Using Behavioral Data to Identify Interviewer Fabrication in Surveys

Benjamin Birnbaum
CSE Department
Univeristy of Washington
birnbaum@cs.washington.edu

Gaetano Borriello
CSE Department
University of Washington
gaetano@cse.washington.edu

Abraham D. Flaxman
IHME
University of Washington
abie@uw.edu

Brian DeRenzi
CSE Department
University of Washington
bderenzi@cse.uw.edu

Anna R. Karlin
CSE Department
University of Washington
karlin@cs.washington.edu

ABSTRACT

Surveys conducted by human interviewers are one of the principal means of gathering data from all over the world, but the quality of this data can be threatened by interviewer fabrication. In this paper, we investigate a new approach to detecting interviewer fabrication automatically. We instrument electronic data collection software to record logs of low-level behavioral data and show that supervised classification, when applied to features extracted from these logs, can identify interviewer fabrication with an accuracy of up to 96%. We show that even when interviewers know that our approach is being used, have some knowledge of how it works, and are incentivized to avoid detection, it can still achieve an accuracy of 86%. We also demonstrate the robustness of our approach to a moderate amount of label noise and provide practical recommendations, based on empirical evidence, on how much data is needed for our approach to be effective.

Author Keywords

user logging; data collection; data quality; behavioral data; supervised classification; surveys; curbstoning; HCI4D

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Surveys conducted by human interviewers are widespread. They are used in many contexts, including—to name a few examples—censuses, academic studies, clinical trials, and vaccination drives. They are especially important in the world's least developed countries, where 72% of people live in rural areas [42] and 41% of adults are illiterate [43]. In these countries, surveys may offer the *only* way to gather critical health and economic data that is needed to make informed resource allocation decisions.

CHI 2013, April 27–May 2, 2013, Paris, France.

Data quality is a major concern for any survey organization. One pernicious data quality problem is interviewer data fabrication, also known as *curbstoning*. Interviewers may fabricate data because they cannot reach some households, are uncomfortable asking sensitive questions, or get paid based on the number of surveys completed. Curbstoning has been discovered and reported in several surveys [6, 32, 38], and an "epidemic of suspected interview falsification" almost derailed at least one large epidemiological study [41]. In one survey from the United States Census Bureau, at least 6.5% of interviewers were found to be falsifying at least some of their data [38], and in another survey, 13% of interviewers admitted to fabricating at least part of an interview—despite being supervised in a telephone call center [28]. Such fabrication threatens the ability of decision-makers to use survey data effectively.

Recently, there has been an increase in the amount of data being collected electronically, either on laptops [1, 12] or on PDAs and mobile phones [15, 16, 21, 33, 34, 36]. Using these devices for data collection has been shown to improve efficiency [5, 23, 30]. We argue here that using them provides an additional advantage: it can make it easier to detect curbstoning. When interviewers conduct a survey with an electronic device, they leave a detailed trace of behavorial data, such as when they select and change answers, move between questions, and scroll. These traces, if recorded, may provide a strong signal indicating whether data is being fabricated. For example, curbstoners might change their answers more frequently or fill out data more quickly than interviewers who are collecting real data.

In this paper, we evaluate the potential of using this behavioral data to identify curbstoning. We instrument widely-used data collection software to record detailed event logs of low-level behavioral data, and then, using features extracted from these logs, we train a classifier on labeled data and use it to make predictions on new data.

We show that our approach can *accurately* identify curbstoning. On independent test data, our classifier achieved an accuracy of 96%, as compared to 77% without the behavioral data. We also show that our approach can *robustly* identify curbstoning. In realistic scenarios, interviewers who fabricate data have an incentive not to get caught and might learn how

their behavioral data is being recorded and monitored. We show that even when interviewers know that our approach is being used, have some knowledge of how it works, and are incentivized to avoid detection, it can detect curbstoning with an accuracy of 86%.

We also address two practical concerns facing a survey organization that wishes to implement our approach. First, we measure the sensitivity of our approach to label noise and show that the accuracy does not decrease very much for moderate (i.e., 10%) levels of label noise. Second, we measure the sensitivity of our approach to the size of the training set and use the results to provide recommendations on the training set size needed.

The rest of this paper is organized as follows. First, we explain how our contributions fit within related work in survey methodology and human–computer interaction. Second, we describe the software that we used for data collection and our methodology for recording behavioral data and extracting features from it. Third, we describe the design of the experiment that gave us ground-truth labels and that allowed us to assess the robustness of our approach to interviewer knowledge and incentive. Fourth, we describe the results of running our classifier on the data from our study. We conclude with a summary, discussion of the practical feasibility of our approach, and directions for future work.

RELATED WORK

Our work fits within two research areas: survey methodology on detecting curbstoning and human–computer interaction on automatically learning from behavioral data. We describe each in turn.

Detecting curbstoning

Survey methodologists began their formal study of curbstoning in 1945 with a seminal paper by Crespi [14]. In this paper, Crespi proposed several reasons that interviewers might fabricate data, including long questionnaires, complex or intrusive questions, unreasonable expectations, and hard to reach subjects. He suggested that survey organizations make careful survey design and management decisions to make cheating unlikely or unattractive. Just how to make these decisions was the subject of much of the early work on curbstoning [2, 17]. Our work complements this important area of research by providing automated support to detect curbstoning when it does occur.

In the late 1980s, researchers started investigating the potential of using the survey data itself—in addition to high-level behavioral data like completion time—to detect curbstoning. These data-driven approaches provide important insight into what may characterize fabricated data, including a bad fit to "Benford's Law" [6, 11, 27, 35, 39], missing contact information [24, 29, 32, 41], fast interviews [8, 29, 31, 32], and low data variance [6, 26, 35, 39]. Unlike our work, however, these studies do not use automatic approaches like supervised classification to identify fabricated data, nor do they validate their hypotheses using independent data sets.

More recently, researchers have begun to investigate the ability of supervised classification to predict fabrication or other data quality problems [4, 10, 31]. These papers show the promise of the approach, but, unlike our work, they do not investigate the potential of low-level behavioral data to improve performance.

