctor could change answer sheets after the test [11, 17]. The ability to detect when someone other than the authorized test-taker completes some or all of the answers on a bubble form could help deter this form of cheating.

Depending on the specific threat and available data, several uses of our techniques exist. Given past answer sheets, test registration forms, or other bubbles ostensibly from the same test-takers, we could train a model as in Section 3.2 and use it to infer whether a surrogate completed some or all of a test.[6] Although the surrogate may not be in the training set, we may rely on the fact that the surrogate is less likely to have bubble patterns similar to the authorized test-taker than to another set of test-takers. Because our techniques are automated, they could flag the most anomalous cases—i.e., the cases that would be rejected even under the least restrictive thresholds—in large-scale datasets for manual review.

If concern exists that someone changed certain answers after the test (for example, a proctor corrected the first five answers for all tests), we could search for questions that are correctly answered at an usually high rate. Given this information, two possible analysis techniques exists. First, we could train on the less suspicious questions and use the techniques of Section 3.2 to determine whether the suspicious ones on a form are from the same test-taker. Alternatively, we could train on the non-suspicious answer choices from each form and the suspicious answer choices from *all* forms other than a form of interest. Given this model, we could apply the techniques of Section 3.1 to see whether suspicious bubbles on that form more closely match less-suspicious bubbles on the same form or suspicious bubbles on other forms.

4.2 Elections

Our techniques provide a powerful tool for detecting certain forms of election fraud but also pose a threat to voter privacy.

Suppose that concerns exist that certain paper ballots were fraudulently submitted by someone other than a valid voter. Although the identity of voters might not be known for direct comparison to past ballots or other

[6] Note our assumption that the same unauthorized individual has not completed both the training bubbles and the current answer sheet.

forms, we can compare batches of ballots in one election to batches from past elections. Accounting for demographic changes and the fact that voters need not vote in all elections, the ballots should be somewhat similar across elections. For example, if 85% of the voters on the previous election's sign-in list also voted during this election, we would expect similarities between 85% of the old ballots and the new ballots.

To test this, we may train on ballots from the previous election cycle and attempt to re-identify new ballots against the old set. We would not expect ballots to perfectly match. Nevertheless, if less than approximately 85% of the old "identities" are covered by the new ballots or many of the new ballots cluster around a small number of identities, this would raise suspicion that someone else completed these forms, particularly if the forms are unusually biased towards certain candidates or issues.

Similarly, analysis could also help uncover fraudulent absentee ballots. Because absentee ballots do not require a voter to be physically present, concerns exist about individuals fraudulently obtaining and submitting these ballots [19]. By training a model on past ballots, we could assess whether suspicious absentee ballots fail to match the diversity expected from the population completing these forms.[7]

Unfortunately, because bubble markings can serve as a biometric, they can also be used in combination with seemingly innocuous auxiliary data to undermine ballot secrecy. Some jurisdictions now release scanned images of ballots following elections with the goal of increasing transparency (e.g., Humboldt County, California [14], which releases ballot scans at 300 DPI). If someone has access to these images or otherwise has the ability to obtain ballot scans, they can attempt to undermine voter privacy. Although elections may be decided by millions of voters, an attacker could focus exclusively on ballots cast in a target's precinct. New Jersey readjusts larger election districts to contain fewer than 750 registered voters [22]. Assuming 50% turnout, ballots in these districts would fall in groups of 375 or smaller. In Wisconsin's contentious 2011 State Supreme Court race, 71% of reported votes cast fell in the 91% of Wisconsin wards with 1,000 or fewer total votes [28].

Suppose that an interested party, such as a potential employer, wishes to determine how you voted. Given the ability to obtain bubble markings known to be from you (for example, on an employment application), that party can replicate our experiment in Section 3.1 to isolate one or a small subset of potential corresponding ballots. What makes this breach of privacy troubling is that it occurs without the consent of the voter and requires no special access to the ballots (unlike paper fingerprinting techniques [4], which require access to the physical ballot). The voter has not attempted to make an identifying mark, but the act of voting results in identifying marks nonetheless. This threat exists not only in traditional government elections but also in union and other elections.

