Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking

Eric D. Ragan
Texas A&M University
eragan@tamu.edu

Hye-Chung Kum
Texas A&M University
kum@tamu.edu

Gurudev Ilangovan
Texas A&M University
ilangurudev@gmail.com

Han Wang
Texas A&M University
hanwang@tamu.edu

ABSTRACT

Effective use of data involving personal or sensitive information often requires different people to have access to personal information, which significantly reduces the personal privacy of those whose data is stored and increases risk of identity theft, data leaks, or social engineering attacks. Our research studies the tradeoffs between privacy and utility of personal information for human decision making. Using a record-linkage scenario, this paper presents a controlled study of how varying degrees of information availability influences the ability to effectively use personal information. We compared the quality of human decision-making using a visual interface that controls the amount of personal information available using visual markup to highlight data discrepancies. With this interface, study participants who viewed only 30% of data content had decision quality similar to those who had full 100% access. The results demonstrate that it is possible to greatly limit the amount of personal information available to human decision makers without negatively affecting utility or human effectiveness. However, the findings also show there is a limit to how much data can be hidden before negatively influencing the quality of judgment in decisions involving person-level data. Despite the reduced accuracy with extreme data hiding, the study demonstrates that with proper interface designs, many correct decisions can be made with even legally de-identified data that is fully masked (74.5% accuracy with fully-masked data compared to 84.1% with full access). Thus, when legal requirements only allow for de-identified data access, use of well-designed interface can significantly improve data utility.

Author Keywords

Human-computer interaction; privacy-preserving interactive record linkage; information privacy

CHI 2018, April 2126, 2018, Montreal, QC, Canada.

https://doi.org/10.1145/3173574.3173900

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g., HCI)—User Interfaces. K.4.1.Computers and society—Public Policy Issues: Privacy

INTRODUCTION

Data systems are everywhere, and more than just a few companies and organizations collect vast quantities of records about people. Data integrity is an obvious requirement for effective data science. Data needs to be complete and of high quality in order for it to be useful, and easy access to records of interest is generally essential for any data management system. However, such completeness and accessibility can cause privacy concerns when dealing with *sensitive personal information* (SPI) or *personally identifiable information* (PII) that can be used to identify a specific person. Personal data records include private information (e.g., social security numbers, identification numbers, names, or date of birth) that is commonly used for identity verification.

Depending on the purpose or meaning of the data set, simply knowing that a person is in that data set could disclose sensitive information [24, 17]. For example, knowing that a particular person is in a cancer patient registry divulges information about that person's medical condition. Databases often also include additional personal data (e.g., credit card numbers, specifics of medical history) for which privacy is paramount.

To increase privacy protection, many data systems commonly employ mechanisms to: limit the information that is accessible (i.e., sanitize the data) [17], limit who can access the data (i.e., access control) [34], and/or increase transparency by continuously monitoring, auditing, and penalizing for misuse (i.e., information accountability) [36]. But such methods do not address several separate issues related to human access of that information. Regardless of how the data is secured and protected, in many cases, people need access to the PII in order to interpret and use the data for real purposes. These are situations involving legitimate inspection or analysis of the data, such as for data work involving medical, financial, welfare, and economics data sources. For example, in order to effectively make use of the data, professionals and researchers often need to handle data inspection and cleaning to

validate datasets, fix errors, standardize values, resolve duplicate records, or integrate data sources [20]. In these scenarios, approved individuals are often granted complete access to needed data sets, but with the cost of privacy risk.

Effective use of data involving personal or subject information often requires different people to have access to the PII, which significantly reduces the personal privacy of those whose data is stored and opens up greater opportunities for personal identification, data leaks, or social engineering attacks. This is a major problem because it seems the primary options are to: (a) increase the number of people with data access to maximize utility of the data at the expense of privacy, or (b) increase restrictions on data access, which limits the throughput for legitimate use of the data.

Our research addresses this problem and studies the tradeoffs between privacy and utility of personal information. More specifically, we focus on the impact of varying degrees of information disclosure on the ability to effectively use personal information. We hypothesize that, in many cases, it should be possible to greatly limit the amount of personal information available to human decision makers without negatively affecting utility or human effectiveness.

To address our research goals, we conducted an experiment involving human decision making in the context of *record linkage*. Record linkage requires judgment about data records—in our case, records containing PII—and whether records coming from different databases refer to the same person. Through a user study, we tested the quality of human decisions using a visual interface that controls the amount of PII available. Additionally, we study the impact of supplemental visual markup designed to facilitate understanding of data discrepancies. The presented research demonstrates that with the appropriate representation of metadata, significant limits can be applied to the available PII without compromising the quality of human decision making.

BACKGROUND

Our research draws from and combines a variety of areas, including privacy, human-computer interaction, visualization, and record linkage. In this section, we provide an overview of key background information that is important for interpreting the presented study.

Information Privacy and Uncertainty

Privacy of personal information is of obvious importance to avoid outcomes such as identify theft or financial loss. Perception of privacy is a major factor that influences how willing people are to use technology [27, 7] or share personal records for medical research [6, 30].

It is well known that different pieces of information can be combined to make inferences and learn more than was intended from a single data source, which is why many sensitive data centers limit what data can be combined from sources that include sensitive data [25, 28, 35]. However, even in those situations, it is not possible to limit previously known background knowledge that can help make new inferences. For example, as potential participants in medical research, patients rarely have concerns about researchers using their data for research, but many are concerned with *local privacy* [11, 18]. Local privacy is concerned with information sharing among people who we know, such as family, friends, coworkers, or neighbors. Because such local members of our lives might be familiar with additional background information about a person, the possibility of combining that knowledge with additional sensitive information can result in undesirable social or professional situations. A recent study on sharing data for medical research found that 81% of the patients (N = 3,516) were somewhat comfortable sharing electronic health data with researchers for research purposes unrelated to their personal health care, but only 60% were comfortable with sharing data for research purposes with someone they know (e.g., friends, neighbors, and coworkers) [29]. This indicates the need for mechanisms to protect individuals whose data is captured against local privacy risk.

To account for such risks in the design and management of data systems involving personal information, the goal is often to limit data access. However, limiting or hiding data can introduce challenges for legitimate data utility. In a review covering data issues across various disciplines, Boukhelifa et al. [3] discuss common sources of data uncertainty (variability, temporality, inconsistency, missing data, and bias). They explain common approaches (understanding, minimizing, exploiting, or ignoring) for coping with uncertainty when working with the data, and that the goal is to minimize rather than ignore uncertainty during decision making. Regarding our research interests in privacy and information disclosure, the concern is that maximizing privacy will increase uncertainty, making decision-making much more difficult.

