Automated Quality Control for Mobile Data Collection

Benjamin Birnbaum
Computer Science Department
University of Washington, Seattle
birnbaum@cs.washington.edu

Brian DeRenzi
Computer Science Department
University of Washington, Seattle
bderenzi@cs.washington.edu

Abraham D. Flaxman
Institute for Health Metrics and Evaluation
University of Washington
abie@uw.edu

Neal Lesh
Dimagi, Inc.
nlesh@dimagi.com

ABSTRACT

Systematic interviewer error is a potential issue in any health survey, and it can be especially pernicious in low- and middle-income countries, where survey teams may face problems of limited supervision, chaotic environments, language barriers, and low literacy. Survey teams in such environments could benefit from software that leverages mobile data collection tools to provide solutions for automated data quality control. As a first step in the creation of such software, we investigate and test several algorithms that find anomalous patterns in data. We validate the algorithms using one labeled data set and two unlabeled data sets from two community outreach programs in East Africa. In the labeled set, some of the data is known to be fabricated and some is believed to be relatively accurate. The unlabeled sets are from actual field operations. We demonstrate the feasibility of tools for automated data quality control by showing that the algorithms detect the fake data in the labeled set with a high sensitivity and specificity, and that they detect compelling anomalies in the unlabeled sets.

1. INTRODUCTION

Countries need accurate population-level health data in order to allocate resources, improve performance, and respond to health emergencies [34]. In low- and middle-income countries, millions of dollars are spent every year to gather health-related data. Surveys and health surveillance efforts involving human interviewers face concerns about low-quality data because of misunderstanding or deliberate falsification. Although little has been published on fabrication rates, in one survey from the United States Census Bureau, at least 6.5% of interviewers were found falsifying data [40]. Because of this, most surveys involve quality control efforts to try to detect and resolve sources of systematic interviewer error [1].

In developing countries, first PDAs, and now mobile phones, have risen in popularity as a platform for data collection and have been shown to improve both efficiency and data quality [45]. Tools such as Open Data Kit (ODK) [20] and EpiSurveyor [17] can immediately upload data to servers. This abillity has enabled improved quality control efforts by teams of people who monitor data [10].

Recent ICTD research has explored approaches for improving data quality in low income countries, such as using call centers for dictation by health workers [36], using probabilistic models of data to prevent and correct data entry errors [7, 8] and creating a hosted paper form digitization service [9]. The objective of our current work is complementary: we seek to create tools to assist data quality control teams to identify suspicious or surprising trends that could be the result of interviewer misunderstanding or falsification. As a first step, we investigate and test algorithms that could form the basis of such tools.

We begin our validation using a unique labeled dataset from a community health organization in Tanzania. In this data set, some of the data is known to be fabricated, and some is believed to be relatively accurate. We explore how well we can detect the fabricated data algorithmically. First, we investigate how well several off-the-shelf supervised machine learning techniques perform. Next, because labeled data sets may not always be available for training, we develop two unsupervised outlier detection techniques and test how well they detect the fake data. We show that both the supervised and unsupervised techniques predict which data is fake with a high sensitivity and specificity.

Because our labeled data sets were obtained under somewhat artificial conditions, we next run our algorithms on two more realistic (but unlabeled) data sets. We show that the unsupervised techniques find compelling anomalies, such as an interviewer who claimed that 89% of her respondents wanted to know more about family planning, even though the overall reported percentage for all interviews by all interviewers was only 29%, and an interviewer who claimed that 91% of her respondents were not home, even though the overall percentage was only 28%. (Figure 5 provides more examples.) We conclude with a brief discussion of how these algorithms could be integrated into tools for data quality control.

To summarize, our main contributions are threefold:

- We develop unsupervised outlier detection techniques for finding anomalous data that, to our knowledge, are

DEV '12, March 11-12, Atlanta, GA

novel.

- We quantitatively validate both the off-the-shelf supervised techniques and the novel unsupervised techniques, by showing that they successfully predict which data is fabricated in the labeled set.

- We informally validate the unsupervised techniques by showing that they find compelling anomalies in realistic but unlabeled data sets.

Although we are interested in systematic interviewer-level errors from any cause, the low-quality data in our labeled data arises specifically out of interviewer fabrication. Nevertheless, we believe our techniques also apply to low-quality data from other causes, such as interviewer misunderstanding.

1.1 Motivation

Our work is motivated by a broad range of data collection situations, but for concreteness we anchor our discussion on data management technologies for community health worker (CHW) programs. In these programs, trained community members who are not healthcare professionals travel between homes in a given area to check up on clients, perform routine care, promote healthy behavior, and make referrals to health facilities when necessary. CHW programs can positively affect health outcomes, but they require strong supervision and support to be effective [30]. To address some of these needs, tools have been developed that run on mobile phones carried by CHWs [14, 16, 27, 32, 46]. For example, CommCare [14, 32] is an open-source software platform for CHWs that is used to register new clients, structure interviews, and record responses directly into the phone. The data received by the application is uploaded to a central server called CommCareHQ, which tracks cases and generates reports for supervisors and funders.

CommCare removes the need for CHWs to manually enter their data into a database after recording it on paper, improving efficiency [32]. The immediate digitization of the data provides other advantages as well. It has the potential to make the data quickly available to doctors and epidemiologists for health surveillance. It also makes it easier for supervisors to manage their operations effectively. For example, a supervisor can use CommCareHQ to monitor the number of visits that CHWs are making to ensure that program objectives are being met. Or, automated SMS reminder systems built on top of the case management logic of CommCareHQ can be used to encourage CHWs to follow up with clients when needed. In fact, it was recently shown that one such system decreased the number of days a CHW's client were overdue for a follow up visit by 86% [15].