Automatically learning from behavioral data

Recorded logs of behavioral data have long been used as a tool by researchers in human–computer interaction. Typically, these logs are used to understand the the usability of interfaces and to inform design choices (see [22] for a survey). Collecting behavioral data from interviewers and respondents in surveys (without using supervised classification to identify fabricated data) has also been proposed [13, 20, 40].

More recently, researchers have explored how supervised classification can be applied to features extracted from logs of behavioral data to make automatic inferences regarding users, such as whether they have pointing problems [25] or what their level of skill in using a software application is [18]. This approach has also been applied to detect data quality problems. Rzeszotarski and Kittur [37] show that supervised classification applied to recorded behavioral data could accurately identify low-quality data submitted by crowd-workers on Mechanical Turk.

Our work builds on the growing body of literature showing the promise of supervised classification on user-trace logs. We demonstrate the feasibility of this approach in a new application: detecting interviewer fabrication in surveys collected on electronic devices. This application is similar to Rzeszotarski and Kittur's in that it concerns data quality, but the context of our work—a live interviewer using a handheld device—differs from the context of theirs—a worker completing a task on Mechanical Turk. In a live interview, for example, there is an extra layer of interaction: instead of directly inputting data, the interviewer first interacts with the respondent and then enters the data. Live interviewers are also typically more focused than workers on Mechanical Turk, who may perform multiple tasks at once or leave a task and come back to it. And finally, the method of interaction differs: in our context, interviewers use their thumbs on a touchscreen device, whereas workers on Mechanical Turk can use a mouse, keyboard, and operations like copy-paste.

Apart from these differences, we note that one of our primary contributions is to show that this approach can detect fabrication even when participants know that it is being used, have some knowledge of how it works, and are incentivized to avoid detection.

SOFTWARE TOOLS AND INSTRUMENTATION

In this section, we describe the software used by interviewers in our study, called Open Data Kit (ODK) Collect [21], how we modified this software to record logs of behavioral data, and how we extracted features from these logs.

Open Data Kit Collect

ODK Collect, which runs on Android devices, allows users to record structured data using the touch screen of their device. In ODK Collect, surveys are specified by an XML form that provides the question text, response types, and branching

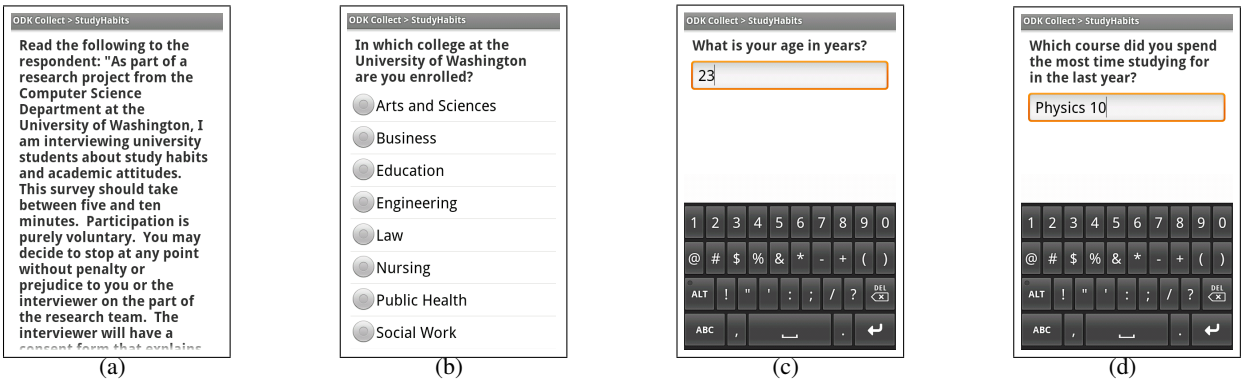

Figure 1. Screenshots from a survey on ODK Collect, showing (a) a note prompt, (b) a categorical question prompt, (c) a numeric question prompt, and (d) a free text question prompt.

Event type	Description	Parameters
answer selected	An answer is selected or changed for a categorical question prompt.	response selected
next	A prompt is reached from a forward (left) swipe.	
previous	A prompt is reached from a backward (right) swipe.	
scroll	The user scrolls up or down on a prompt.	amount scrolled in pixels (direction indicated by sign)
text changed	The text is changed for a numeric or free text question prompt.	answer text prior to change, answer text after change

Table 1. Log event types. In addition to the parameters listed in this table, each event was also tagged with a millisecond-precision timestamp, an instance identifier, and a prompt identifier.

logic of the survey. Each interview that a user conducts corresponds to one *instance* of the form. Each form consists of multiple *prompts*, which can be read-only *note prompts* or interactive *question prompts*. To move through the survey, users swipe the touch screen left (forwards) or right (backwards). To choose a response for a categorical question, users select the desired response with their finger. To enter a numeric or free text response, they use the keyboard on their phone (either a physical keyboard, if one exists, or the "soft" keyboard that appears on the touch screen.) If the entire prompt is too big to fit in the phone's screen, users can scroll through the prompt vertically. Figure 1 shows screenshots from four different prompts from the survey that we used for the study (described shortly).

Recording logs of behavioral data

To collect behavioral data on how interviewers interacted with ODK Collect, we modified the software to record detailed event logs. Each entry consisted of an *event type*, a millisecond-precision timestamp, an instance identifier, and a prompt identifier—along with various parameters specific to the event type. The possible event types were *answer selected*, *next*, *previous*, *scroll*, and *text changed*. Table 1 lists the event types, their semantics, and parameters. The logs were stored in SQLite databases on participants' phones.

Feature extraction

From these logs, we extracted a rich set of features. First, for each prompt, we extracted a set of *prompt-level features* that pertained only to that prompt. A description of some of these features is given in Table 2.

Second, we used the log files and the values of the prompt-level features to extract a set of *instance-level features* that pertained to the entire interview. A description of some of these features is given in Table 3. These features were either statistical aggregations of the prompt-level features, such as total-answer-time, or features extracted directly from the logs that involved more than one prompt at once, such as times-old-response-checked.