Privacy-Preserving Visualization

Numerous researchers have studied the design of visual interfaces to facilitate understanding data while retaining privacy (e.g., [14, 15, 9]). A common approach is to use visual aggregation to present general data trends and relationships of groups while preventing the identification of any specific individual. This can be achieved through different visualization designs. For example, Chou et al. [9] demonstrated the use of representation similar to a Sankey diagram or a banded parallel-coordinates plot to summarize lifestyles and common sequences of people visiting common locations. Dasgupta et al. [13] present methods for clustering and binning similar ranges of values that can be shown through parallel coordinate views, binned scatterplots. They discuss the importance of consideration for perception of uncertainty when viewing aggregated and anonymized information, and they also discuss how their clustering approach might be applied to other types of visual designs to highlight the most prominent relationships.

Other work by Dasgupta et al. [15] studied the use of different visual encodings to show collections of personal health records. They explain that privacy-preserving visualizations might contradict the traditional goal of maximizing accurate presentation of data values. They discuss how visual encodings that are less accurate for numerical representation (e.g., area, color) might be preferred over representations that

can be read more accurately (e.g., position, length) in order to make it more difficult to discern population patterns that might compromise privacy.

Considering privacy with sensitive images, Çiftçi et al. [10] studied the use of color distortion to mask facial features and make it more difficult to identify faces in photographs. While their goals are similar to our own, our work focuses on tabular data rather than images, and substantially different methods are needed for the different data types.

Our work is motivated by the assertion that awareness of privacy is important—especially for those who have access to PII. A related hypothesis is that limiting the amount of information disclosure could encourage those with data access to have greater respect for privacy. This concept was supported in study by Chang et al. [8], which found that demonstrating discretion of disclosure of personal information can increase awareness and respect for personal privacy. In this study, participants reviewed user profile pictures as used in social media applications. Some participants viewed more revealing or sexualized profile images, while others viewed more conservative or reserved images. When participants viewed the less revealing images, they were less likely to share personal information or to recommend that others also share more information. If this effect generalizes to other data disciplines, it could mean visible methods for reducing information disclosure might promote conscientious care with personal information.

Record Linkage

Properly integrated person-level population data can provide invaluable insights into the collective impact of the myriad of policies and individual decisions in our society. These insights can inform new policy decisions, inform resource allocation decisions, identify opportunities for early intervention, and identify root causes of social and public health problems. However, the integration of population-level data across a diverse set of sources is a challenging task [25].

Integrating data from diverse, heterogeneous systems requires entity resolution, more widely known as record linkage. Record linkage is the process of identifying record pairs that belong to the same real-world entity (e.g., a person). While linkage is trivial when records share a common identifier (e.g., a social security number that can act as a primary key), such identifiers are commonly missing across data sources, and the addition of data errors or formatting differences can introduce further uncertainty. Thus, full access to personal identifying information is often required to ensure proper linkage in multiple stages of record-linkage processes such as proper tuning of parameters in automatic linkage algorithms, data preparation for automatic algorithms (e.g., cleaning, standardization), developing training and test datasets, and manually making quality linkage decisions to refine results from automatic methods. Record linkage is a critical task for data-intensive biomedical and social science research to reap the benefits of big data for social good, and record linkage studies are becoming more and more common in fields such as health, child welfare, and economics [25].

However, privacy is a major concern for record linkage due to the necessary use of PII to make linkage decisions. Existing research on privacy-preserving record linkage is based on the application of cryptology to link data securely given a predetermined linkage function, which is unknown in most real situations [19, 34]. Recently, there has been a push towards encryption-based record linkage that promises to preserve privacy by guaranteeing anonymity. However, little is known about the linkage accuracy achieved by such software. Privacy protection in record linkage is fundamentally different from other privacy-preserving data operations because the goal is to exactly identify the entity represented by the data being linked so that data sources can be accurately merged. For instance, one must be able to distinguish between family members or twins in the data [24].

Absence of a common, error-free, unique identifier makes exact matching solutions inadequate, leading to approximate methods (probabilistic or deterministic) that require data cleaning as well as manual resolution of ambiguous matches [26]. High quality data integration requires human interaction to tune the results from these machine-only systems [23, 12]. For example, in a 2011 study linking cancer registry data to Medicare and Medicaid data, of a total of 109,925 individuals that were being linked, 16,288 needed to be confirmed through manual verification [2]. In another study linking cancer registry data, over 4,000 of 131,000 matches were found manually after reviewing many more false matches between twins and family members [4]. A systematic-comparison study in 2017 found that automated linking results in high rates of erroneous matches ranging from 17% to over 60% across multiple real datasets. In addition, the study suggests that more modern automatic algorithms based on machine learning need training data that must be constructed by people [1].

Regardless of the method, systematic linkage errors are inevitable in automatic algorithms and can result in selection bias [5, 1]. Human involvement is essential to obtain high quality, bias-free linkages. This means that some information must be revealed to trusted persons to produce accurate linkages. In the manually intensive process, linkage experts spend months using software to clean and tune the linkage models, during which many choices and assumptions are made. These are typically difficult to document and verify. As a result, most linkages are not reproducible because the tuning step is difficult to replicate. Human intervention is essential, and including more human reviewers is often necessary to verify linkage accuracy. In practice, such trusted personnel are often trained temporary workers, interns, or graduate students who may have a high rate of turnover. The goal is to limit the amount of PII that people can access to only the minimum necessary, but the problem is that limiting data can reduce its utility for accurate decision making.

Interfaces for Interactive Record Linkage

Many researchers have studied the user of interactive systems to assist in data review and verification efforts. For example, Kandel et al. [21] presented *Profiler*, which combines multiple statistical methods with interactive visualization and

data cleaning capabilities to help address data problems such as anomalies, duplicate records, or missing data. This application demonstrates the power of visual analytics for automatically flagging issues and providing detailed data views, though it might be better used by experts in data science than by domain specialists (or novices) who are often tasked with the final decisions on record linkage tasks.

Other researchers have looked to more visual designs. For instance, the *D-Dupe* system [22] uses multiple views to support interactive comparisons of database items that are possible duplicates of entities (such as records of people). The tool shows potential duplicate pairs along with similarity metrics, and it allows users to inspect full details of any entity. The tool also includes a relational view that shows connections to other entities (as an example with paper authors, connections could be based on co-authorship or organizational affiliation). Following a similar example, Shen et al. [33] presented *NameClarifier*, an interactive record-linkage system for author names in publications. The system is designed to help resolve author ambiguity due to authors having common names or incomplete information. The application uses multiple views, including a graph view to show co-authorship and a temporal view to help contextualize a paper's topic compared to topics of other publications by the author at different periods of time. The visual relational views of [22] and [33] leverage designs that summarize mappings to larger entity or attribute spaces (see [31]).