Finally, one known threat against voting systems is pattern voting. For this threat, an attacker coerces a voter to select a preferred option in a relevant race and an unusual combination of choices for the other races. The unusual voting pattern will allow the attacker to locate the ballot later and confirm that the voter selected the correct choice for the relevant race. One proposed solution for pattern voting is to cut ballots apart to separate votes in individual contests [6]. Our work raises the possibility that physically divided portions of a ballot could be connected, undermining this mitigation strategy.[8]

4.3 Surveys

Human subjects research is governed by a variety of restrictions and best practices intended to balance research interests against the subjects' interests. One factor to be considered when collecting certain forms of data is the level of anonymity afforded to subjects. If a dataset contains identifying information, such as subject name, this may impact the types of data that should be collected and procedural safeguards imposed to protect subject privacy. If subjects provide data using bubble forms, these markings effectively serve as a form of identifying information, tying the form to the subject even in the absence of a name. Re-identification of subjects can proceed in the same manner as re-identification of voters, by matching marks from a known individual against completed surveys (as in Section 3.1).

Regardless of whether ethical or legal questions are raised by the ability to identify survey respondents, this ability might affect the honesty of respondents who are aware of the issue. Dishonesty poses a problem even for commercial surveys that do not adhere to the typical practices of human subjects research.

The impact of this work for surveys is not entirely negative, however. In certain scenarios, the person responsible for administering a survey may complete the forms herself or modify completed forms, whether to avoid the work of conducting the survey or to yield a desired outcome. Should this risk exist, similar analysis to the standardized test and election cases could help uncover the issue.

[7] If a state uses bubble form absentee ballot applications, analysis could even occur on the applications themselves.

[8] We thank an anonymous reviewer for suggesting this possibility.

4.4 Authentication

Because bubble markings are a biometric, they may be used alone or in combination with other techniques for authentication. Using a finger or a stylus, an individual could fill in a bubble on a pad or a touchscreen. Because a computer could monitor user input, various details such as velocity and pressure could also be collected and used to increase the accuracy of identification, potentially achieving far stronger results than in Section 3.2. On touchscreen devices, this technique may or may not be easier for users than entry of numeric codes or passwords. Additional testing would be necessary for this application, including tests of its performance in the presence of persistent adversaries.

5 Mitigation

The impact of this paper's techniques can be both beneficial and detrimental, but the drawbacks may outweigh the benefits under certain circumstances. In these cases, a mitigation strategy is desirable, but the appropriate strategy varies. We discuss three classes of mitigation strategies. First, we consider changes to the forms themselves or how individuals mark the forms. Second, we examine procedural safeguards that restrict access to forms or scanned images. Finally, we explore techniques that obscure or remove identifying characteristics from scanned images. No strategy alone is perfect, but various combinations may be acceptable under different circumstances.

5.1 Changes to Forms or Marking Devices

As explored in Section 3.3, changes to the forms themselves such as a gray background can impact the accuracy of our tests. The unintentional protection provided by this particular change was mild and unlikely to be insurmountable. Nevertheless, more dramatic changes to either the forms themselves or the ways people mark them could provide a greater measure of defense.

Changes to the forms themselves should strive to limit either the space for observable human variation or the ability of an analyst to perceive these variations. The addition of a random speckled or striped background in the same color as the writing instrument could create difficulties in cleanly identifying and matching a mark. If bubbles had wider borders, respondents would be less likely to color outside the lines, decreasing this source of information. Bubbles of different shapes or alternate marking techniques could encourage less variation between users. For example, some optical scan ballots require a voter simply to draw a line to complete an arrow shape [1], and these lines may provide less identifying information than a completed bubble.

The marking instruments that individuals use could also help leak less identifying information. Some Los Angeles County voters use ink-marking devices, which stamp a circle of ink for a user [18]. Use of an ink-stamper would reduce the distinguishability of markings, and even a wide marker could reduce the space for inadvertent variation.

5.2 Procedural Safeguards

Procedural safeguards that restrict access to both forms themselves and scanned images can be both straightforward and effective. Collection of data from bubble forms typically relies on scanning the forms, but a scanner need not retain image data for any longer than required to process a respondent's choices. If the form and its image are unavailable to an adversary, our techniques would be infeasible.

In some cases, instructive practices or alternative techniques already exist. For example, researchers conducting surveys could treat forms with bubble markings in the same manner as they would treat other forms containing identifying information. In the context of elections, some jurisdictions currently release scanned ballot images following an election to provides a measure of transparency. This release is not a satisfactory replacement for a statistically significant manual audit of paper ballots (e.g., [2, 5]), however, and it is not necessary for such an audit. Because scanned images could be manipulated or replaced, statistically significant manual confirmation of the reported ballots' validity remains necessary. Furthermore, releasing the recorded choices from a ballot (e.g., Washington selected for President, Lincoln selected for Senator, etc.) without a scanned ballot image is sufficient for a manual audit.

