

## Privacy beyond anonymity : Decoupling data through encryption

Hye-chung Kum, kum@email.unc.edu  
Department of Computer Science, School of Social Work, UNC-CH

Darshana Pathak dpathak@cs.unc.edu  
Department of Computer Science, UNC-CH

Gautam Sanka, gausanka@cs.unc.edu  
Department of Computer Science, UNC-CH

Stanley Ahalt, ahalt@renci.org  
RENCI, Department of Computer Science, UNC-CH

**Abstract—Objective** There is a constant need for record linkage to integrate uncoordinated databases in health informatics but privacy protection is difficult to balance with the need for manual resolution of ambiguous links. We introduce a simple but powerful decoupled data system which can provide both data integration and privacy protection recognizing that the desire for privacy protection is for the sensitive data rather than the identifying data. Identity disclosure without sensitive attribute disclosure has little potential for harm. **Methods** We analyzed the potential insider threat model and present methods to reduce the risk of disclosure by controlling the ways information is displayed during the review process. We conducted a survey to evaluate the impact of chaffing, falsification, and nondisclosure of the universe on inferences people make. **Results** We confirmed that only name should be displayed during review and found that chaffing and either falsifying or not defining the universe around the data were effective in introducing uncertainty to the information disclosed during review. When the universe around the data was not defined, 56% of the participants were uncertain about the identity given a common name. Even for rare names, if the list is chaffed and the universe is not defined, 66% of the participants were uncertain on the identity. **Conclusion** When chaffing is used in combination with nondisclosure of the universe, even rare names can be displayed with minimum risk of attribute disclosure during clerical review. Our results show that these methods are effective in missing and erroneous data as well.

**Keywords-** *privacy preseriving data integration, privacy preseriving record linkage, identity inference, decoupled data, chaffing, clerical review*

### I. INTRODUCTION

With the tremendous advancements in technology during last few decades, information systems in the public health sector have undergone significant infrastructure changes. Computerized databases and internet-based data collection methods have made it possible to collect, store, and process huge amounts of data. However, data often ends up in heterogeneous and uncoordinated systems. The information derived from these systems is often redundant, fragmented over multiple databases and sometimes incomplete [1, 2]. This introduces the need for record linkage – the process of identifying record pairs which belong to the same real-world entity.

The record linkage process is complicated by the inherent factors observed in the real-world data, such as missing data (Missing SSN), erroneous data (transpose of DOB), non-standardized form of data (prefixes as Mr. Ms.), and change in the data over time (changed last name). Absence of common, error-free and unique identifier makes exact matching solutions inadequate leading to methods for approximate record linkage to deal with these issues [3, 4, 5]. In a case study of linking cancer registries from two institutions, 10% more matches were found using a simple deterministic approximate match compared to the exact match methods due to typos in names or missing SSNs [5]. A more sophisticated approximate match method, six pass probabilistic record linkage, to link NY state cancer registry with Medicaid data reported that only 83% of the matches could be identified with exact match [6]. In these approximate

methods, real world entities such as twins and family, which share similar identifying information, make it very difficult to identify false positives automatically [7]. All such methods require careful management of errors introduced during the linkage process and manual resolution of ambiguous matches. Both studies above carried out clerical reviews to identify false positives on ambiguous matches. Bosce reported that most of the false positives were spouses, relative, or twins [6].

Clerical review of identifying information required for approximate record linkage seems to be in direct conflict with the privacy of the subjects of the data. To be precise, it is in direct conflict with identity disclosure of the individuals. Although attribute disclosure occurs mostly through identity disclosure, it is important to distinguish between the two because identity disclosure *without* sensitive attribute disclosure has low potential for harm [8]. If we move beyond anonymity as privacy protection, and recognize that the desire for privacy protection is for sensitive data rather than identifying data, there is a relatively simple solution to privacy preserving data integration.

Decoupling the identifying information (PII) from the sensitive data can give a practical solution to this seemingly difficult problem (figure 1). Decoupling data follows the minimum necessary standard for privacy protection and removes the unnecessary information, the connection information from PII to sensitive data, during the record linkage process. The innovation in decoupling data is the focus on revealing information rather than hiding it. The key is to understand the minimum information required for acceptable linkage, and then to design protocols to reveal, in a secure manner, only that information.

The main contribution of the paper is to introduce the simple but powerful decoupled data system for privacy preserving record linkage, describe the different techniques used to reduce risk of sensitive attribute disclosure during the clerical review, and present an evaluation of the techniques under a variety of situations, including the effectiveness of these methods in the presence of missing and erroneous data. The evaluation reveals how best to minimize the risk by controlling what information should be displayed and how it should be displayed during clerical review.