Taking a similar approach but focusing on a different purpose, *PeerFinder* [16] prioritizes the comparison of a single person to related groups of people rather than comparisons of an entity to other individual records. The tool demonstrates how to take advantage of multiple views to assist in comparing multiple attributes and groups, and it supports parameter tuning for query criteria.

These examples take advantage of additional known group or network data about entities to assist in decision making. However, it is important to note that the ability to show entity relationships is not available for all data scenarios—particularly those where information access is already limited. Viewing such additional information would often be counter to the goals of privacy preservation and limiting information access.

EXPERIMENT TASK AND APPLICATION

To study the tradeoffs between level of information disclosure and information utility for decision making, we designed an experiment using an interactive record linkage task. The study scenario assumes it is necessary to perform record linkage and remove duplicate entities for two data sources that include records with personal information. Each record in our scenario included a name (first name and family name), date of birth, gender, race, and an identification number (as a proxy for a social security number or other sensitive identifier). The scenario assumes an automated record linkage algorithm has already handled the simple linkage cases, but there are remaining pairs of data records that could not be automatically linked due to uncertainty. Therefore, participants were asked to review pairs of records and decide whether the two records refer to the same person or come from two different people with similar identifying information.

Figure 1 shows an example of the base case for viewing record pairs. The study application showed a list of pairs; in this example, two pairs are shown. Even when pairs are mostly similar, numerous factors can influence the level of uncertainty or difficulty in making linkage decisions. For example, the following are common problems with data records associated with the same person: (1) variations in first name (common due to nicknames or spelling variations), (2) last name changes (common for women upon marriage), (3) swaps between first name and last name (common when names are entered in wrong fields), or (4) errors in date of birth (common due to typos and different date formatting conventions). Common problems are also found for different people who might be mistakenly believed to be the same person due to having similar information. Examples include: (1) a father and son who have the same name with a different suffix (e.g., Jr., III), (2) twins with different first names but the same last name and birthday, or (3) people with common names who share the same birthday.

In Figure 1, we can see pair #1 has the same values for all fields except the name fields, where the first and last names are swapped (*Jason Boyle* and *Boyle Jason*). Since *Jason* is a common first name, the name swap and other similarities indicate a high probability that these records refer to the same person. Pair #2 also has similar values but with a missing ID number and similar birth dates but with different years (1930 and 1903), which could easily be attributed to a mistype. However, the names in pair #2 are highly common names for the originating population (i.e., these names occur frequently in the United States). As a result, we cannot be confident that these two records refer to different people.

As we can see from this example, decision making for this task often involves thinking in terms of probability and consideration for different types of real-world scenarios. For each pair, participants were tasked with deciding whether each pair referred to the same or different person, and they marked each decision with an indicator of their level of confidence. Responses used a six-point scale ranging from "highly confident the people are the same" to "highly confident the people are different", with lower levels of confidence in between. Figure 1 includes the response input panel in the rightmost column of each record pair. The letters H, M, and L stand for *high*, *moderate*, and *low* confidence levels.

The study application also includes features to provide supplemental information to assist human reviewers in easily identifying the similarities and differences between records in each pair, as visual data descriptors can help people to more easily inspect tabular data [32]. The interface can show visual markup to highlight minor character differences, significant differences in field values, transposed values within a field, swapped entity values, and missing values. Examples of the visual markup features in our application are shown in Figures 2 and 3. Icons denote different types of discrepancies, and different characters can be highlighted in different colors to make it easy to identify differences. In addition, the sys-

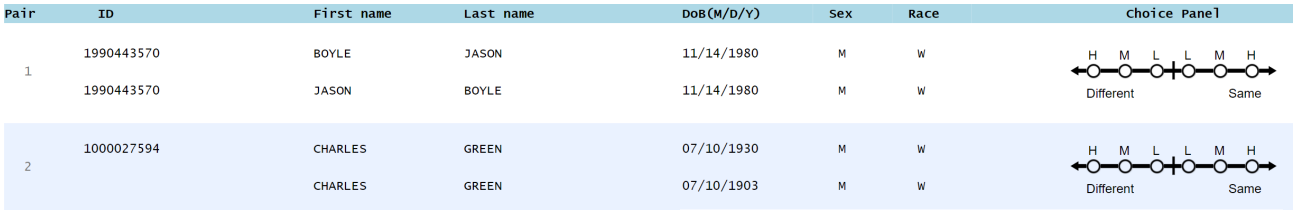

Figure 1. Above are two pairs of data records in the study application in the *baseline* condition with all information visible.

Figure 2. Example from the study application showing supplemental markup and value masking. The two pairs in this example are shown in the *moderate* condition using the same pairs shown in Figure 1. The visual markup highlights discrepancies, provides information about name frequency, and hides common values.

tem can include extra information under the *FFreq* and *LFreq* columns relating to the frequency of first and last names (respectively) in each data source. For example, a circled number *one* indicates that the corresponding name is a unique instance in the data source where that record was from, while an infinity icon represents a high frequency (more than 100) of the name. Other icons indicate rare (2–5 instances) or moderate frequencies (6–100 instances). Finally, we note the study application can hide data values in records. Identical fields can be replaced with check marks, and identical characters can be masked with other characters (asterisks, ampersands, or pound signs) to denote identical or swapped values without revealing real values. We use this functionality to test different levels of information visibility in our experiment. Further explanation of the tested combinations of visual markup and different levels of information hiding will be explained in the following sections.

EXPERIMENT

Using a record-linkage scenario, we conducted a controlled experiment to evaluate how varying degrees of information disclosure can affect decision making for tasks that require interpretation of personal data.

Hypotheses

This research is motivated by the need to understand the extent to which it is feasible to de-identify personal data without negatively affecting the utility of the data for decision making. Our high-level hypothesis is that even legally de-identified data—where personal details are hidden—can be effectively used for decision making tasks that generally rely on personal details, but we expect that achieving this will require an appropriate interface that can sufficiently convey the most important meta-information for the decision-making task. However, we also expect that availability of certain data details is sometimes necessary for some difficult data decisions. Therefore, we hypothesize that complete data hiding can negatively affect decision quality, but partial reductions to information disclosure that legally de-identify entities can be sufficient for most quality decision.