However, the data management capabilities of CommCare are only useful if it is managing high-quality data — and this might not always be the case. At one CommCare deployment, at an organization we will call "Organization A," surprising patterns in the data were discovered. Further investigation revealed that a significant number of CHWs were entering fabricated home visits into their phones. At a different program, at an organization we will call "Organization B," we found that supervisors were performing extensive data quality control measures that exhibited the degree to which they were concerned about the issue. Survey methodologists use the term "curbstoning" to describe interviewer data fabrication, and as discussed in the related work section, the phenomenon has been a concern of these researchers for over 60 years.

Complete fabrication of data is just an extreme case of the type of data quality problems that can arise in low-resource settings. In our experience, most data collectors are conscientious and hardworking. However, a field worker acting in good faith may easily introduce bad data into the pipeline because of a misunderstanding or miscommunication. This may happen especially easily in the developing world because of a pervasive lack of expertise, low training, high turnover, mismatched incentives, and cultural differences regarding data [9].

Low-quality data in a system like CommCare negates the intended benefit of providing decision makers a sound basis for their decisions. Even more importantly for the community health setting, it hinders the ability of supervisors to manage their operations effectively, which can directly affect the health of the client population. Supervisors already spend a portion of their limited time to address data quality issues; the development of automated tools to assist them in their efforts is of primary importance.

1.2 Organization

We begin in Section 2 by describing related work. In Section 3, we describe the data sets that we use to validate our algorithms. We briefly define terminology in Section 4. In Section 5, we describe the supervised machine learning techniques and the results of our validation of these techniques. We do the same for the unsupervised techniques in Section 6. In Section 7, we describe the results of running the unsupervised technique on the unlabeled data sets. We conclude in Section 8 by briefly discussing how these algorithms could be integrated into useful tools.

2. RELATED WORK

Data quality has motivated a number of ICTD interventions (see, for example, [9, 35, 36, 38]). Perhaps most relevant to our work is Usher [7, 8], a software program that improves data quality at the form level by creating a probabilistic model of the form data and using this to guide data entry clerks away from making mistakes. This work is similar to ours in its use of machine learning to improve data quality. Although its use case is predicated on the assumption that users are cooperative and not deliberately falsifying data, the probabilistic model that it builds could be extended to detect fabrication. We note that our unsupervised techniques are different from Usher's supervised approach.

Standard data cleaning practice involves finding and removing numeric outliers [21]. Our novel unsupervised techniques can be considered an extension of this idea to outlying *distributions* of categorical data. Of course, standard numeric outlier detection algorithms should also be included in a working system for monitoring data quality.

Another relevant body of work comes from researchers in survey methodology. Traditional methods for improving survey quality emphasize a multi-level process-oriented approach [1]. This line of work draws on both the statistical and psychological literature to understand how errors are introduced and what can be done to prevent them. Error prevention techniques at the interviewer level include matching interviewers to respondents by socioeconomic characteristics, recruiting only experienced interviewers, carefully

training interviewers in the concepts and objectives of the survey, and re-interviewing a small random sample of respondents to detect anomalies [1]. All of these techniques may be difficult in resource-constrained settings.

As mentioned in the introduction, survey methodologists use the term "curbstoning" to describe data falsification by interviewers. This problem first received attention in a 1945 paper by Crespi [12], who gave several reasons that interviewers might fabricate data, including long questionnaires, complex or intrusive questions, unreasonable expectations, and hard to reach subjects. All of these are potential problems faced by fieldworkers in developing regions that should be considered when programs are designed. We see our work as complementary to a careful consideration of these important human factors.

More recently, researchers with the United States Census Bureau have published a series of papers on curbstoning by census interviewers [2, 5, 23, 28, 31, 40]. Techniques to detect curbstoning are varied. Some researchers have explored the use of timing information recorded by data entry programs used during the survey [5, 31, 33]. For example, Li et al. show that falsified interviews in the Current Population Survey in the United States Census are more likely to be faster than average and to occur shortly before deadlines [31]. The ability of tools such as ODK and CommCare to immediately upload metadata such as timing information provides an opportunity to leverage this observation. Indeed, in Section 5, we show that timing data can help us detect fabrication in the community health setting.

Other papers have looked at anomalous distributions for particular questions that might indicate someone minimizing effort or the ability of supervisors to monitor performance, such as an unusual amount of unavailable respondents or missing telephone numbers that would prevent supervisor followups [23, 28, 33, 44]. Some researchers hypothesize that interviewers who falsify data may choose less extreme values and have answer distributions with lower variance [37, 42]. A large body of work examines the potential of Benford's Law [22] to detect low-quality or falsified data [3, 26, 39, 42, 43]. This technique does not extend to non-numeric data.

Similar to our work, some survey methodologists have explored the use of machine learning methods, such as logistic regression, to detect data that is likely to have been faked [31]. This work uses models trained on metadata, such as timing information, whereas our work uses the data itself in addition to metadata.

To summarize, our research is based on a number of techniques that have appeared in related contexts. Our contribution is to bring these techniques together and systematically explore them in the context of mobile tools for data collection in developing regions. We join others [9] in arguing that the results presented here and related work [8] can form the basis of an important research agenda: to develop algorithms and tools to ensure that the valuable data from the developing world is of high quality.

3. DATA SETS

We use three data sets for our validation. The first is a labeled set from a community health program in Tanzania ("Organization A"), in which some CHWs were known to be fabricating data and some CHWs were believed to be obtaining accurate data. This data set is used in Sections 5 and 6. The second is a large unlabeled set, also from Organization A, collected from actual field operations. This data set is used in Section 7. The third is from a community outreach program in Uganda ("Organization B"). This data set is also used in Section 7. We describe each in turn.

Organization A (Labeled)

This data is from a CommCare-supported CHW program in Tanzania that specialized in maternal and neonatal health. The workers travelled between approximately 100 houses in the period of a month, asking basic health and demographic questions, such as whether anyone had a cough, what kind of family planning methods were being used, and whether there were any children under 2 years old. If a CHW found an urgent health problem, she would refer the household to a clinic.