Whether the perceived transparency provided by the release of ballot scans justifies the resulting privacy risk is outside the scope of this paper. Nevertheless, should the release of scanned images be desirable, the next section describes methods that strive to protect privacy in the event of a release.

5.3 Scrubbing Scanned Images

In some cases, the release of scanned bubble forms themselves might be desirable. In California, Humboldt County's release of ballot image scans following the 2008 election uncovered evidence of a software glitch causing certain ballots to be ignored [13]. Although a manual audit could have caught this error with high probability, ballot images provide some protection against unintentional errors in the absence of such audits.

The ability to remove identifying information from scanned forms while retaining some evidence of a respondent's actions is desirable. One straightforward approach is to cover the respondent's recorded choices with solid black circles. Barring any stray marks or misreadings, this choice would completely remove all identifying bubble patterns. Unfortunately, this approach has several disadvantages. First, a circle could cover choices that were not selected, hiding certain forms of errors. Second, suppose that a bubble is marked but not recorded. While the resulting image would allow reviewers to uncover the error, such marks retain a respondent's identifying details. The threat of a misreading and re-identification could be sufficient to undermine respondent confidence, enabling coercion.

An alternative to the use of black circles is to replace the contents of each bubble with its average color, whether the respondent is or is not believed to have selected the bubble. The rest of the scan could be scrubbed of stray marks. This would reduce the space for variation to color and pressure properties alone. Unfortunately, no evidence exists that these properties cannot still be distinguishing. In addition, an average might remove a respondent's intent, even when that intent may have been clear to the scanner interpreting the form. Similar mitigation techniques involve blurring the image, reducing the image resolution, or making the image strictly black and white, all of which have similar disadvantages to averaging colors.

One interesting approach comes from the facial image recognition community. Newton et al. [23] describe a method for generating k-anonymous facial images. This technique replaces each face with a "distorted" image that is k-anonymous with respect to faces in the input set. The resulting k-anonymous image maintains the expected visual appearance, that of a human face. The exact details are beyond the scope of this paper, but the underlying technique reduces the dimensionality using Principal Component Analysis and an algorithm for removing the most distinctive features of each face [23].

Application of facial anonymization to bubbles is straightforward. Taking marked and unmarked bubbles from all ballots in a set, we can apply the techniques of Newton et al. to each bubble, replacing it with its k-anonymous counterpart. The result would roughly maintain the visual appearance of each bubble while removing certain unique attributes. Unfortunately, this approach is imperfect in this scenario. Replacement of an image might hide a respondent's otherwise obvious intent. In addition, distinguishing trends might occur over multiple bubbles on a form: for example, an individual might mark bubbles differently near the end of forms (this is also a problem for averaging the bubble colors). Finally, concerns exist over the guarantees provided by k-anonymity [3], but the work may be extensible to achieve other notions of privacy, such as differential privacy [9].

We caution that the value of these images for proving the true contents of physical bubble forms is limited: an adversary with access to the images, whether scrubbed or not, could intentionally modify them to match a desired result. These approaches are most useful where the primary concern is unintentional error.

6 Related Work

Biometrics. Biometrics can be based on physical or behavioral characteristics of an individual. Physical biometrics are based on physical characteristics of a person, such as fingerprints, facial features, and iris patterns. Behavioral biometrics are based on behavioral characteristics that tend to be stable and difficult to replicate, such as speech or handwriting/signature [15]. Bubble completion patterns are a form of behavioral biometric.

As a biometric, bubble completion patterns are similar to handwriting, though handwriting tends to rely on a richer, less constrained set of available features. In either case, analysis may occur on-line or off-line [21]. In an on-line process, the verifying party may monitor characteristics like stroke speed and pressure. In an off-line process, a verifying party only receives the resulting data, such as a completed bubble. Handwriting-based recognition sometimes occurs in an on-line setting. Because off-line recognition is more generally applicable, our analysis occurred purely in an off-line manner. In some settings, such as authentication, on-line recognition would be possible and could yield stronger results.

Document re-identification. Some work seeks to re-identify a precise physical document for forgery and counterfeiting detection (e.g., [7]). While the presence of biometrics may assist in re-identification, the problems discussed in this paper differ. We seek to discover whether sets of marked bubbles were produced by the same individual. Our work is agnostic towards whether the sets come from the same form, different forms, or duplicates of forms. Nevertheless, our work and document re-identification provide complementary techniques. For example, document re-identification could help determine whether the bubble form (ballot, answer sheet, survey, etc.) provided to an individual matches the one returned or detect the presence of fraudulently added forms.