Applied to the record linking scenario of our study, we expect the use of value masking and visual markup can sufficiently portray differences to limit the amount of personal information needed to make linkage decisions. This means we predict records with hidden details can be linked with a level of accuracy similar to the base case with unlimited information disclosure, but we expect to see a reduction in quality for extreme data masking. We summarize these hypotheses as follows:

H1: With an appropriate interface, significant limits on data availability can be enforced without compromising decision quality.

H2: There is a limit to how much data can be hidden before negatively influencing the quality of judgment in decisions involving person-level data.

H3: The addition of supplemental visual information can help expedite record linkage decisions by making it easier to identify the types of differences.

Experimental Design

The experiment followed a between-subjects design with five conditions. We summarize the differences among the conditions below, but the differences can be seen more easily in Figure 3.

- *Baseline (full disclosure with no markup)*: This condition displayed the full information from all records. As the baseline condition, no visual markup was available to highlight differences, and name frequency indicators were not included. This condition represents how record linkers would normally view records without any privacy protec-

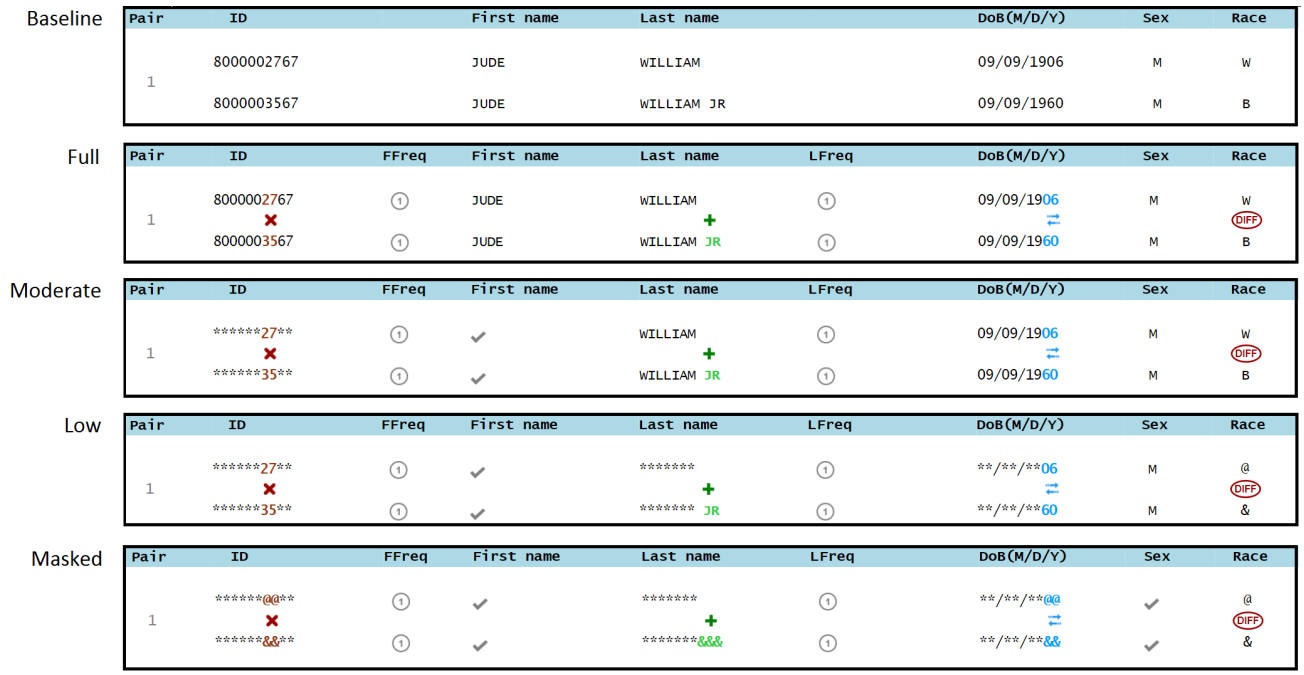

Figure 3. Examples showing one record pairs in the five different experimental conditions. These views show the same underlying data, but the visuals and amount of symbol substitution varies based on the viewing condition.

tion, as it is similar to the conventional method used at most record linkage centers worldwide.

- *Full* (*full disclosure with markup*): This condition also displayed the full information from all records. No data values were hidden. In this view, pairs were augmented with graphical icons and color-coded text to highlight the differences, and frequency icons were included.

- *Moderate* (*moderate disclosure with markup*): The goal for this condition was to hide information except for the most relevant items believed to assist decision making. Information was hidden in columns for pairs having the same values for both records, and check marks were instead shown to indicate matching values. Because ID numbers are often considered highly sensitive types of information (such as social security numbers) and the raw value is not useful information for linkage decisions, full IDs were never revealed in this mode. Supplemental visual markup (difference icons, colored text, and frequency icons) was again used to highlight differences (the same as in *full*).

- *Low* (*low disclosure with markup*): The goal for this condition was to reveal as few data characters as possible while showing how pairs were different. As in the *moderate* condition, check marks were shown instead of values when the columns were the same, and visual markup (difference icons, colored text, and frequency icons) was again used to highlight differences. Unlike in the *moderate* condition, little information was shown for different columns. If a small number of characters in a field were different, asterisks (∗) were used to indicating matching characters, and only the values of the different characters were shown. For greater differences, no characters were shown, and the red *different* icon was shown. Gender was always visible in this mode to support decisions that required knowing the gender of the person without seeing the full name.

- *Masked* (*masked disclosure with markup*): This condition represents legally de-identified data, which shows no data values and fully prioritizes privacy over information disclosure. Not a single actual character is revealed, and users must rely entirely on the supplemental visual markup (icons, colored symbols, and frequency icons). Check marks again denote matching columns. Representation of differing fields is most similar to the *low* condition, except the characters that are different are represented by different symbols (& and @) rather than their actual values and values for the gender are always hidden.

These conditions allowed us to test our hypotheses about the effects of different levels of information disclosure and the influence of supplemental markup. Figure 4 shows the percentages of characters disclosed in the different conditions. The *baseline* and *full* conditions both show the values of all characters in the records, but the value hiding and character masking of the other conditions greatly reduce the amount of visible characters. Exact percentages for this figure were calculated as an average among the ten data sets used for the experiment, which is explained in the next section.