The data consists of 846 forms filled out by 40 CHWs. There were 12 questions in total, of which one was numerical ("How many people are in your household?") and 11 were categorical. CommCare recorded the time it took CHWs to fill out each form, but it did not record per-question timing information. Of the 846 rows in our data set, 317 were labeled *real* (from 15 unique CHWs) and 529 were labeled *fake* (from 25 unique CHWs).

The *real* part of our labeled data set came from a study measuring how often different interviewers reported the same answers on surveys of the same household [13]. Unlike normal day-to-day operations at Organization A, these interviews were conducted in a central location, to which mothers from households in the area would travel. The interviews were conducted on seven different days in November and December of 2010. Because of the unusually high amount of supervision during these interviews, we have high confidence that the interviewers were faithfully interviewing the clients and recording their responses accurately.

The *fake* part of our labeled data set was obtained from a "fake data party" that we set up for the purpose of this study. Twenty-five CHWs from Organization A were gathered and given a small payment to fill out several copies of the same form used for the real data. They did not actually interview any households while filling out these forms. Instead, to approximate the experience of a CHW who was trying to fabricate data realistically enough not to get caught, they were instructed in Swahili to "pretend they were a client and answer as a client would." Over two and a half hours, these 25 CHWs faked a total of 529 forms. A brief manual inspection of this data revealed no obvious misunderstandings and no obvious shortcuts taken by the CHWs to fill out the forms.

We emphasize that the labels on this data set are not perfect. We cannot be certain that every form in the real data set is accurate, and a CHW who is fabricating data in a fake data party might do so differently than a CHW who is fabricating data in the field. We believe, however, that the labels are accurate enough and the data realistic enough to provide a meaningful test for our algorithms.

Organization A (Unlabeled)

In addition to the labeled data set from Organization A, we also use an unlabeled data set from the same program. We believe that this data set is more representative than the labeled data set because it comes from actual CHW field

operations during June and July of 2011. The form used is similar to the form in the labeled set. It has 15 questions, all of which are categorical. We filtered the data to include rows only from CHWs who submitted at least 20 forms during the two-month period. This resulted in a set of 4321 forms submitted by 33 unique CHWs.

Organization B (Unlabeled)

This data set comes from the field operations of a community outreach program in Uganda. A number of questions in this data set were "check all that apply." To transform these into a multiple choice format, we created a separate yes/no question for each possibility. This resulted in 103 different questions. The raw data set consists of 328 rows submitted by 42 interviewers. We filtered this to include rows only from interviewers who submitted at least 8 forms, which resulted in a final set of 265 forms submitted by 26 interviewers.

4. TERMINOLOGY

In Sections 5 and 6, we evaluate the ability of several algorithms to predict whether survey data is real or fake. The unit of prediction varies: in some cases predictions are made on individual forms (form-level); in some cases predictions are made on all of the forms filled out by a particular CHW (interviewer-level); and in some cases predictions are made on all of the forms from a particular CHW, but only for one question at a time (question-level). For each case, a *positive* instance is one that contains fake data. For a given algorithm, the *true positive rate* is the proportion of fake instances that are correctly labeled as such. The *false positive rate* is the proportion of real instances that are incorrectly labeled fake. The *sensitivity* is the true positive rate, and the *specificity* is one minus the false positive rate.

All of the classifiers we evaluate output a continuous-valued score for each instance, where a higher score indicates a stronger belief that the instance is fake. To transform this score into a binary classification (either real or fake), a cutoff score must be chosen: all instances with a score higher than the cutoff are predicted to be fake, and all instances with a score lower than the cutoff are predicted to be real. The choice of cutoff presents a tradeoff. A high cutoff will result in fewer false positives (higher specificity), but also fewer true positives (lower sensitivity). A low cutoff will result in more false positives (lower specificity), but also more true positives (higher sensitivity).

Receiver Operating Characteristic (ROC) curves [18] show this tradeoff graphically. They plot the true positive rate versus the false positive rate for all possible cutoffs. A point on the diagonal line $y = x$ is what we would expect if the classifier was guessing randomly with a fixed probability of outputting "fake."[1] If a classifier has an ROC curve that is substantially above this line, than this is evidence that it is making predictions that are more accurate than would be expected by chance. The more an ROC curve is above this line, the better the classifier's predictive ability is. We use ROC curves extensively in this paper to compare the predictive abilities of various algorithms.

[1]For example, if a classifier output "fake" for all instances, it would have both a true and false positive rate of 1. If it output "fake" with probability 0.5 for each instance, it would have both a true and false positive rate of 0.5. Clearly, this line of reasoning can be extended to any point on the line $y = x$ by choosing the appropriate fixed probability of outputting "fake." Also note that any ROC curve will pass through both $(0, 0)$—for a threshold higher than any score returned by the algorithm—and $(1, 1)$—for a threshold lower than any score returned by the algorithm.

5. SUPERVISED METHODS

If data with labels indicating quality can be obtained, it could be used to train a classifier to make predictions about the quality of new data as it arrives. In this section, we investigate the feasibility of such an approach by testing the ability of off-the-shelf machine learning algorithms to create classifiers for the labeled data set from Organization A.

To perform this investigation, we used Weka [19], an open-source software distribution that implements a large number of machine learning algorithms. After a brief initial exploration, we settled on three widely-used machine algorithms that seemed to perform fairly well for this data set: logistic regression, the K2 algorithm for learning a Bayesian network [11] and random forests [4].[2] Our purpose was not to perform on exhaustive investigation to find the absolute best classifier, since such an approach would be unlikely to generalize beyond our data set. Rather, we wanted to see how well commonly-used off-the-shelf machine algorithms performed. We hoped to show that it only takes simple techniques to make good predictions.

We tested each of our algorithms twice, once with the completion-time attribute and once without. In Section 5.1 we test the ability of the algorithms to determine which *CHWs* were faking data. In Section 5.2, we test the ability of the algorithms to determine which *forms* were fake, a task that is presumably harder.

5.1 Interviewer-level Experiments

To test the ability of these algorithms to determine which of the 40 CHWs were creating fake data, we first aggregated the data by CHW as follows. We expanded each categorical question with $k > 2$ categories into k binary questions, one for each category. Then, we took the mean of each of these columns for each interviewer, creating a new table of labeled data consisting of 40 rows (CHWs) and 29 columns.