Cheating Detection. Existing work uses patterns in answer choices themselves as evidence of cheating. Myagkov et al. [20] uncover indicators of election fraud using aggregate vote tallies, turnout, and historical data. Similarly, analysis of answers on standardized tests can be particularly useful in uncovering cheating [10, 17].

For example, if students in a class demonstrate mediocre overall performance on a test yet all correctly answer a series of difficult questions, this may raise concerns of cheating. The general strategy in this line of research is to look for answers that are suspicious in the context of either other answers or auxiliary data.

Bubble-based analysis is also complementary to these anti-cheating measures. Each technique effectively isolates sets of suspicious forms, and the combination of the two would likely be more accurate than each independently. Although our techniques alone do not exploit contextual data, they have the advantage of being unbiased by that data. If a student dramatically improves her study habits, the resulting improvement in test scores alone might be flagged by other anti-cheating measures but not our techniques.

7 Future Work

Although a variety of avenues for future work exist, we focus primarily on possibilities for additional testing and application-specific uses here.

Our sample surveys allowed a diverse set of tests, but access to different datasets would enable additional useful tests. We are particularly interested in obtaining and using longitudinal studies—in which a common set of respondents fill out bubble forms multiple times over some period—to evaluate our methods. While providing an increased number of examples, this could also identify how a respondent's markings vary over time, establish consistency over longer durations, and confirm that our results are not significantly impacted by writing utensil. Because bubble forms from longitudinal studies are not widely available, this might entail collecting the data ourselves.

While we tested our techniques using circular bubbles with numbers inscribed, numerous other form styles exist. In some cases, respondents instead fill in ovals or rectangles. In other cases, selection differs dramatically from the traditional fill-in-the-shape approach—for example, the line-drawing approach discussed in Section 5 bears little similarity to our sample forms. Testing these cases would not only explore the limits of our work but could also help uncover mitigation strategies.

Section 4 discusses a number of applications of our techniques. Adapting the techniques to work in these scenarios is not always trivial. For example, Section 6 discusses existing anti-cheating techniques for standardized tests. Combining the evidence provided by existing techniques and ours would strengthen anti-cheating measures, but it would also require some care to process the data quickly and merge results.

Use of bubble markings for authentication would require both additional testing and additional refinement of our techniques. Given a dataset containing on-line information, such as writing instrument position, velocity, and pressure, we could add this data to our feature vectors and test the accuracy of our techniques with these new features. This additional information could increase identifiability considerably—signature verification is commonly done on-line due to the utility of this data—and may yield an effective authentication system. Depending on the application, a bubble-based authentication system would potentially need to work with a finger rather than a pen or stylus. Because the task of filling in a bubble is relatively constrained, this application would require cautious testing to ensure that an adversary cannot impersonate a legitimate user.

8 Conclusion

Marking a bubble is an extremely narrow task, but as this work illustrates, the task provides sufficient expressive power for individuals to unintentionally distinguish themselves. Using a dataset with 92 individuals, we demonstrate how to re-identify a respondent's survey with over 50% accuracy. In addition, we are able to detect an unauthorized respondent with over 92% accuracy with a false positive rate below 10%. We achieve these results while performing off-line analysis exclusively, but on-line analysis has the potential to achieve even higher rates of accuracy.

The implications of this study extend to any system utilizing bubble forms to obtain user input, especially cases for which protection or confirmation of a respondent's identity is important. Additional tests can better establish the threat (or benefit) posed in real-world scenarios. Mitigating the amount of information conveyed through marked bubbles is an open problem, and solutions are dependent on the application. For privacy-critical applications, such the publication of ballots, we suggest that groups releasing data consider means of masking respondents' markings prior to publication.

Acknowledgements

We are grateful to Ian Davey, Deven Desai, Ari Feldman, Adam Finkelstein, Joe Hall, Josh Kroll, Tim Lee, Szymon Rusinkiewicz, Harlan Yu, and Bill Zeller for helpful suggestions and assistance in conducting this research. We thank Joel Trachtenberg, Jennifer Minsky, and Derrick Higgins for assistance in obtaining test data. We thank Carolyn Crnich, Kevin Collins, and Mitch Trachtenberg of the Humboldt County Election Transparency Project for discussing the implications of our work on that project. We thank our anonymous reviewers and shepherd, Claudia Diaz, for comments on the paper.