Third, we computed a set of *interviewer-normalized features* based on the prompt- and instance-level feature. The purpose of computing these features was to control for interviewer-level trends in feature values. (For example, one interviewer's phone might have a larger screen size than another's, and therefore what might be a large amount of scrolling for one interviewer might not be for another.) For each interviewer i, numeric feature j, and instance x completed by i, we computed (1) the difference and (2) the squared difference, in standard deviations, between the value of feature j for instance x and the mean value of feature j over all of interviewer i's instances.

We chose some of these features based on prior work from survey methodologists that proposed characteristics that might be exhibited by fabricated data. For example, the features measuring timing were inspired by the idea that curbstoners might complete interviews more quickly [8, 29, 31, 32], and the features num-conditional and total-time-conditional were inspired by the idea that curbstoners might choose short paths through the survey [6, 24]. Other features were not inspired by prior work, but seemed potentially useful. Our philosophy was to be inclusive about the features we extracted and then to rely on the feature selection built into our learning algorithm, random forest, to hone in on the most important features.

The survey that we used contained 52 prompts. For each prompt, there were between 1 and 6 prompt-level features,

response: the actual value of the response for the prompt, if it was a question prompt.
ord: for ordinal question prompts, a positive integer indicating the response's position in the order.
time: the total number of milliseconds spent on the prompt.
delay-to-first-edit: the number of milliseconds between when the prompt was first swiped to and when the the first edit was made.
num-contiguous-edits: the number of times the user re-edited the response to a question prompt contiguously without changing prompts in between.
num-non-contiguous-edits: the number of times the user edited the response to a question prompt immediately after swiping to the prompt.

Table 2. A selection of prompt-level features extracted from the logs.

total-time: the value of the last timestamp seen in the logs for the instance minus the value of the first timestamp seen in the logs for the instance. Because the set of possible timestamps includes timestamps from re-editing sessions, the value of this feature could be much higher than the actual time spent actively editing the form.
(**total, median, min)-answer-time**: the sum (median, minimum) of the time features for all question and note prompts.
(**total, median, min)-delay-to-first-edit**: the sum (median, minimum) of delay-to-first-edit features for all question prompts that were edited at least once.
mean-string-length: the mean length of all non-empty free text responses.
note-time: the total time in milliseconds spent on note prompts.
num-conditional: how many of the conditionally-appearing question prompts were answered.
total-time-conditional: the total time spent on conditionally-appearing question prompts.
(**mean, max)-select-one-contiguous-edits**: the mean (max) of the num-contiguous-edits feature values for the categorical question prompts that were edited at least once.
(**mean, max)-select-one-non-contiguous-edits**: the mean (max) of the num-non-contiguous-edits feature values for the categorical question prompts that were edited at least once.
num-swipes: the total number of forward and backward swipes on the form.
num-previous: the number of backward swipes on the form.
total-scrolled: the total number of pixels scrolled up and down while the form was being filled out.
times-old-response-checked: the number of times the following event happened: when filling out the response to a question, the user moved backward some number of questions and then forward to the original question, without changing any response along the way.

Table 3. A selection of instance-level features extracted from the logs.

giving a total of 209 prompt-level features for each instance. There were also a total of 22 instance-level features and 410 interviewer-normalized features. Thus, for each instance of the survey, we extracted a total of 641 features.[1]

EXPERIMENT DESIGN AND DATA COLLECTION

In this section, we describe how we obtained labeled data. First, we describe the survey itself; second, we detail our experiment protocol; third, we summarize the data that we collected; and fourth, we provide an empirical validation that part of our protocol worked as designed.

The survey

We created a survey specifically for our study. We called the survey the *Study Habits* survey. It contained 44 questions and was designed to take between 5 and 10 minutes to complete. To be eligible to take it, a respondent had to be a university student and be between ages 18 and 25.

The survey started by asking for details about the respondent's major. It then asked for details about the hardest class she had taken in the last year, including its name, when she took it, whether she liked it, and how many hours a week she spent on it. Next, it asked how much time she spent on obligations outside of school, including paid work, volunteer work, research, and family. Following that, it asked whether and how much she would be willing to pay for a major she was interested in if it cost extra. Then it asked a series of questions to determine how often she sought help in her studies from faculty, TAs, tutors, and advisors. It concluded by asking whether she believed that she studied more than the average student, procrastinated more than the average student, and had more obligations outside of school than the average student. (The full survey can be found in Appendix A of [3].)

We considered several factors in the design of the *Study Habits* survey. First, we chose our questions in consultation with the undergraduate advisors in our department and communicated our collaboration verbally and in writing to the interviewers and respondents. We made this choice to encourage interviewers and respondents to take the survey as seriously as they would take most surveys.

Second, we had the questions require a variety of answer types—including 27 categorical questions, 7 numeric questions, and 10 free text questions. We also included a mixture of questions requiring little thought, like "What is your sex?" and questions requiring more thought, like "If tuition was more expensive for certain majors, would that stop you from pursuing a more costly major?" We made these choices to make our survey representative of many real surveys.

Third, we incorporated a rich branching logic. Seventeen of the questions were conditional; that is, they were asked only if an earlier question was answered in a particular way. (For example, respondents were asked what their major was only if they had indicated earlier that they had decided on a major.) We made this choice because we believed that curbstoners would choose answers so that fewer of the conditional questions appeared [6, 24].

[1]Not surprisingly, these features were far from independent. According to a Principal Components Analysis (PCA), 99% of the variation of our data was contained in a space having dimension 305.

Label	Description	Explanation	When Collected	Data Sets
`real`	*real*	Real interviews between participants and respondents, including the interviews that occurred at the end of the training session.	interview period	$\mathcal{D}_0, \mathcal{D}_1, \mathcal{D}_2$
$\texttt{fake}_0$	*uninformed fake*	Fabricated interviews from the training session. Participants told only to "pretend they were interviewing five different people, and answer as they would."	training session	$\mathcal{D}_0$
$\texttt{fake}_1$	*informed fake*	Fabricated interviews from the first round of the follow up session. Participants knew the true purpose of the study and were given a monetary incentive to fabricate data realistically.	follow up session	$\mathcal{D}_1$
$\texttt{fake}_2$	*better-informed fake*	Fabricated interviews from the second round of the follow up session. Participants knew the true purpose of the study, received an incentive to fabricate data well, and were given feedback on features used to identify their fabricated data.	follow up session	$\mathcal{D}_2$

Table 4. Summary of data labels.