To test the algorithms, we used 10-fold cross validation. That is, we randomly partitioned the 40 rows into 10 groups of 4 rows. Then, for each group of 4 rows, we trained the classifier using the 36 other rows and obtained a score from the trained classifier on each of the 4 rows in the set. In this way, scores were obtained for all rows using separate training data.

The ROC curves for the scores obtained are shown in Figure 1. The results are quite positive, although the small size of the interviewer-level data should be kept in mind. Without including the mean form completion time, two of the algorithms were able to achieve 80% sensitivity at a specificity of greater than 90%. Including the mean form completion time improved the results further.

By drilling down into the data and the structure of the classifiers, we can gain some insight into how they are able

[2]The logistic regression algorithm used a ridge estimator [29]. The Bayesian network algorithm used the `SimpleEstimator` class implemented in Weka. The random forest algorithm used 10 trees, with each node considering $\log M + 1$ features, where M is the total number of features in the data set.

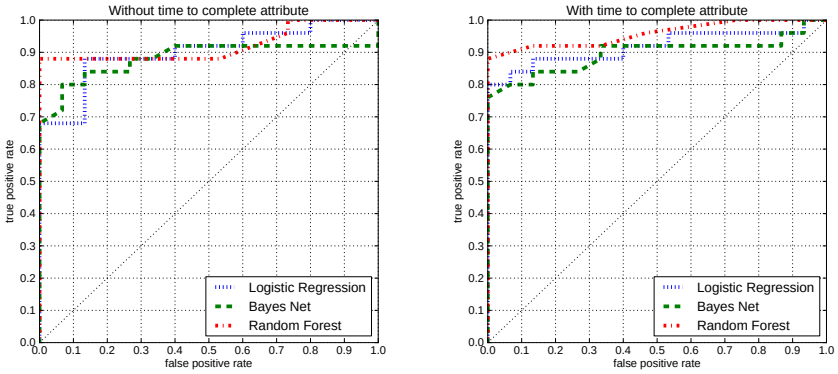

Figure 1: Interviewer-level ROC curves for supervised methods.

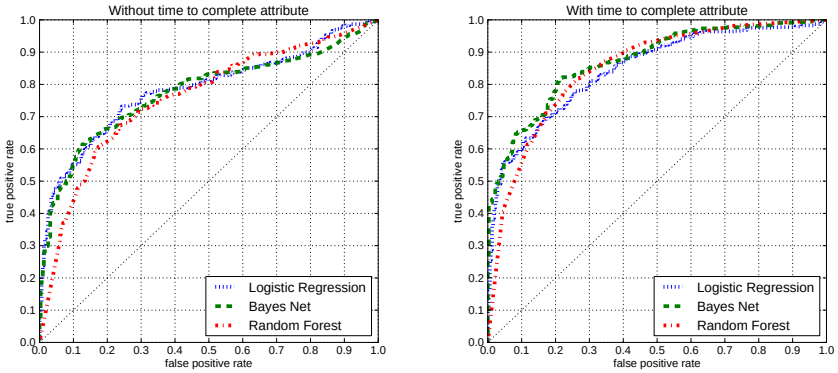

Figure 2: Form-level ROC curves for supervised methods.

to make their predictions. Completion time is a good predictor of whether a form was fake: the mean completion time of the fake forms was 148 seconds, whereas the mean completion time of the real forms was 240 seconds. This result is consistent with previous work using timing information to detect fabrication [5, 31, 33]. Of course, one should not discount the possibility that the lower completion time for the fabricated data is an artifact of the structure of the "fake data party."

As another example, the largest coefficient from the logistic regression classifier (apart from completion time) was for the question asking whether someone in the household had diarrhea in the last month. The reason for this is that CHWs who were faking data significantly overestimated the actual proportion of households for which that was true. The proportion of households with diarrhea in the fake data was 35%, whereas the proportion of households with diarrhea in the real data was only 5%.

5.2 Form-level Experiments

In the previous section, we leveraged the fact that in our data, CHWs either faked all or none of their forms. In practice, this might not be true: cheating interviewers might fake only some of their forms. In this section, we test the ability of machine learning algorithms to predict whether individual forms are real or fake. We use the same three machine learning algorithms from the previous section, and again test the algorithms both including and not including form completion time. For this case, there is no need to aggregate the data, so the input now consists of 846 rows of data, with 13 attributes (12 questions plus the form completion time).

The ROC curves for a 10-fold cross validation are shown in Figure 2. As we might expect, the results here are not as strong as when data is aggregated for each CHW. However, the classifiers still perform at a level substantially above what we would expect by chance. At a 90% specificity, the best-performing classifier using the completion-time attribute had a sensitivity of about 65%, and the best-performing classifier not using the completion-time attribute had a sensitivity of about 55%.

In both the interviewer-level and form-level settings, the success of the classifiers in distinguishing between the real and fake data suggests the feasibility of using them in the field to highlight potential fake data, given that labeled data can be obtained. We conjecture that such methods may also be useful in detecting low-quality data due to other types of interviewer error. The potential helpfulness of the completion-time attribute in making predictions highlights

an advantage of using mobile technology such as CommCare or ODK that can immediately upload this metadata to servers. In future work, we plan to investigate the potential of other meta-attributes of the data to predict fabrication.

6. UNSUPERVISED METHODS

In practice, it might be infeasible to collect the labeled data necessary for training classifiers. In this section, we investigate unsupervised techniques that do not rely on any labeled training data. We develop the techniques in Section 6.1 and validate them against the labeled data set from Organization A in Section 6.2.

6.1 The Techniques

In developing our techniques, we hypothesized that CHWs might not have a good sense for the true distribution of the quantities that they were measuring. If this were true, the answer distributions of CHWs who were faithfully interviewing clients would tend on average to be close to the true distribution, but CHWs who were fabricating data would come up with different answer distributions – and these might vary from the norm in different ways for different CHWs. Making the standard assumption from the unsupervised anomaly detection literature that there are fewer CHWs faking data than not [6], outlying answer distributions would tend to be fake.