Generation of Test Data

In order to evaluate the influence of system design factors and level of disclosure on linkage accuracy, it is necessary to know whether each linkage decision is correct or not. However, for real scenarios, it is not possible to know the "true"

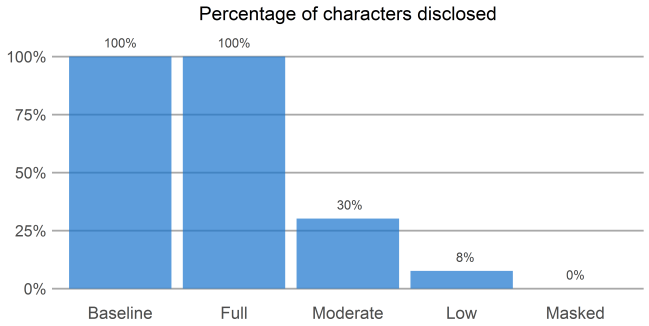

Figure 4. The different experimental conditions controlled the level of information disclosed to participants. This bar charts shows the differences in percentage of character values revealed with the different conditions applied to the generated test data used for the experiment. Percentages are relative to the number of characters in the *baseline* condition, which shows 100% disclosure of all characters. The *moderate*, *low*, and *masked* conditions hide matching characters and use character masks to greatly reduce the amount of visible characters.

answer. Thus, it was necessary to create a test set of records of personal information to use as "ground truth" for the study

We generated realistic pairs of data based on publicly available voter registry data from a large county in the United States from two time periods (four years apart). We used the registry number and street address information to build the ground truth of whether record pairs corresponded to the same or different people, and we modified some of the data to control for types of differences and data errors that would be available to participants. By controlling this process, we knew whether each generated pair of records referred to the same person or to different people.

The perturbed data was based on a variety of common scenarios that make linkage difficult, such as name changes, typographical errors, family relationships, and name swaps. We also generated pairs with missing information for both the "same" and "different" personal record pairs. The test data had a total of 747 record pairs with known "same" or "different" classifications and labels for the type of linkage scenario. For the specific record pairs the participants would see in the experiment, we sampled from the test data to create 10 different sets, where each set had 36 record pairs with the same composition of scenario labels. Of the 36 pairs in each set, 6 were chosen as "easy" pairs to be used to check whether participants sufficiently understood the decision-making task and were putting forth reasonable effort during the study. For example, a pair with two records where they shared all attributes most likely refers to the same person, and two records having no common attributes most likely refers to different people.

Procedure

The study was approved by our organization's Institutional Review Board (IRB). All study sessions took place as group sessions in a computer lab, where all participants conducted the study on identical computers running Windows 7 with 23-inch displays at 1920x1080 resolution. The experiment's application was implemented as a web app, and all participants completed the study using Google Chrome.

At the start of the study, participants were given an overview and asked to provide informed consent before participation. Participants then completed a background questionnaire asking for age, gender, level of education, academic specialization, experience with data analysis, and primary language.

Next, a member of the research team presented an overview of the record linkage task and explained the interface. After the introductory instructions, the application guided participants through additional instructions and practice questions, which took most participants approximately 15–25 minutes to complete. This section walked participants through examples related to specific factors (e.g., family relationships, name frequency, missing ID numbers) that can make linkage decisions more difficult. The application provided feedback for practice questions that were answered incorrectly.

Following the practice phase, participants immediately moved on to the main assessment phase, which included a set of 36 linkage questions. After the main phase of questions, participants continued to answer additional questions until near the end of the study session, and then concluded by completing a closing questionnaire that asked for comments about the linkage task and interface elements. The entire study session lasted 90 minutes.

Participants

The study had a total of 104 participants. Each participant completed one condition, and participant numbers were distributed as follows: 20 in *baseline*, 20 in *full*, 23 in *moderate*, 21 in *low*, and 20 in *masked* (the minor variations in numbers across conditions are due to the experimental design using group sessions in a computer lab). There were 61 males and 42 females, and one participant did not specify gender. Ages ranged from 18 to 43 years, and the median age was 24 years. About 65% of the participants were from the United States and had English as their native language. The participants came from a variety of academic disciplines. Recruitment was done by university-wide emails asking interested participants to select their availability from the given set of scheduled study times. When scheduling, we distributed participants across conditions in an effort to balance the level of education and academic discipline among groups. About 57% of the participants were either pursuing or already had a graduate degree, and the remaining participants were undergraduate university students. All participants received a $15 gift card for compensation, and an additional $35 incentive award was offered for the best performers to encourage participant engagement and effort.

RESULTS

We present an overview of the study results, and we analyzed the results to test for differences based on the previously explained hypotheses. Hypotheses 1 and 2 are concerned with the amount of information disclosure, so we compared outcomes from the *full*, *moderate*, *low*, and *masked* conditions to address these hypotheses. Hypothesis 3 is concerned with differences between the baseline and the addition of supplemental markup, so we compared outcomes from the *baseline* and *full* conditions to test this hypothesis. We did not

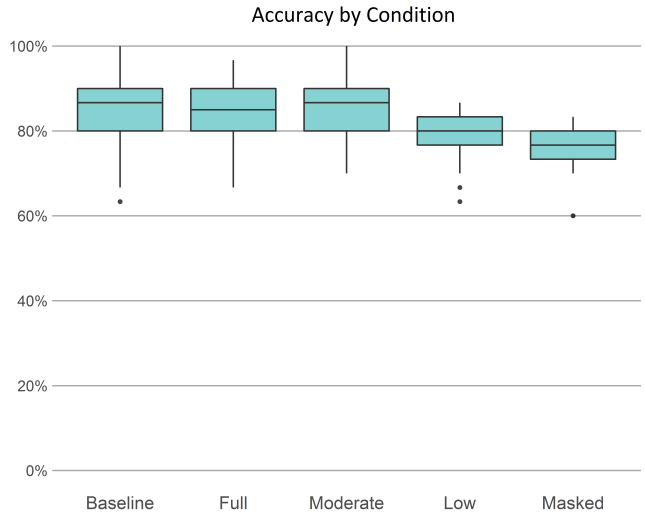

Figure 5. Record linkage accuracy for the five conditions.

conduct statistical comparisons of all five conditions together because this would confound the presence/absence of supplemental markup and level of disclosure.

Performance Overview

As previously mentioned, participants completed 36 questions, which included six additional easier questions to verify that participants understood the task and that they were paying attention throughout the study. From pilot testing, we set a performance requirement such that any participants who incorrectly answered more than one of the easy questions would not have their data included for analysis. All participants met the acceptance threshold. These easy questions were not included in the analysis of study results.