Experiment protocol

In realistic scenarios, interviewers who fabricate data have an incentive not to get caught and might learn about curbstoning detection methods that are used. We designed our validation experiment to simulate these conditions.

Specifically, we created three different settings under which interviewers fabricated data. In the first setting, interviewers were not told anything about the purpose of our experiment, and they were given no incentive to fabricate data realistically—instead, they were told only to "pretend they were interviewing someone and answer as he or she would." In the second setting, they were informed that the purpose of our study was to test an algorithm to detect curbstoning, and they were given a monetary incentive to fabricate data realistically; however, they were given no specific information about how our algorithm worked. In the third setting—in addition to knowing the purpose of the study and being given an incentive to fabricate data realistically—they were also given personalized feedback about how the algorithm was identifying their fabricated data.

The four different types of data that we collected (one real, and three fabricated) are summarized in Table 4. Each interviewer's participation lasted approximately one week and consisted of three phases: (1) the *training session*, which took place in the lab; (2) the *interview period*, which took place on the participant's own time; and (3) the *follow up session*, which took place in the lab. We describe each of these phases in more detail below.

Training session

Interviewers began their participation in our study by coming to a training session that lasted between one and two hours. When they came, we told them that our research group was running a survey about study habits on behalf of the undergraduate advisors in our department. They would be interviewers for this survey, and for the week between the training and follow up sessions, they were to administer the survey using ODK Collect on any 10 eligible participants who they could find. We told them that in addition to being interested in the results of the survey, we were also testing some new (unspecified) features of ODK Collect involving data quality. We did not give them any additional information about the purpose of our study.

After demonstrating to the participants how to use the software and conduct the interview, we told them that we wanted them to practice on their own phones by "pretending that they were interviewing five different people and answering as they would." The data collected during this portion of the training session was the first fabricated data created by the participants. We refer to this data, labeled $\texttt{fake}_0$, as the *uninformed fake* data.

After the participants finished generating the $\texttt{fake}_0$ data, we instructed them to take turns interviewing each other. They were to do everything that they would in a normal survey, including obtaining consent. This data was the first data that we gave the label `real`. (Most of the `real` data came from the interview period described below.) At the end of the training session, we gave participants a $20 gift card as compensation for their time.

Interview period

Following the training session, we asked participants to administer the survey to 10 eligible respondents over the course of a week. (There was, however, no enforcement of this request, and compensation was not dependent on the number of respondents interviewed.) Participants conducted the surveys whenever they wanted, using their own Android phone or a loaner Android phone if they did not have their own. The data collected during this period was given the label `real`.

Because the participants were not supervised during this period, we took two measures to ensure that the data that they were collecting was actually real. First, we performed reinterviews on a random sample of respondents (and informed the interviewers that we would do this). Second, at the end of the follow up session, which we describe below, we asked if there were any data quality issues that we should know about. We were careful to convey that we would value any revelations about fabricated data or other deviations from the protocol and that such revelations would not affect compensation. As we will discuss below, interviewers gave us some valuable information during this debriefing.

Follow up session

After the interview period, we asked participants to come back to the lab for a two-hour follow up session, which consisted of two rounds.

In the first round, we began by informing participants of the true purpose of our study, telling them that we were "designing algorithms to automatically detect fabricated survey data" and "testing how well our algorithms predict which data is fabricated." Now that they knew the true purpose of the study,

Our algorithm logs how many seconds you spend on the screens in which you are supposed to read something to the respondent. For each submission you gave us, it found the total number of seconds you spent on these screens. This number tended to be higher in your fabricated data than in the real data.

Our algorithm logs whenever you move backwards through the form (by swiping to the right). For each submission you gave us, it computed the total number of times you moved backwards. This number tended to be lower in your fabricated data than in the real data.

Our algorithm logs the total amount you scrolled up or down on a screen (in pixels) on each submission you gave us. This number tended to be lower in your fabricated data than in the real data.

Figure 2. Example feedback email sent to interviewer.

they were to spend the next 40 minutes fabricating between 4 and 10 more forms.[2] To incentivize them to fabricate data realistically, we said,

> [We] will be giving you feedback at the end of the 40 minutes on how well you fooled the algorithm. Each submission that you give [us] will be given a score by the algorithm, where a low score indicates that it could not tell your data was fabricated—or, in other words, that you did a good job fabricating data.

They were to think of this as a "friendly challenge" to see who could fabricate the data the best, and the person who got the lowest score would get an extra $10 gift certificate. We called the data that we collected during this round the *informed fake* data and gave it the label $\texttt{fake}_1$.

After this data was collected and uploaded, we ran a script that extracted 9 features for each form, called the *feedback-eligible features* (max-select-one-edits, mean-select-one-edits, mean-string-length, median-answer-time, note-time, num-conditional, num-previous, num-swipes, and total-scrolled). We chose these features because—based on previous research and our own intuition—we thought that they would be both predictive of fabrication and easy to explain to participants. After the script extracted the feedback-eligible features, it trained a simple classifier on these features and used its output to generate the fabrication scores. We read the scores for each participant aloud to the group and immediately gave the $10 gift card to the winning participant.

In the second round, we said to the interviewers,

> Now [we're] going to give each of you some personalized feedback on how you did. The algorithm looks at various measures of how you entered the data for each form. High or low values of these measures may indicate that your data is likely to be fabricated. For each of you, [we're] going to send you an email that gives the top three measures that were used to predict which of your forms were fabricated. You can think of these as the clues that the algorithm used to figure out which forms you were fabricating.

For each participant, we ran a script that calculated the three feedback-eligible features that were most correlated with his or her own data being fake, along with the direction of that correlation. Then, we generated emails that explained these correlations and sent them to the participants. An example of an email sent during one of the follow up sessions is shown in Figure 2.