It is important to remember that the techniques developed in this section do not detect fabrication directly. Rather, they detect interviewer-level anomalies in the data. Such anomalies could be due to other factors besides fabrication, such as varying interview styles, levels of experience, or geographies. All of these causes are potentially interesting to supervisors, however. Furthermore, in Section 6.2 we show that the fabricated portions of our data do tend to score highly according to the techniques developed here.

We start with some notation that will be used to define both techniques. Consider an interviewer i and a question j with a finite, discrete set of answer values. Let X_j be the range of values that question j can take. For any $x \in X_j$, let $g_j^i(x)$ be the number of times that interviewer i reports the value x for question j. For all i, j, and $x \in X_j$, let

$$f_j^i(x) = \frac{g_j^i(x)}{\sum_{x \in X_j} g_j^i(x)}$$

be the fraction of times that interviewer i reports the value x for question j.

Multinomial Model Technique

Given g_j^i, we can ask what the chance of seeing this distribution is, supposing that the answers are chosen according to a multinomial probability distribution with parameters estimated from the other interviewers. The lower this probability is, the more surprising interviewer i's answers are.

Specifically, we compute a score m_j^i for each interviewer as follows. For all $x \in X_j$, let

$$g_j^{-i}(x) = \sum_{i' \neq i} g_j^{i'}(x)$$

be the frequency count for value x over all interviewers besides i. Let

$$f_j^{-i}(x) = \frac{g_j^{-i}(x)}{\sum_{x' \in X_j} g_j^{-i}(x')}$$

be the normalized version of these counts. From this distribution, we can calculate the *expected* count for value x from interviewer i to be

$$E_j^i(x) = f_j^{-i}(x) \sum_{x' \in X_j} g_j^i(x') \ .$$

These expected counts can form the basis for a χ^2-test giving a p-value for the chance of seeing frequency counts that deviate at least as much as g_j^i, given the null hypothesis that the counts are drawn from a multinomial distribution with parameters given by f_j^{-i}. That is, we calculate the statistic

$$Q_j^i = \sum_{x \in X_j} \frac{(g_j^i(x) - E_j^i(x))^2}{E_j^i(x)}$$

and determine the probability of seeing a value at least this large in a χ^2 distribution with $|X_j|-1$ degrees of freedom. A smaller value of this probability indicates a greater amount of surprise for interviewer i's distribution. The final score, m_j^i, is the negative logarithm of this value. Thus, to summarize, we compute the final score to be

$$m_j^i = -\log \Pr\left[\chi^2_{|X_j|-1} \geq Q_j^i\right] \ ,$$

where a higher value indicates a greater amount of surprise.

S-Value Technique

The technique just described forms an expected distribution for an interviewer by taking the weighted mean of the distributions for all other interviewers. Using the mean as an estimator is sensitive to outliers: if there are a large number of interviewers who have bad data, the expected distribution obtained in this way might not be representative of the true population distribution. We take inspiration for our second measure from the field of robust statistics [24], which prefers medians to means because of their decreased sensitivity to outliers.

For any vector v, define $\mu_{1/2}(v)$ to be the median of the values in v. For a question j and a value $x \in X_j$, let $f_j(x)$ be the vector of values $f_j^i(x)$ for all interviewers i. The surprise for interviewer i's distribution for question j can be characterized by

$$d_j^i = \sum_{x \in X_j} \left| f_j^i(x) - \mu_{1/2}(f_j(x)) \right| \ .$$

We would like to use the d_j^i values to detect outlying distributions, but for some questions, these values might naturally be high for all interviewers. This could happen if, for example, the question's answer depends strongly on geography and interviewers are widely dispersed. For this reason, we normalize the d_j^i by their median. More precisely, let d_j be the vector of d_j^i values over all i, and let

$$s_j^i = \frac{d_j^i}{\mu_{1/2}(d_j)} \ .$$

This value, which we named the *s-value* to indicate surprise, will be higher the more interviewer i's distribution for question j is an outlier.

This technique is methodologically similar to the recently developed "Tariff Method" of verbal autopsy [25]. Here the goals are reversed, however. Instead of starting with signs and symptoms and hoping that we can identify the cause of

death, we start with the survey responses, and hope that we *cannot* identify the interviewer.

6.2 Validation

In this section, we use our outlier detection techniques as classifiers. The scores m_j^i and s_j^i can be thought of as indicating the degree to which these classifiers predict that question j from interviewer i is faked. The sums $m^i = \sum_j m_j^i$ and $s^i = \sum_j s_j^i$ can be thought of as indicating the degree to which these classifiers predict that interviewer i is faking all of her questions. Just as in Section 5, we can use ROC curves to show the sensitivity and specificity tradeoff given by these scores.

One complication is that we expect the outlier detection techniques to perform better when there is more real data than fake data. Our labeled data set, however, is 63% fake. To remedy this, we form test data sets by randomly choosing a subset of real and fake CHWs and only including forms from these CHWs when calculating the outlier scores. We vary the ratio of real to fake CHWs: for our first data set (5:1 ratio), we chose all 15 of the real CHWs and a random subset of 3 of the fake CHWs. For our second data set (1:1 ratio), we randomly chose 9 real CHWs and 9 fake CHWs. (In fact, we repeated this 1000 times to achieve a representative sample. The ROC curves that we show are the aggregate ROC curves over all 1000 repetitions.)