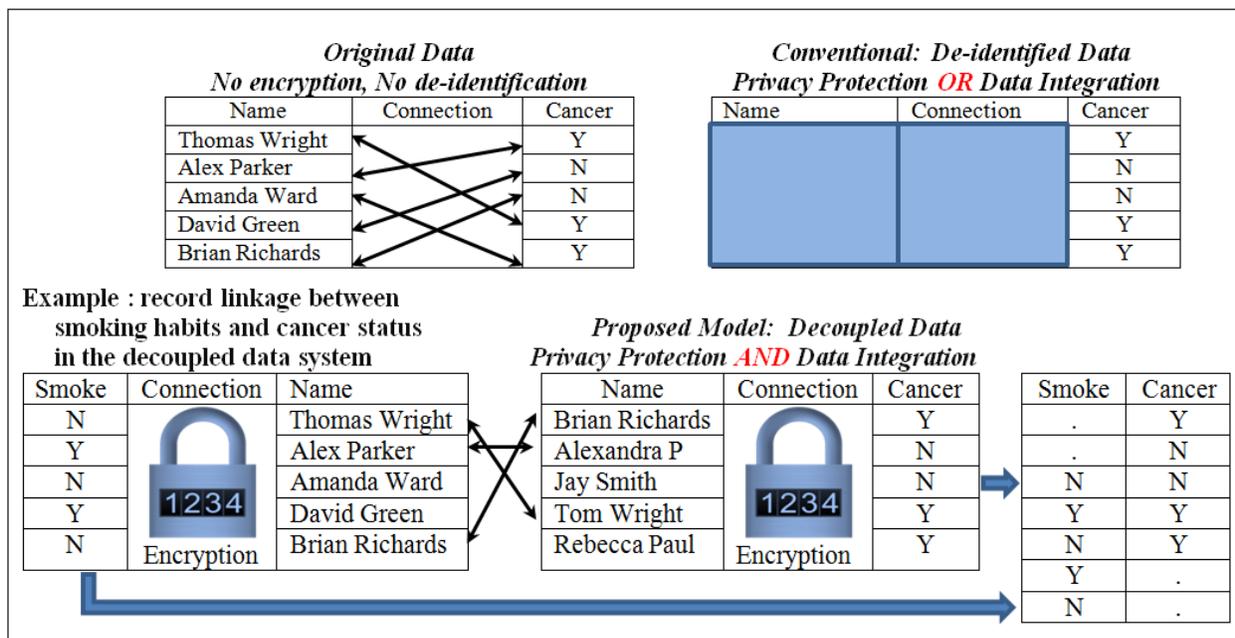


Figure 1. Three modes of information sharing in data, with an example for decoupled systems

## II. METHODS TO CONTROL INFORMATION EFFECTIVELY

The decoupled data system allows researchers to integrate data using isolated identifying information which provides both error management and privacy protection [7]. In essence, with proper protocols in place, all de-identified data could be released as decoupled data to these secure systems, which enable flexible record linkage. With so much information being on the system including PII, the

system should be maintained as a restricted access system where all activities on and off the computer are restricted and researchers have little control to manipulate or view the data. They simply specify the data they want to integrate using the metadata and the method to integrate it and the bulk of the work is done by the decoupled data integration software. As the software works through the approximate record linkage specifications, it will interact with the researchers to resolve the ambiguous regions by asking them to manually review some PII and to determine which are true positives versus false positives (Figure 2). A decoupled data information system is essentially a computerized third party that strictly controls information.

A. *Shuffling and Encrypting Connection Information for Decoupling.*

The first step to building a computerized third party is to decouple the PII information from the sensitive information to block attribute disclosure. The decoupling occurs in three steps. First the full table is split into two tables, one for PII and another for the remaining information, mostly sensitive data, denoted as  $T = T_{(PII)} + T_{(SD)}$ . Second, the rows in the PII table are shuffled randomly so that the row association between  $T_{(PII)}$  to the  $T_{(SD)}$  cannot be easily derived. Finally, asymmetric encryption is used to lock the row association information from the  $T_{(PII)}$  to the  $T_{(SD)}$  [9]. In the full decoupled information system, which has multiple decoupled tables, the table association information of the  $T_{(PII)}$  to the  $T_{(SD)}$  is also locked using asymmetric encryption. Thus, only those who have both the private key to decrypt the table association and the private key to decrypt the row association for a particular table can access the connection information between the PII and the sensitive data [9]. More details of the encryption scheme are given in the appendix.