We use the percentage of correct responses to report the accuracy of linkage decisions. Across all conditions, accuracy ranged from 60% to 100%, with an overall mean of 81.28% (SD = 8.57%). Figure 5 shows the accuracy results broken down by condition. We present quantitative results graphically with standard box-and-whisker plots where the box represents the interquartile range (IQR) with a horizontal line for the median. Each whisker extends to the most extreme value falling within an additional half-IQR beyond the IQR (in both directions). Dots represent outlier values falling outside the range of the whiskers.

We also consider completion time, which includes only the portions of the study spent answering the main questions. Mean completion time was 11.07 minutes (SD = 5.65).

Along with linkage performance metrics (time and accuracy), we also analyzed differences in participant confidence in linkage decisions based on the confidence indicators for each response. We quantified confidence responses by encoding low, moderate, and high responses with values of 1, 2, and 3 (respectively). Overall, participants were significantly more likely to indicate low confidence (responses of 1) for questions they answered incorrectly compared to correct decisions, with significance indicated by chi-squared test results of $\chi^2(1) = 136.23$ and $p < 0.001$. This effect is seen easily in Figure 6, which shows the distribution of confidence outcomes broken down by questions that were answered correctly and incorrectly. This result suggests that when participants were wrong, they were more likely to suspect they might be wrong—so they were right to lack confidence.

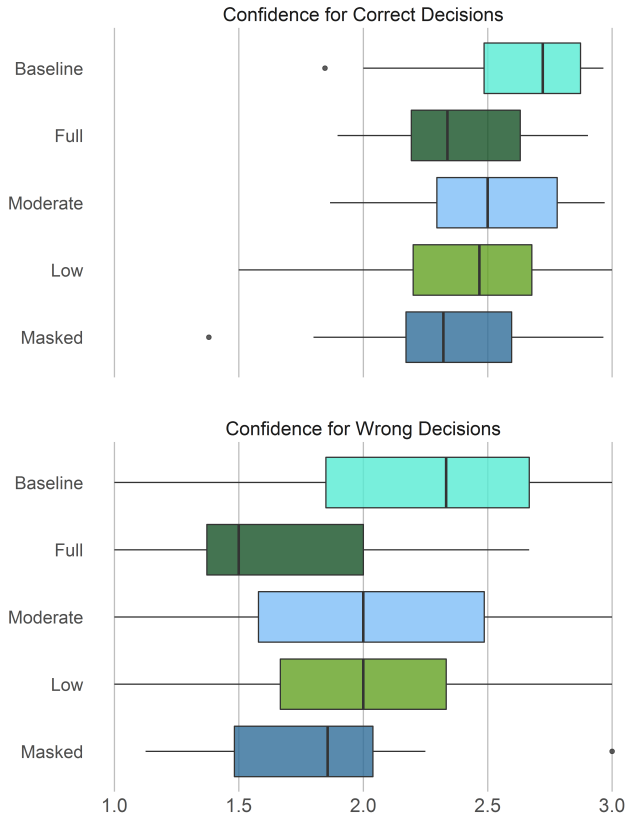

Figure 6. Confidence results for the five conditions separated by correct and incorrect responses. Confidence was significantly lower for incorrect decisions.

Effects of Level of Information Disclosure

We compared the results from the *full*, *moderate*, *low*, and *masked* conditions to study the effects of varying level of information disclosure and data hiding. The accuracy results did not meet the assumptions for parametric testing and we were unable to correct with transformations. We therefore tested for differences in accuracy due to level of disclosure using a non-parametric Kruskal-Wallis test. The test found a significant main effect with $\chi^2(3) = 23.31$ and $p < 0.001$. A posthoc Dunn test showed accuracy in the *low* and *masked* conditions were significantly worse than the *full* and *moderate* conditions ($p < 0.05$). The posthoc test showed a near-significant difference ($p = 0.087$) between *low* and *masked*.

We also tested for differences in completion time across the levels of information disclosure. The raw completion times did not meet the assumptions for parametric testing; the time

data were transformed with the log function to address normality. A one-way independent ANOVA found no evidence of different times across disclosure conditions, with test results of $F(3, 80) = 0.43$ and $p = 0.733$. Thus, the study did not find evidence of different levels of disclosure affecting the amount of time taken to complete the record linkage activity.

In addition to the performance measures, we tested for confidence differences among the four levels of information disclosure. A Kruskal-Wallis test failed to detect a difference with $\chi^2(3) = 3.42$ and $p = 0.331$. Figure 6 shows average confidence results across conditions separated by correct and incorrect decisions.

Effects of Supplemental Markup

To study the effects of the supplemental visual markup, we tested for differences between the *baseline* and *full* conditions, which both had full data disclosure with no data hiding. The accuracy results were similar between the two (see Figure 5). Since the accuracy results met the assumptions for parametric testing, we tested for the effects of supplemental visual markup with a Student's t-test. No significant difference was found with $t(33.47) = 0.05$ and $p = 0.96$.

For completion times, we again applied a log transformation to adjust normality for parametric testing. A t-test found no evidence of an effect of markup, with $t(32.87) = -0.56$ and $p = 0.58$. Thus, with no detected effects on time or accuracy, we reject the hypothesis that the additional supplemental markup improved linkage performance.

We also tested for differences in confidence of responses based on the presence or absence of the supplemental markup. Since confidence indicators were ordinal, we analyzed confidence results with a non-parametric Kruskal-Wallis test. The results yielded $\chi^2(1) = 7.99$ and $p = 0.005$, showing the addition of supplemental visual markup significantly reduced confidence. This difference can be seen between the top two rows in each of the plots in Figure 6.

This was initially confusing because the markup provides additional information that is intended to make it easier to identify differences. However, the markup also provides name frequency information, thus increasing the amount of information available for decision making and encouraging consideration for more factors. We suspect it was the addition of frequency information—rather than the visual icons and highlighting—that caused the accuracy effect. Because our supplemental markup combined both name frequency and the visual discrepancy highlighting, our experimental design is not able to separate these elements for quantitative analysis.

From the post-study questionnaires, we do know that most participants found the frequency information valuable. Participants were directly asked whether they considered name frequency when making decisions ("Did you take into account rarity/commonness of the first name and last names when you made linkage decisions?") and to explain why or why not. From the results, 94% of participants in the conditions with supplemental markup responded that they did consider frequency for their judgments. Thus, it is clear that frequency information is important for making decisions, even if it increases uncertainty.