Figure 3 shows the ROC curves generated by the interviewer-level m^i and s^i values. For the 5:1 data set, both the multinomial model technique and the s-value technique perform quite well, achieving approximately an 80% sensitivity at a 90% specificity. Both methods continue to perform well at the 1:1 real to fake ratio. In this case, the s-value method achieves a 70% sensitivity at a 90% specificity, and the multinomial model method achieves a 60% sensitivity at a 90% specificity. It is perhaps surprising that the outlier detection technique works this well even when the amount of fake data is the same as the amount of real data. We speculate that this is because in aggregate, the CHWs come up with a fairly accurate estimate of the population averages, but the fake CHWs tend to deviate more from this estimate than the real CHWs. We also speculate that the reason for the s-value technique's better performance is that its use of medians to determine expected distributions is less sensitive to the extra noise introduced by the large number of fake CHWs.

Figure 4 shows the ROC curves generated by the m_j^i and s_j^i values, which correspond to predictions made at the level of CHW-question pairs. Because the scores are not aggregated by CHWs, the results in this case are not as strong. Nevertheless, for both the 5:1 and 1:1 data sets, both methods perform substantially above chance levels, achieving 30-40% sensitivity at 90% specificity. As with the interviewer-level predictions, the two techniques have roughly the same performance with the 5:1 data, but the s-value technique does better than the multinomial model technique for the 1:1 data.

7. EVALUATION ON FIELD DATA

A limitation of the experiments described so far is that the data from a fake data party might differ from the type of low-quality data that we care about in practice, which might err due to miscommunication or misunderstanding in addition to deliberate falsification. Furthermore, even if we limit our interests to detecting fake data, the data from a fake data party might be different than the data from interviewers faking data in the field who are highly motivated to avoid being caught. An ideal experiment for our methods would be to run them on true field data that is labeled according to quality, but we do not have any such data. We do, however, have a large corpus of unlabeled data, namely the unlabeled sets from Organizations A and B described in Section 3.

Figure 5 shows, for both data sets, the five most surprising interviewer-question pairs according to the multinomial model method described in Section 6. Each entry in the figure shows the question text, along with the interviewer's distribution compared to the overall distribution from all interviewers. For example, the figure shows that the most surprising distribution from an interviewer at Organization A, with an m_j^i value of 412.2, is "Do you want to know more about family planning?" Interviewer 1 said that 89% of her clients wanted to learn more about family planning, whereas the proportion of clients, over all interviewers, who responded this way was only 29%. It seems clear that a supervisor would want to follow up on an anomaly like this. Indeed, after showing a similar table to Organization B as part of a demo, they informed us that they found information like this very useful and were interested in the possibility of using algorithms such as these as part of their operations.

One could ask whether the types of anomalies shown in Figure 5 are what we would expect from chance fluctuations. We do not think this is the case. One justification for this is shown in Figure 6, which helps to visualize interviewer correlations in the m_j^i values at organization B. Each bar in this figure corresponds to an interviewer; the height is the number of questions for which the interviewer's distribution has an m_j^i value in the top 95th percentile of the values for all interviewer-question pairs. The bars are sorted from left to right according to this frequency, and only the top 15 interviewers are shown.

Figure 6 shows that two interviewers have many more surprising answer distributions than other interviewers. If the m_j^i values were independently and identically distributed, then for a given interviewer, the chance that a particular question distribution would make it in the 95th percentile is 0.05. Thus, for the first interviewer in Figure 6, the chance that at least 27 questions would randomly appear in the 95th percentile is the chance that a binomial random variable with parameters $n = 103$ (the number of questions) and $p = 0.05$ would achieve a value of at least 27, which is 8.5×10^{-13}. Since there are 26 interviewers, the chance that this would occur for any interviewer is less than $26 \cdot 8.5 \times 10^{-13} = 2.2 \times 10^{-11}$. Hence, it seems highly likely that something besides chance alone is causing this interviewer's distributions to be so different; a supervisor would be well-advised to investigate further.

8. DISCUSSION

In this paper, we have demonstrated that our algorithms have the potential to detect low-quality data collected by human interviewers. We showed that both the supervised and unsupervised methods can detect the fake data in the labeled set with a high sensitivity and specificity. We also showed that the unsupervised techniques can find compelling anomalies in real data sets from two separate community-based programs using mobile data collection.

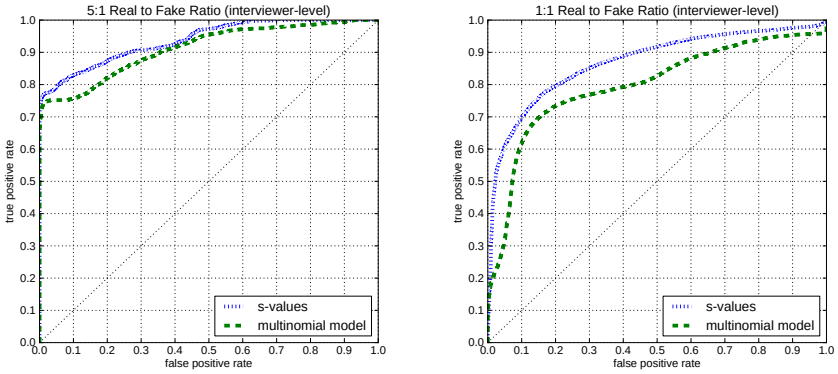

Figure 3: Interviewer-level ROC curves for unsupervised methods.

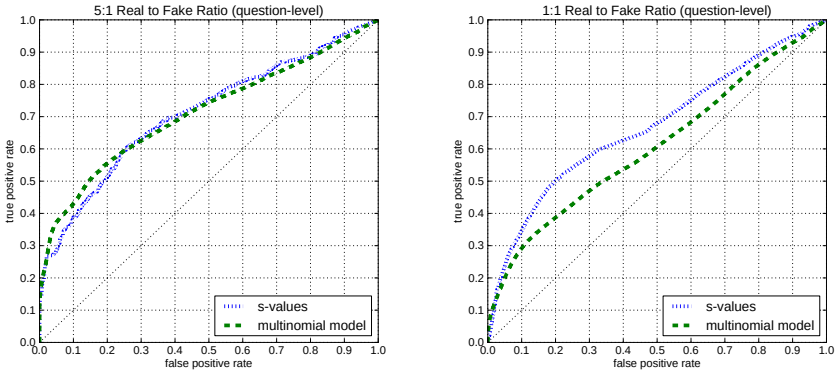

Figure 4: Question-level ROC curves for unsupervised methods.