After they received these emails, we gave them another 40 minutes to fabricate between 4 and 10 surveys. We said, "When you fabricate these surveys, keep in mind the report you [got]. This may help you to better fool the algorithm." As in the first round, we gave them a reward of $10 if they got the best score. We called the data collected during this second round the *better-informed fake* data and gave it the label $\texttt{fake}_2$. At the end of the follow up session, we gave participants a $30 gift card—in addition to any of the additional award money received—as compensation for their time.

Data collection

We collected data for our validation experiment between April 12, 2012 and May 14, 2012. Twenty-eight interviewers participated (18 male, 10 female), collectively interviewing 256 respondents and generating 448 fabricated forms. Interviewers were between the ages of 18 and 27 and were undergraduate or Masters students at the University of Washington. Most of them had an engineering or technical major.

All of the 28 participants who attended a training session also came to a follow up session. There were 11 training sessions and 11 follow up sessions, which varied in size between one and five participants.[3] Although each interviewer's involvement with the study was intended to last exactly one week, because of last-minute scheduling changes, this period actually varied between 5 and 9 days.

Our data quality measures gave us confidence that most of the data labeled `real` was actually real. For 26 of the 28 interviewers, the first random respondent we chose to reinterview verified that the interview actually occurred. (When this happened, we did not reinterview any other respondents.) The other two had only one respondent leaving contact information, who could not be reached.

[2] We gave this flexibility to participants to avoid artificially constraining the amount of time that they had to fabricate data. Given our instructions, the particpants had between 4 and 10 minutes to fabricate each form. We chose this range of times based on how long it took participants to collect their first `real` data in the training sessions. Of the 24 `real` forms filled out in the training sessions, the completion times of 22 fell within this range. Additionally, the instructions given to respondents included explicit wording to keep them from rushing: "There is not necessarily an advantage to creating a lot of fake data. You may—but do not have to—use the entire 40 minutes." Regardless, in most of the follow up sessions, respondents finished fabricating their data before the 40 minutes had elapsed.

[3] If there was only one participant in the follow up session, we awarded the $10 gift certificate to the participant if he or she beat a predetermined cutoff score.

Label	Training	Test
real	203 (38%)	53 (33%)
$\texttt{fake}_0$	98 (18%)	26 (16%)
$\texttt{fake}_1$	134 (25%)	49 (30%)
$\texttt{fake}_2$	106 (20%)	35 (21%)

Table 5. Distribution of labels for training and test sets. There were 541 instances in the training set and 163 instances in the test set.

During the debriefing at the end of the follow up session, we learned some important information from interviewers. Three of them told us that for a total of 8 forms, they collected data on paper during the interview and entered it into the phone later; three of them told us that for a total of 5 forms, they handed the phone to the respondent to self-administer; and finally, one of them told us that he fabricated 2 forms in order to reach the goal of 10 interviews. Although these types of deviations from protocol could occur in a real survey, it was unclear what label should be given to these forms. Depending on the survey, having a respondent self-administer the survey may or may not be acceptable, and the fabricated data from the interview period that we discovered does not fit cleanly into one of our three labels of fabricated data. Therefore, to ensure that our data accurately reflected the study protocol, we removed these 15 forms.

As we collected the data, we divided it into a training and test set. To ensure that each interviewer's forms were divided into the training and test sets in a similar way, we performed the following procedure. For each interviewer i and for each label ℓ in $\{\texttt{real}, \texttt{fake}_0, \texttt{fake}_1, \texttt{fake}_2\}$, we looked at all of the forms completed by interviewer i having label ℓ. If the number of these forms was between 0 and 2, we did not remove any of the forms for the test set. If it was between 3 and 5, we randomly chose 1 of the forms to be in test set. If it was between 6 and 10, we randomly chose 2 of the forms to be in test set. Otherwise, if it was some number $r > 10$, we randomly chose $\lceil r/5 \rceil$ forms to be in the test set.

This procedure created a training set of size 541 and a test set of size 163. The distribution of the labels in each of these sets is shown in Table 5. To avoid over-fitting, we did all of our exploratory data analysis on the training set (using cross-validation when necessary).

Validation of study protocol

The purpose of giving feedback between the two rounds of the follow up session was to simulate what happens as interviewers learn and adapt to an algorithm to detect curbstoning. However, just because the interviewers received personalized feedback does not mean that they necessarily understood it or reacted to it. Before we tested how well a classification algorithm could identify fabricated data, we validated that interviewers did indeed react to the feedback that they were given.

Recall that there were a set of nine *feedback-eligible* features that were used by the algorithm in the follow up session and that the interviewers were sent an email that told them which three of these features were most correlated with fabrication, along with the direction of the correlation. There were $28 \times 3 = 84$ interviewer–feature pairs (i, j) in which interviewer i received feedback about feature j. If the interviewers understood and reacted to this feedback, then for each such pair (i, j), the mean value of feature j for interviewer i should change in the *opposite* direction of the initial correlation between the two rounds of the follow up session. The null hypothesis—that interviewers either did not understand or did not react to the feedback—would imply that the average value of feature j would be just as likely to go up as it would be to go down, regardless of whether the feature was negatively or positively correlated with fabrication.

We found that 71 times out of 84, the mean value of j changed in the opposite direction of the initial correlation. Thus, the null hypothesis could be rejected ($p < 0.001$), suggesting that, in aggregate, interviewers did react to the feedback in a manner suggesting that they understood it.

The magnitude of the changes of the features between the two rounds is important as well. For a graphical exploration of the magnitude of the changes, we refer the interested reader to Figure 6.1 in [3]. This figure, which we omitted to save space, provides further support for the thesis that interviewers understood and reacted to the feedback that they were given.

RESULTS

In this section, we describe the results of running a classifier on the data we collected: we summarize the overall performance, investigate how this performance depends on label noise and training set size, and discuss limitations.

To evaluate a classifier separately on each of the three types of fake data, we created three data sets: $\mathcal{D}_0$, which consisted of the real data and the $\texttt{fake}_0$ data; $\mathcal{D}_1$, which consisted of the real data and the $\texttt{fake}_1$ data; and $\mathcal{D}_2$, which consisted of the real data and the $\texttt{fake}_2$ data. We divided each of the data sets into a training and test set according to the overall split between training and test described in Table 5.