DISCUSSION

Our research is based on the need to protect privacy in situations where utility of PII data for legitimate purposes typically requires full access for approved data workers, developers, and decision makers. We studied the balance between privacy and data utility when limiting the amount of disclosed PII in record linkage, a type of data verification task that is essential to ensure accurate and high-quality data sources for biomedical, social science, and economic research [25].

The results of the experiment show that it is possible to greatly reduce the availability of data details without noticeably affecting decision quality. As shown in Figure 4, the *moderate* condition showed only 30% of all characters in the data records, yet the linkage performance was similar to that of the *full* condition, which had all characters visible. The 70% reduction in data disclosure certainly improves privacy by greatly limiting access to details in PII, and with the supplemental markup, the experiment did not detect any negative impact to data utility. This result is promising for the potential to use value hiding and masking to reduce PII access through a method that did not significantly interfere with human interpretation and judgment.

However, it is important to understand that complete anonymization can reduce data utility for tasks that depend on PII information for data work and decision making. Some methods of privacy preservation aim to optimize privacy by entirely hiding data details, but our experiment contributes strong evidence that restricting details can be detrimental for certain types of data work that require access to personal information. While considering the different levels of detailed disclosure, the results show strong evidence of a significant accuracy reduction in *low* and even a greater reduction in *masked* (see Figure 5). As such, the study suggests that complete anonymization may not be a sufficient solution to privacy for data purposes for which data veracity and accurate decision making are the highest priorities.

On the other hand, although average linkage accuracy in the *masked* condition was significantly reduced, participants were still able to make many correct decisions with the given interface. Participants in the *masked* condition had 74.5% accuracy, as compared to the 84.1% average accuracy with *full* access, and 74.5% will still result in a large number of correct decisions that otherwise would likely be missed. This is especially important for situations where legal requirements only allow for de-identified data access. The study results show that although quality of decisions do suffer with 100% data masking, appropriate interface design can make it possible to improve data utility and judgments even for legally de-identified data that are fully masked.

When taking approaches to increase privacy through partial reduction of data visibility, care must be taken when choosing which properties and values to hide or reveal. This assertion is supported by the difference in results between the *moderate* and *low* conditions. While both of these conditions sub-

stantially reduce the percentage of characters disclosed (see Figure 4), the *low* condition saw a drop in linkage accuracy compared to *full*, while the *moderate* condition did not. This difference highlights the tradeoffs between privacy and data utility when limiting information disclosure. Among our conditions, *moderate* seems to provide an appropriate balance between these factors for our record-linkage scenario, but the threshold for data hiding without sacrificing utility will depend on the specific nature of the data and the needs of the data inspection or decision-making task.

We also hypothesize that an appropriate interface is essential for visually representing meta-data for specific data tasks when limiting disclosure. For our case, we posit that the addition of the supplemental markup was necessary to understand the discrepancies in the conditions with varying levels of disclosure. We suspect the effects of the different disclosure levels would not persist without the markup. The visual markup was used to explain differences between records in such a way that full review of the PII was not necessary (at least for the *moderate* case). Though the design of appropriate visual representations will depend on the specifics of the dataset and intended tasks, our study also serves as an example to demonstrate that it is possible.

The results of the study also demonstrate that providing supplemental metadata can influence how human review and decision making occurs, as evidenced by the significantly reduced confidence when supplemental markup was provided to the baseline case where the records were fully disclosed. For record-linkage decision making, it is important to consider whether people's name are rare or common in order to assess the probability of duplicate entities. Participant feedback from the study suggests the provided name-frequency information was helpful for linkage decisions, though the additional information reduced participants' confidence in their decisions. In general, increased awareness of uncertainty is a positive outcome for data analysis and verification tasks, and the effect on confidence might be indicative of more careful thinking about probability and real-world scenarios involving name changes and data discrepancies.

We posit that other data tasks involving sensitive or personal information could also take advantage of supplemental data descriptors and metadata to help raise awareness of information necessary for decision making when taking steps to hide details to preserve privacy. As a concept, the visual markup approach is similar to other visual methods to provide relevant summary information about the dataset or information about how an entity relates to others in the dataset (e.g., [15, 32]). However, because design specifics will depend on the data analysis or decision tasks, it is valuable to demonstrate and communicate successful case studies in designing data descriptors and privacy-preserving interfaces in specific domains. While the record linkage scenario is one such domain, opportunities remain for future work to consider similar methods in other areas.

Besides record linkage (i.e., integrating multiple databases without a common identifier), the visual masking approach can be applied to a variety of common data cleaning tasks for data scenarios involving PII. Data cleaning involves tasks such as anomaly detection, deduplication of a single database, missing data imputation, and data standardization. All of these tasks ultimately require human judgment that would require inspection of personal information that opens up high risk for privacy.

Also, to further generalize our findings to other tasks, one goal for future research is to better understand how to visually summarize probability and frequency information about record values in an understandable way, as this may be one of the most important pieces of meta-data needed for other decision-making such as threshold determination in anomaly detection and data standardization. Our study included simple representations for frequency thresholds relevant for data linkage based on consultation with record linkage experts, but to apply the visual masking approach to other scenarios, it will be important to design methods for summarizing different types of probability and frequency information.

CONCLUSION

For legitimate data work such as data integration and verification using PII data, different people need to have access to personal information, which sacrifices the personal privacy of those whose data is stored. Often, the primary methods for handling privacy concerns are either to restrict data access at the expense of data utility, or to open the data to more people to improve throughput and utility at the expense of reduced privacy. We study the use of data hiding and visual masking as a means of limiting the amount of PII available for human review while providing supplemental markup to help communicate essential properties needed for effective decision making. In a controlled experiment, we found evidence of tradeoffs between data restriction and decision quality. The results demonstrate that extreme limits to data disclosure can significantly reduce the quality of decision making. However, when legal requirements only allow for de-identified data access, use of an appropriate interface can significantly improve data utility, as participants achieved 74.5% accuracy with fully-masked data compared to 84.1% with unrestricted data access in the *full* condition.

Moreover, the results demonstrate that it is possible to significantly reduce PII disclosure without noticeably affecting decision accuracy. Through the use of visual indicators of metadata and data discrepancies, participants who made data decisions while viewing only 30% of PII content had average decision quality similar to those who had full 100% access to the data. The findings of this work are important for understanding how to design privacy-preserving data systems for data workers.

ACKNOWLEDGMENTS

This work was funded in part by Patient Centered Outcomes Research Institute (PCORI) methods program contract ME-1602-34486. We also thank Ashok Krishnamurthy and Mary Whitton for their constructive feedback on earlier iterations of the design.