A direction for future work is to obtain labeled data with fake forms that are generated in a more realistic way. At the fake data party, the CHWs did not know that their data was being evaluated by its quality. In actual operations, CHWs might be motivated to spend more effort faking their data in order to avoid detection. We plan to test if our methods remain successful when CHWs are rewarded for avoiding detection.

We view this study as a preliminary step in the design of tools for automated data quality control for mobile data collection. The design and testing of such tools is left for future work, but here we briefly discuss considerations regarding how they might use the algorithms discussed in this paper.

Tools that use the supervised learning methods would have the advantage of flexibility, since the characteristics of low-quality data are likely to vary from one context to another. If labeled data can be obtained, then a tool could be trained on this data and then provide real-time feedback to supervisors as data is collected. This feedback could be given at the form-level, as in Section 5.2, or at the interviewer-level, as in Section 5.1. In the latter case, data would have to be aggregated by interviewer over a given time period. This time period could be a sliding window, or the predictions could be made at specified time intervals, such as before weekly meetings with interviewers.

Tools that use the unsupervised methods would have the advantage of not requiring labeled data. The methods presented in Section 6 aggregate data at both the question and interviewer level. The question-level aggregation may be useful if interviewers are likely to misunderstand or misreport specific questions. Because of the interviewer-level aggregation of the methods from Section 6, they require aggregation of data over a window of time. However, one could also imagine unsupervised methods that work at the form-level and which could therefore provide immediate feedback as forms are completed. These methods might look at completion time or unusual combinations of answers, for example.

Tools could also take advantage of a richer set of metadata than timing information alone. Mobile data collection tools could record attributes from interviewers' interaction traces, such as skipped questions, questions answered out of order, the amount of scrolling (which could be a proxy for effort), or attempts to input out-of-range values (which could be a proxy for competence). These features could be leveraged by both supervised and unsupervised techniques.

Finally, one could imagine an adaptive system combining

Organization A (Tanzania)

```
Question 1 (score = 412.2): Do you want to know more about
family planning?
                            No   Yes
Interviewer 1 (312 forms):  11%  89%
Everyone (3131 forms):      71%  29%

Question 2 (score = 335.4): Is the client available?
                            No   Yes
Interviewer 2 (258 forms):  91%  9%
Everyone (4321 forms):      28%  72%

Question 3 (score = 101.9): Do you want to know more about
family planning?
                            No   Yes
Interviewer 3 (100 forms):  11%  89%
Everyone (3131 forms):      71%  29%

Question 4 (score = 95.5): Did you give any referral in
this household?
                            No   Yes
Interviewer 3 (100 forms):  62%  38%
Everyone (3131 forms):      93%  7%

Question 5 (score = 85.6): Did anyone in the household
have a fever yesterday or today?
                            No   Yes
Interviewer 3 (100 forms):  68%  32%
Everyone (3131 forms):      94%  6%
```

Organization B (Uganda)

```
Question 1 (score = 41.3): If yes, which of these
marketing activities do you do with this group?
(Choosing crops or varieties to grow and market)
                            No   Yes
Interviewer 1 (16 forms):   44%  56%
Everyone (265 forms):       94%  6%

Question 2 (score = 36.5): Why did you find yourself
without food? (There was no food distribution)
                            No   Yes
Interviewer 2 (8 forms):    0%   100%
Everyone (265 forms):       94%  6%

Question 3 (score = 35.5): Why did you find yourself
without food? (No one was willing to offer us some food)
                            No   Yes
Interviewer 2 (8 forms):    0%   100%
Everyone (265 forms):       93%  7%

Question 4 (score = 32.1): What did you do with the
information you got from the FIELD OFFICER? (I started
storing my produce at a warehouse or collection point)
                            No   Yes
Interviewer 3 (9 forms):    11%  89%
Everyone (265 forms):       89%  11%

Question 5 (score = 31.7): What information have you
ever got from a FIELD OFFICER? (Supplying to WFP)
                            No   Yes
Interviewer 3 (9 forms):    0%   100%
Everyone (265 forms):       84%  16%
```

Figure 5: The five most surprising questions from Organization A and Organization B.

supervised and unsupervised methods. To start, it could use unsupervised methods to alert supervisors to suspicious patterns. These alerts might be based a weighted combination of factors, such as how outlying the answer distributions are (using one of the methods from Section 6), form completion time, or the number of interview refusals. As supervisors follow up on these alerts, they would provide feedback to the system regarding the true cause of the suspicious data. The system could then use this feedback to adaptively modify the weights of the different factors. Indeed, one could frame this problem as an instance of active learning [41] in which labels can be obtained, but for the cost associated with the follow ups.

9. ACKNOWLEDGMENTS

We thank Gilbert Agaba, Nick Amland, Ken Bayona, Gaetano Borriello, Deborah Gitonga, Kayur Patel, Amelia Sagoff, Paul Ssengooba, the Grameen Foundation's AppLab, Dimagi, and ITIDO. We thank the nurses, community health workers, knowledge workers, mothers, and farmers for their time and effort. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-0718124).

10. REFERENCES

[1] P. P. Biemer and L. E. Lyberg. *Introduction to Survey Quality*. Wiley-Interscience, 2003.

[2] P. P. Biemer and S. L. Stokes. The optimal design of quality control samples to detect interviewer cheating. *J. Official Statistics*, 5(1):23–39, 1989.

[3] S. Bredl et al. A statistical approach to detect cheating interviewers. Technical report, 2008.