References

[1] ALAMEDA COUNTY, CALIFORNIA. Alameda County Voting System Demonstration. http://www.acgov/org/rov/votingsystemdemo.htm.

[2] ASLAM, J. A., POPA, R. A., AND RIVEST, R. L. On estimating the size and confidence of a statistical audit. In *Proc. 2007 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT '07)*.

[3] BRICKELL, J., AND SHMATIKOV, V. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (August 2008).

[4] CALANDRINO, J. A., CLARKSON, W., AND FELTEN, E. W. Some consequences of paper fingerprinting for elections. In *Proceedings of EVT/WOTE 2009* (Aug. 2009), D. Jefferson, J. L. Hall, and T. Moran, Eds., USENIX/ACCURATE/IAVoSS.

[5] CALANDRINO, J. A., HALDERMAN, J. A., AND FELTEN, E. W. Machine-assisted election auditing. In *Proc. 2007 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT '07)*.

[6] CARBACK, R. How secret is your secret ballot? part 1 of 3: Pattern voting. https://scantegrity.org/blog/2008/06/16/how-secret-is-your-secret-ballot-part-1-of-3-pattern-voting/, June 16 2008.

[7] CLARKSON, W., WEYRICH, T., FINKELSTEIN, A., HENINGER, N., HALDERMAN, J. A., AND FELTEN, E. W. Fingerprinting blank paper using commodity scanners. In *Proc of IEEE Symposium on Security and Privacy* (May 2009).

[8] COLLEGE BOARD. 2010 college-bound seniors results underscore importance of academic rigor. http://www.collegeboard.com/press/releases/213182.html.

[9] DWORK, C. Differential privacy. In *Proc of the 33rd International Colloquium on Automata, Language and Programming* (July 2006).

[10] GABRIEL, T. Cheaters find an adversary in technology. *New York Times* (December 27 2010).

[11] GABRIEL, T. Under pressure, teachers tamper with tests. *New York Times* (June 10 2010).

[12] GARNER, S. R. Weka: The waikato environment for knowledge analysis. In *In Proc. of the New Zealand Computer Science Research Students Conference* (1995), pp. 57–64.

[13] GREENSON, T. Software glitch yields inaccurate election results. *Times-Standard* (December 5 2008).

[14] HUMBOLDT COUNTY ELECTION TRANSPARENCY PROJECT. http://www.humetp.org/.

[15] JAIN, A., HONG, L., AND PANKANTI, S. Biometric Identification. *Communications of the ACM 43*, 2 (February 2000), 91–98.

[16] JOLLIFFE, I. T. *Principal Component Analysis*, second ed. Springer, October 2002.

[17] LEVITT, S. D., AND DUBNER, S. J. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. HarperCollins, 2006.

[18] LOS ANGELES COUNTY REGISTRAR-RECORDER / COUNTY CLERK. InkaVote Plus Manual. http://www.lavote.net/voter/pollworker/PDFS/INKAVOTE_PLUS_HANDBOOK.pdf, 2011.

[19] MCCUTCHEON, C. Absentee voting fosters trickery, trend's foes say. *Times Picayune* (October 2006).

[20] MYAGKOV, M., ORDESHOOK, P. C., AND SHAKIN, D. *The Forensics of Election Fraud: Russia and Ukraine*. Cambridge University Press, 2009.

[21] NALWA, V. S. Automatic on-line signature verification. In *Proceedings of the Third Asian Conference on Computer Vision-Volume I - Volume I* (London, UK, 1997), ACCV '98, Springer-Verlag, pp. 10–15.

[22] NEW JERSEY LEGISLATURE. N.J.S.A. 19:4-13 (1976).

[23] NEWTON, E., SWEENEY, L., AND MALIN, B. Preserving privacy by de-identifying facial images. *IEEE Transactions on Knowledge and Data Engineering 17* (2005), 232–243.

[24] PEURA, M., AND IIVARINEN, J. Efficiency of simple shape descriptors. In *In Aspects of Visual Form* (1997), World Scientific, pp. 443–451.

[25] PLATT, J. C. Sequential minimal optimization: A fast algorithm for training support vector machines, 1998.

[26] U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES. IRB guidebook. http://www.hhs.gov/ohrp/irb/irb_guidebook.htm.

[27] VERIFIED VOTING FOUNDATION. The verifier. http://www.verifiedvoting.org/verifier/.

[28] WISCONSIN GOVERNMENT ACCOUNTABILITY BOARD. Spring 2011 election results. http://gab.wi.gov/elections-voting/results/2011/spring.