For classification, we chose the random forest algorithm [7] because it is efficient, is considered to be one of the most accurate off-the-shelf supervised classification algorithms [9], exhibits little sensitivity to parameter-choice, and performed well in our initial experiments on the training data. We used the implementation provided by Weka [19], an open source data mining library.

Overall performance

We trained a random forest classifier on the training set and evaluated it on the test set for each of the data sets.[4] On data set $\mathcal{D}_0$, the accuracy was 96%; on data set $\mathcal{D}_1$, the accuracy

[4]Random forest takes two parameters: I, the number of decision trees; and K, the number of features considered at each split in the tree. Performance increases as I increases; therfore the standard recommendation is to choose as high a value for I as is practical [7]. Based on preliminary experiments on training data, we determined that $I = 200$ was high enough to obtain maximal performance and low enough to be efficient. We chose this value for all of our experiments. To optimize performance, we initially performed a search through values of K (2, 4, 8, 16, 32, 64, 128), using as an optimization criterion the accuracy achieved during 10-fold cross-validation on the training set (averaged over 10 repetitions). Once we did this, we saw that performance did not depend strongly on K. Thus, for simplicity, we set $K = 32$ in the remainder of our experiments.

Data	Prec.	Rec.	Spec.	F_1	Acc.
$\mathcal{D}_0$	0.93	0.96	0.96	0.94	**0.96**
$\mathcal{D}_1$	0.88	0.92	0.89	0.90	**0.90**
$\mathcal{D}_2$	0.85	0.80	0.91	0.82	**0.86**

Table 6. Performance of random forest.

Data	Prec.	Rec.	Spec.	F_1	Acc.
$\mathcal{D}_0$	0.75	0.46	0.93	0.57	**0.77**
$\mathcal{D}_1$	0.71	0.31	0.89	0.43	**0.61**
$\mathcal{D}_2$	0.90	0.26	0.98	0.40	**0.69**

Table 7. Performance of random forest without behavioral data.

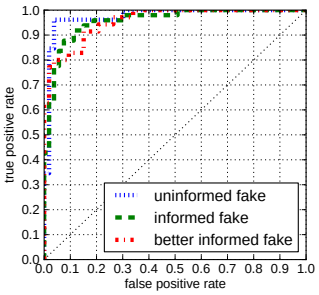

Figure 3. ROC curves of random forest.

was 90%; and on data set $\mathcal{D}_2$, the accuracy was 86%.[5] More detailed performance measures—including the precision, recall, specificity, and F_1-score—are shown in Table 6. ROC curves of the classifiers' performance are shown in Figure 3.

Next, to see how much the behavioral data helped, we removed all of the behavioral features. That is, we kept only the response features described in Table 2 and the interviewer-normalized features derived from them (which resulted in 47 features total). We re-trained a random forest classifer on the training data, and re-tested it on the test data. The accuracies achieved were 77% for $\mathcal{D}_0$, 61% for $\mathcal{D}_1$, and 69% for $\mathcal{D}_2$. More detailed performance measures are shown in Table 7.

Discussion

Our results in Table 6 show that our approach can accurately identify curbstoning and that the accuracy decreases as interviewers learn about the algorithms and are given an incentive to fabricate data realistically. However, even for data set $\mathcal{D}_2$, when interviewers had specific knowledge about how the algorithm worked and were incentivized to avoid detection, the accuracy was still 86%.

These high accuracies can be contrasted with the much lower accuracies shown in Table 7, demonstrating how important it was to use the behavioral data. In fact, the performance without the behavioral data was not much above what would be achieved by a baseline classifier that always output `real`—especially for the data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, when interviewers knew the algorithm was being used and had an incentive to fabricate data realistically.

[5]One might wonder whether these numbers depended on the fact that every interviewer was represented equally in the training and test sets. To validate that our results did not depend too strongly on how we chose test data, we performed an additional 10-fold cross-validation of random forest on the entire data set, in which we divided the data without considering the interviewer. In this analysis, random forest had an accuracy of 0.93, 0.91, and 0.89 on the uninformed, informed, and better-informed fake data, respectively.

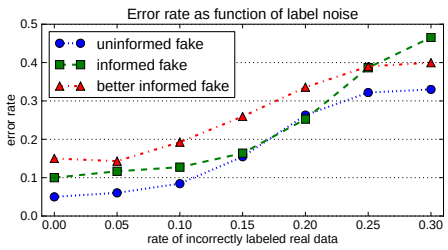

Figure 4. Effect of label noise.

The most important behavioral features were the timing-based ones, in particular the delay-to-first-edit features. As might be expected, fabricated forms tended to be completed faster. Perhaps more surprisngly, even though 22 out of the 28 participants were told that they were fabricating data too quickly in the first round of the follow up session, timing remained a strong predictor of fabrication in the second round. For example, the interviewer-normalized version of the feature delay-to-first-edit had a significant ($p < 0.001$) negative correlation with the $\texttt{fake}_2$ data ($r = -0.56$). Timing was not the only predictive feature, however. Other features that were significantly ($p < 0.001$) correlated with fabrication included the num-edits features (positively correlated), num-swipes (negatively correlated), and total-scrolled (negatively correlated).

In Chapter 6 of [3], we performed a systematic analysis that listed the features most correlated with fabrication. From this analysis, we made a few general conclusions, including (1) as interviewers gain experience and motivation to fabricate data well, behavioral data becomes a more important indicator of fabrication, relative to response data; (2) to predict fabrication, it is better to use aggregated form-level features than individual prompt-level features; and (3) timing information from hard questions predicts fabrication more accurately than timing information from easy ones. (For example, the time taken to answer the question "If tuition was more expensive for certain majors, would that stop you from pursuing a more costly major?" was a better predictor of fabrication than the time taken to answer the question "What is your sex?".)

Robustness to label noise

If a survey organization were to apply our approach to detect curbstoning, they would need labeled data on which to train a classifier. It is easy to ensure that the fake data is actually fake because it can be generated in a lab. It is less easy, however, to ensure that real data is actually real because it is typically gathered in less-supervised scenarios. Therefore, one might expect some inaccuracy in the data labeled `real`.