[4] L. Breiman. Random forests. *Machine Learning*,

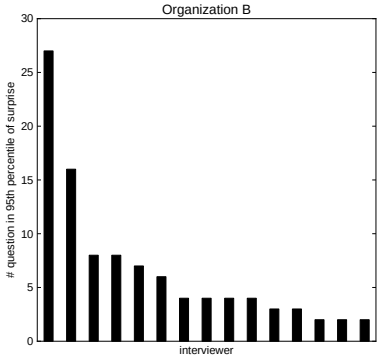

Figure 6: The number of questions per interviewer at Organization B that are in the 95th percentile of surprise, according to the m_j^i scores. The interviewers are sorted from left to right according to the number of questions in the 95th percentile. (Only the top 15 interviewers are shown.) Note the abnormally high frequency of the first two interviewers, which suggests the m_j^i values are not due to chance alone.

45:5–32, October 2001.

[5] J. M. Bushery et al. Using date and time stamps to detect interviewer falsification. *ASA Section on Survey Research Methods*, pages 316–320, 1999.

[6] V. Chandola et al. Anomaly detection: A survey. *ACM Comput. Surv.*, 41:15:1–15:58, July 2009.

[7] K. Chen et al. Designing adaptive feedback for improving data entry accuracy. In *UIST*, 2010.

[8] K. Chen et al. Usher: Improving data quality with dynamic forms. In *TKDE*, 2010.

[9] K. Chen et al. Data in the first mile. In *CIDR*, 2011.

[10] E. K. Chiteri et al. Quality control on HDSS automated application systems. `http://groups.google.com/group/ict4chw/browse_thread/thread/688fe36f845d56a5/`, Acc. 20 Oct. 2011.

[11] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[12] L. P. Crespi. The cheater problem in polling. *Public Opinion Quarterly*, 9(4):431–445, 1945.

[13] B. DeRenzi. *Technology for Workforce Performance Improvement of Community Health Programs*. PhD thesis, University of Washington, 2011.

[14] B. Derenzi et al. Mobile phone tools for field-based health care workers in low-income countries. *The Mount Sinai J. Medicine*, 78(3):406–18, 2011.

[15] B. DeRenzi et al. Improving community health worker performance through automated SMS reminders. In *ICTD*, 2012.

[16] eMocha, `http://emocha.org/`. Acc. 20 Oct. 2011.

[17] EpiSurveyor, `http://www.episurveyor.org/`. Acc. 20 Oct. 2011.

[18] T. Fawcett. An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27:861–874, June 2006.

[19] M. Hall et al. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[20] C. Hartung et al. Open Data Kit: Tools to build information services for developing regions. In *ICTD*, 2010.

[21] J. M. Hellerstein. Quantitative data cleaning for large databases. Technical report, 2008.

[22] T. P. Hill. The difficulty of faking data. *Chance Magazine*, 12(3):31–37, 1999.

[23] C. C. Hood and J. M. Bushery. Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *ASA Section on Survey Research Methods*, pages 820–824, 1997.

[24] P. J. Huber. *Robust Statistics*. Wiley, 1981.

[25] S. L. James et al. Performance of the Tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31, 2011.

[26] G. Judge and L. Schechter. Detecting problems in survey data using Benford's law. *J. Human Resources*, 44(1):1–24, 2009.

[27] R. Khan. ClickDiagnostics experiences from Bangladesh. `http://groups.google.com/group/ict4chw/browse_thread/thread/96dd344a921f124e`, Acc. 20 Oct. 2011.

[28] E. A. Krejsa et al. Evaluation of the quality assurance falsification interview used in the census 2000 dress rehearsal. *ASA Section on Survey Research Methods*, pages 635–640, 1999.

[29] S. le Cessie and J. C. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1997.

[30] U. Lehmann and D. Sanders. Community health workers: What do we know about them? Technical report, World Health Organization, 2007.

[31] J. Li et al. Using statistical models for sample design of a reinterview program. *ASA Section on Survey Research Methods*, pages 4681–4695, 2009.

[32] G. Mhila et al. Using mobile applications for community-based social support for chronic patients. In *HELINA*, 2009.

[33] J. Murphy et al. A system for detecting interview falsification. In *Am. Assoc. for Public Opinion Research*, 2004.

[34] C. J. L. Murray and J. Frenk. Health metrics and evaluation: strengthening the science. *Lancet*, 371(9619):1191–9, Apr. 2008.

[35] T. Parikh et al. Mobile phones and paper documents: Evaluating a new approach for capturing microfinance data in rural india. In *CHI*, 2006.

[36] S. Patnaik et al. Evaluating the accuracy of data collection on mobile phones: A study of forms, SMS, and voice. In *ICTD*, pages 74–84, 2009.

[37] J. Porras and N. English. Data-driven approaches to identifying interviewer data falsification: The case of health surveys. *ASA Section on Survey Research Methods*, pages 4223–4228, 2004.

[38] A. L. Ratan et al. Managing microfinance with paper, pen and digital slate. In *ICTD*, 2010.

[39] J.-P. Schraepler and G. G. Wagner. Identification, characteristics and impact of faked interviews in surveys. *All. Stat. Archive*, 89(1):7–20, 2005.

[40] I. Schreiner et al. Interviewer falsification in census bureau surveys. *ASA Section on Survey Research Methods*, pages 491–496, 1988.

[41] B. Settles. Active learning literature survey. Comp. Sci. Tech. Report 1648, University of Wisconsin–Madison, 2009.

[42] C. Shäafer et al. Automatic identification of faked and fraudulent interviews in surveys by two different methods. *ASA Section on Survey Research Methods*, pages 4318–4325, 2004.

[43] D. Swanson et al. Detecting possibly fraudulent or error-prone survey data using Benford's law. *ASA Section on Survey Research Methods*, pages 4172–4177, 2003.

[44] C. F. Turner et al. Falsification in epidemiologic surveys: Detection and remediation. Technical Report 53, Research Triangle Institute, 2002.

[45] P. Yu et al. The development and evaluation of a PDA-based method for public health surveillance data collection in developing countries. *Int. J. of Med. Informatics*, 78(8):532–42, 2009.

[46] M. Zimic et al. Can the power of mobile phones be used to improve tuberculosis diagnosis in developing countries? *Trans. Roy. Soc. of Tropical Medicine and Hygiene*, 103(6):638–40, 2009.