REFERENCES

1. Martha Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. 2017. *How Well Do Automated Linking Methods Perform in Historical Samples? Evidence from New Ground Truth*. Technical Report.
2. Francis P Boscoe, Deborah Schrag, Kun Chen, Patrick J Roohan, and Maria J Schymura. 2011. Building capacity to assess cancer care in the Medicaid population in New York State. *Health services research* 46, 3 (2011), 805–820.
3. Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. 2017. How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3645–3656.
4. Cathy J Bradley, Charles W Given, Zhehui Luo, Caralee Roberts, Glenn Copeland, and Beth A Virnig. 2007. Medicaid, Medicare, and the Michigan Tumor Registry: a linkage strategy. *Medical Decision Making* 27, 4 (2007), 352–363.
5. Janet M Bronstein, Charles T Lomatsch, David Fletcher, Terri Wooten, Tsai Mei Lin, Richard Nugent, and Curtis L Lowery. 2009. Issues and biases in matching medicaid pregnancy episodes to vital records data: the Arkansas experience. *Maternal and child health journal* 13, 2 (2009), 250–259.
6. Kelly Caine and Rima Hanania. 2012. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association* 20, 1 (2012), 7–15.
7. Kelly E Caine, Marita O'Brien, Sung Park, Wendy A Rogers, Arthur D Fisk, Koert Van Ittersum, Muge Capar, and Leonard J Parsons. 2006. Understanding acceptance of high technology products: 50 years of research. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. SAGE Publications Sage CA: Los Angeles, CA, 2148–2152.
8. Daphne Chang, Erin L Krupka, Eytan Adar, and Alessandro Acquisti. 2016. Engineering Information Disclosure: Norm Shaping Designs. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 587–597.
9. Jia-Kai Chou, Yang Wang, and Kwan-Liu Ma. 2016. Privacy preserving event sequence data visualization using a Sankey diagram-like representation. In *SIGGRAPH ASIA Symposium on Visualization*. ACM.
10. Serdar Çiftçi, Pavel Korshunov, Ahmet Oguz Akyuz, and Touradj Ebrahimi. 2015. Using false colors to protect visual privacy of sensitive content. In *Human Vision And Electronic Imaging Xx*, Vol. 9394. Spie-Int Soc Optical Engineering, 93941L.
11. Federal Trade Commission and others. 2008. Innovations in health care delivery. (2008).
12. Gordon Darroch. 2002. Semi-Automated Record Linkage with Surname Samples: a Regional Study of Case LawLinkage, Ontario 1861–1871. *History and Computing* 14, 1-2 (2002), 153–183.
13. Aritra Dasgupta, Min Chen, and Robert Kosara. 2013. Measuring Privacy and Utility in Privacy-Preserving Visualization. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 35–47.
14. Aritra Dasgupta and Robert Kosara. 2011. Adaptive privacy-preserving visualization using parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2241–2248.
15. Aritra Dasgupta, Eamonn Maguire, Alfie Abdul-Rahman, and Min Chen. 2014. Opportunities and challenges for privacy-preserving visualization of electronic health record data. In *Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records*.
16. Fan Du, Catherine Plaisant, Neil Spring, and Ben Shneiderman. 2017. Finding similar people to guide life choices: Challenge, design, and evaluation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5498–5544.
17. Stephen E Fienberg. 2005. Confidentiality and disclosure limitation. *Encyclopedia of Social Measurement* 1 (2005), 463–69.
18. Daniel J Gilman and James C Cooper. 2009. There is a Time to Keep Silent and a Time to Speak, The Hard Part is Knowing Which is Which: Striking the Balance Between Privacy Protection and the Flow of Health Care Information. (2009).
19. Rob Hall and Stephen E Fienberg. 2010. Privacy-Preserving Record Linkage.. In *Privacy in statistical databases*, Vol. 6344. Springer, 269–283.
20. Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4 (2011), 271–288.
21. Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 547–554.
22. Hyunmo Kang, Lise Getoor, Ben Shneiderman, Mustafa Bilgic, and Louis Licamele. 2008. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. *IEEE transactions on visualization and computer graphics* 14, 5 (2008), 999–1014.
23. Hanna Köpcke, Andreas Thor, and Erhard Rahm. 2010. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 484–493.

24. Hye-Chung Kum, Stanley Ahalt, and Darshana Pathak. 2013. Privacy-preserving data integration using decoupled data. In *Security and Privacy in Social Networks*. Springer, 225–253.

25. Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, and Stanley C Ahalt. 2014a. Social genome: Putting big data to work for population informatics. *Computer* 47, 1 (2014), 56–63.

26. Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, Michael K Reiter, and Stanley Ahalt. 2014b. Privacy preserving interactive record linkage (PPIRL). *Journal of the American Medical Informatics Association* 21, 2 (2014), 212–220.

27. Pin Luarn and Hsin-Hui Lin. 2005. Toward an understanding of the behavioral intention to use mobile banking. *Computers in human behavior* 21, 6 (2005), 873–891.

28. National Cancer Institute NIH. 2017. SEER Research Data Use Agreement – Surveillance, Epidemiology and End Results Program. (2017).

29. E.C. O'Brien, A.M. Rodriguez, H.-C. Kum, L. Schanberg, S.M. O'Brien, and S. Setoguchi. 2017. Patient perspectives on the linkage of health data for clinical research: insights from a survey in the United States. Presentation abstract at the 2017 World Congress of Epidemiology. (2017).

30. Vaishali Patel, Penelope Hughes, Wesley Barker, and Lisa Moon. 2016. *Trends in Individuals Perceptions regarding Privacy and Security of Medical Records and Exchange of Health Information: 2012-2014*. Technical Report. ONC Data Brief, no.33. Office of the National Coordinator for Health Information Technology: Washington DC.

31. George G Robertson, Mary P Czerwinski, and John E Churchill. 2005. Visualization of mappings between schemas. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 431–439.

32. Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2017. A systematic view on data descriptors for the visual analysis of tabular data. *Information Visualization* 16, 3 (2017), 232–256.

33. Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu, and Weiwei Cui. 2017. NameClarifier: a visual analytics system for author name disambiguation. *IEEE transactions on visualization and computer graphics* 23, 1 (2017), 141–150.

34. Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38, 6 (2013), 946–969.

35. Joan L Warren, Carrie N Klabunde, Deborah Schrag, Peter B Bach, and Gerald F Riley. 2002. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical care* 40, 8 (2002), IV–3.

36. Daniel J Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Gerald Jay Sussman. 2008. Information accountability. *Commun. ACM* 51, 6 (2008), 82–87.