We ran an experiment to simulate the effect of label noise in the `real` data. For each data set $\mathcal{D}_0$, $\mathcal{D}_1$, and $\mathcal{D}_2$, we randomly chose a subset of the fake data in the training set and switched its label to `real`. Then we trained the random forest algorithm on this partially mislabeled training data. For each value of $\rho \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$, we switched

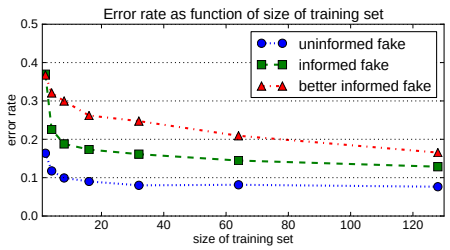

Figure 5. Effect of training set size.

the label on the amount of fake data needed to make ρ fraction of the real data mislabeled. We repeated this 50 times for each value of ρ and plotted the average error rate of the resulting classifier on the test set in Figure 4.

Discussion

Figure 4 shows that, not surprisingly, as the noise in the labeled data increases, the accuracy of the resulting classifier decreases. Once 30% of the `real` data is mislabeled, the accuracy is barely above what one would get with the baseline classifier that always outputs `real`. However, for a more moderate level of noise, say 10%, the accuracy of all three classifiers was still more than 80%. This result suggests that our approach is robust to a moderate level of label noise.

Amount of training data required

Another practical concern facing a survey organization implementing our approach is the question of how much training data to use. To help answer this question, we ran an experiment to test how the accuracy varied as the amount of training data varied. For each of the three data sets $\mathcal{D}_0$, $\mathcal{D}_1$, and $\mathcal{D}_2$, and for each number $m \in \{2, 4, 8, 16, 32, 64, 128\}$, we created a random subset of the training set consisting of $m/2$ real forms and $m/2$ fake forms. Then we trained a random forest classifier on this data and measured its accuracy on the test set. We repeated this procedure 50 times for each value of m and plotted the average error rate in Figure 5.

Discussion

Figure 5 shows that the amount of training data required to achieve performance close to what we have reported depends on the type of fake data. Therefore, the amount of training data that a survey organization collects should depend on the anticipated sophistication of curbstoners. If it is mostly interested in catching careless curbstoning, like the $\texttt{fake}_0$ data, then it may need only a few dozen labeled instances. On the other hand, if it is interested in catching more sophisticated curbstoning, like the $\texttt{fake}_1$ or $\texttt{fake}_2$ data, then it may need closer to 100 labeled instances.

Limitations

There are at least two limitations of our study. The first is that the participants were relatively inexperienced as interviewers. When they fabricated the $\texttt{fake}_0$ data, they had not yet conducted any real interviews, and when they fabricated the $\texttt{fake}_1$ and $\texttt{fake}_2$ data, they had conducted approximately 10 interviews. Consequently, they may not have been as good at fabricating data as more experienced interviewers, and this inexperience may have caused our results to be overly optimistic.[6] Further research is needed to determine the long-term effectiveness of our approach.

The second limitation is that, in our study, if any data in a survey was fabricated, then the entire survey was fabricated. In real surveys, interviewers might fabricate only certain questions of a survey, which would be harder to detect. Further research is also needed to evaluate the effectiveness of our approach in detecting this type of fabrication.

CONCLUSION

In this paper, we described a new approach to detecting interviewer fabrication in surveys: instrumenting electronic data collection software to record logs of low-level behavioral data and applying supervised classification on features extracted from these logs. We showed that this approach could identify fabrication with an accuracy of up to 96% and that even when interviewers had specific knowledge of how the algorithm worked and were incentivized to avoid detection, it could still achieve an accuracy of 86%. We also demonstrated the robustness of our approach to a moderate amount of label noise and provided practical recommendations, based on empirical evidence, on how much training data is needed.

One might wonder whether the benefit of our approach outweighs the cost. Further research is needed to answer this question definitively, but we argue briefly here that our approach may be economically justified. Managers of large surveys, such as the Demographic and Health Surveys (conducted over 260 times in 90 countries over the last 30 years) already spend their time and budget on ensuring data quality using costly techniques such as re-interviews. For a small amount of initial overhead, they could obtain labeled data as we outlined—our results show that there does not have to be a lot. Then they could use pre-built software to record behavioral data, build a classifier, and output predictions that would dramatically improve their ability to identify low-quality data.

Even in less elaborate surveys, interviewers still train for hours, some of which are spent fabricating surveys to gain familiarity with the questionnaire and software. In a deployment using our approach, these fabricated forms could be used as the labeled fake data. The labeled real data could be obtained from surveys that are known to be high quality, such as ones from trusted interviewers or verified by re-interviews. These labels would not be perfect, but our approach is robust to moderate amounts of label noise. Thus, the overhead needed to apply our techniques may actually be quite small.

There are several directions for future research; we describe three here. First, as mentioned earlier, it is important to study how well our approach works on larger surveys with more experienced interviewers and partial fabrication of surveys. Second, the approach that we have proposed is black-box in

[6]There is, however, at least one argument to suggest that this effect may be small: the fact that random forest did so poorly without the behavioral data for $\mathcal{D}_1$ and $\mathcal{D}_2$ (Table 7) suggests that, at least by the follow up session, our interviewers had enough experience to simulate realistic responses (if not realistic behavior).

the sense that it does not incorporate expert knowledge regarding what tends to characterize fabricated data. However, people familiar with a survey might have important knowledge, such as lower bounds on the time required to answer certain questions. Investigating how this type of knowledge can be incorporated into a curbstoning detection system (both from an interface and algorithmic perspective) is an important direction for future work. Third, it would be interesting to study how data from the many sensors built into smart phones could be leveraged to detect fabrication or other data quality issues. GPS is an obvious candidate (but not a panacea, since interviewers being in the right place at the right time does not alone guarantee data quality). One could also imagine employing other sensors, such as the accelerometer, light sensor, microphone, or camera.

ACKNOWLEDGMENTS

We thank Saleema Amershi, Adrienne Andrew, Michael Buettner, Dan Weld, Luke Zettlemoyer, and the UW Change group for helpful discussions; Nicki Dell, Kayur Patel, and the anonymous reviewers for their insightful comments on the paper; Carl Hartung for writing an early version of the logging code; and Crystal Eney and the other CSE advisors for their help in designing the survey. This research was supported by grants from Google and Yahoo!.

REFERENCES

1. Baker, R. P. New technology in survey research: Computer-assisted personal interviewing (CAPI). *Social Science: Computer Review 10*, 2 (1992), 145–157.
2. Bennett, A. Toward a solution of the "cheater problem" among part-time research investigators. *J. Marketing 2* (1948), 470–474.
3. Birnbaum, B. *Algorithmic Approaches to Detecting Interviewer Fabrication in Surveys*. PhD thesis, U. Washington, 2012.
4. Birnbaum, B., et al. Automated quality control for mobile data collection. In *DEV* (2012), 1:1–1:10.
5. Blaya, J. A., et al. E-health technologies show promise in developing countries. *Health Aff. (Millwood) 29*, 2 (2010), 244–51.
6. Bredl, S., et al. A statistical approach to detect cheating interviewers. Tech. Rep. 39, University Giessen, Center for International Development and Environmental Research (ZEU), 2008.
7. Breiman, L. Random forests. *Machine Learning 45* (2001), 5–32.
8. Bushery, J. M., et al. Using date and time stamps to detect interviewer falsification. *Proc. ASA (Survey Research Methods)* (1999), 316–320.
9. Caruana, R., et al. An empirical evaluation of supervised learning in high dimensions. In *ICML* (2008).
10. Chen, K., et al. USHER: Improving data quality with dynamic forms. *IEEE Trans. Knowledge and Data Engineering 23*, 8 (2010), 1138–1153.
11. Cho, M. J., et al. Inferential methods to identify possible interviewer fraud using leading digit preference patterns and design effect matrices. *Proc. ASA (Survey Research Methods Section)* (2003), 936–941.
12. Couper, M. P. Usability evaluation of computer-assisted survey instruments. *Social Science: Computer Review 18*, 4 (2000), 384–396.
13. Couper, M. P., and Kreuter, F. Using paradata to explore item level response times in surveys. *J. Royal Statistical Society: A* (2012).
14. Crespi, L. P. The cheater problem in polling. *Public Opinion Quarterly 9*, 4 (1945), 431–445.
15. DeRenzi, B., et al. Mobile phone tools for field-based health care workers in low-income countries. *Mt. Sinai J. Medicine 78*, 3 (2011), 406–418.
16. EpiSurveyor. **http://www.episurveyor.org/**.
17. Evans, F. B. On interviewer cheating. *Public Opinion Quarterly 25* (1961), 126–127.
18. Ghazarian, A., and Noorhosseini, S. M. Automatic detection of users skill levels using high-frequency user interface events. *User Modeling and User-Adapted Interaction 20*, 2 (2010), 109–146.
19. Hall, M., et al. The WEKA data mining software: An update. *SIGKDD Explorations 11*, 1 (2009).
20. Hansen, S. E., and Marvin, T. Reporting on item times and keystrokes from Blaise audit trails. Tech. rep., 2001.
21. Hartung, C., et al. Open Data Kit: Tools to build information services for developing regions. In *ICTD* (2010).
22. Hilbert, D. M., and Redmiles, D. F. Extracting usability information from user interface events. *ACM Comp. Surveys 32*, 4 (2000), 384–421.
23. Hong, H. S., et al. Adoption of a PDA-based home hospice care system for cancer patients. *Comput. Inform. Nurs. 27*, 6 (2009), 365–71.
24. Hood, C. C., and Bushery, J. M. Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *Proc. ASA (Survey Research Methods Section)* (1997), 820–824.
25. Hurst, A., et al. Automatically detecting pointing performance. In *IUI* (2008).
26. Inciardi, J. A. Fictitious data in drug abuse research. *Intl. J. Addictions 16* (1981), 377–380.
27. Judge, G., and Schechter, L. Detecting problems in survey data using Benford's Law. *J. Human Resources 44*, 1 (2009), 1–24.
28. Kiecker, P., and Nelson, J. E. Do interviewers follow telephone survey instructions? *J. Market Research Society 38* (1996), 161–176.
29. Krejsa, E. A., et al. Evaluation of the quality assurance falsification interview used in the Census 2000 dress rehearsal. *Proc. ASA (Survey Research Methods Section)* (1999), 635–640.
30. Lal, S. O., et al. Palm computer demonstrates a fast and accurate means of burn data collection. *J. Burn Care Rehabil. 21*, 6 (2000), 559–61.
31. Li, J., et al. Using statistical models for sample design of a reinterview program. *Proc. ASA (Survey Research Methods Section)* (2009), 4681–4695.
32. Murphy, J., et al. A system for detecting interview falsification. In *American Assoc. Public Opinion Research 59th Ann. Conf.* (2004).
33. Parikh, T. S., et al. Mobile phones and paper documents: Evaluating a new approach for capturing microfinance data in rural india. In *CHI* (2006), 551–560.
34. Pendragon Forms. **http://pendragonsoftware.com/**.
35. Porras, J., and English, N. Data-driven approaches to identifying interviewer data falsification: The case of health surveys. *Proc. ASA (Survey Research Methods Section)* (2004), 4223–4228.
36. Ratan, A. L., et al. Managing microfinance with paper, pen and digital slate. In *ICTD* (2010).
37. Rzeszotarski, J. M., and Kittur, A. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *UIST* (2011).
38. Schreiner, I., et al. Interviewer falsification in census bureau surveys. *Proc. ASA (Survey Research Methods Section)* (1988), 491–496.
39. Shäfer, C., et al. Automatic identification of faked and fraudulent interviews in surveys by two different methods. *Proc. ASA (Survey Research Methods Section)* (2004), 4318–4325.
40. Stieger, S., and Reips, U.-D. What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior 26*, 6 (2010), 1488–1495.
41. Turner, C. F., et al. Falsification in epidemiologic surveys: Detection and remediation. Tech. Rep. 53, Research Triangle Institute, 2002.
42. United Nations Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects*, 2011.
43. United Nations Development Programme. *The Human Development Report*, 2011.