B. *Display Control During Clerical Review*

What information is displayed during clerical review has much impact on the risk of disclosure during record linkage. Figure 2 depicts what a typical screen might display during the review process with only the name being revealed as untouched data. When possible, we recommend displaying the difference between the attributes that are meaningful for record linkage instead of the raw data. For example, the gender field would only indicate, same, different, or missing in one or both fields. Id numbers such as SSN would never be revealed to the researchers. Instead, the difference between two SSNs can be conveyed to the reviewer as number of different digits and transposes. A discussion of our decision to reveal names but not DOB is provided in the appendix.

C. *Chaffing and Universe Manipulation*

With strict decoupling, the researchers will not be able to associate a particular row of data with any PII viewed during clerical review. But researchers can combine what they know with the PII data shown during review to make inferences and learn sensitive information about people they know. This can lead to attribute disclosure via group membership [8]. The threat model is described formally in Figure 3. Thus, strict decoupling via encryption is not sufficient to protect privacy when identities could be revealed during clerical review. We employ additional methods to interfere with possible inferences by manipulating the universe around the data that is displayed during review.

The probability of attribute disclosure through group membership is dependent on a variety of factors including any pre-existing information that is known by the observer, the knowledge on the nature of the list, and the uniqueness of the PII in the universe of the data. For example, in our threat model a researcher who memorized the name for the target subject could spot the same name during clerical review. It is important to note that, identifying the person is not a privacy violation. Rather, the violation occurs when an identified person's disease status becomes known. Whether spotting the same name during clerical review can lead to confirmation that the real person of interest ( $I_{an_{Bob}}$ ) has cancer depends on two factors. First does the list represent everyone with cancer? And second, how likely is it that the viewed name ( $I_{an_{PII}}$ ) represents the actual real world entity ( $I_{an_{Bob}}$ ), which depends on factors such as the rarity of the name. The goal of the methods in this section is to further introduce uncertainty in inferring that  $I_{an_{Bob}}$  has cancer by intentionally manipulating the universe around the displayed data point ( $I_{an_{PII}}$ ) with little impact on the matching decision. There are three methods of modification: (1)

chaffing: literally change the nature of the universe by adding fake data, (2) fabrication: change the label/name of the universe presented to the researcher, and (3) nondisclosure: hide the identity of the universe from the researcher to reduce confidence (Figure 4).

Chaffing is the process of adding fake data to a dataset to enlarge the universe to such an extent where group membership no longer reveals sensitive information. It is comparable to a person concealing themselves by becoming part of a large crowd. In particular, by adding a certain type of fake data, we can fundamentally block attribute disclosure through group membership. Disclosure through group membership can only occur when the list displayed for review represents a homogenous group, such as cancer registry. By adding real names to the list that do not have cancer, and letting the researcher know that the chaffing has occurred, the list is no longer homogenous and membership on the list has no meaning. In sum, the researcher can no longer be certain of sensitive information based on group membership even if the identity has been disclosed. The key to chaffing is to add the appropriate fake data so as to introduce uncertainty to attribute disclosure, but not to interfere with the matching decision. We can do so by adding rows of real people from the same region as the target population effectively changing the nature of the list being viewed. More research is needed as to how much fake data is required, and how distinct the fake data needs to be from the real data as to not confuse the matching process.

File	FirstName	LastName	DOB	SSN	Gender	Address
LA	John	Gray	07/13/1978	986-65-3210	F	512, Academy St. - 94525
LB	Jon	Grey	07/23/1968	987-65-4210		
LA	Alex	Parker II	05/02/1977	111-22-3333	M	23, N Fort Dr. - 12345
LB	Alex	Parker	02/05/1977	111-22-3333	M	23, Fort Drive - 12345
LA	Donna	Balmer	11/27/1981	777-66-2134	F	43, Westem Pky - 98765
LB	Ms. Donna	Palmer	11/27/1981	777-66-1234	F	
...	....	....	....	....	....	....

Table A: Raw data from files LA and LB. The researcher doing clerical review will not be able to access or read raw data. Instead, a subset of the PII attributes used for record linkage will be converted into codes and displayed during clerical review as shown below. Except for names, the information will be converted into coded characters for same(\_), different(D), transposed(TX) and missing(M) fields as discussed below.

**D=Different, M=Missing (either or both records), TX=Transpose, \_=Same** 

**Type Y in the right most column if you would like to link the two records.**

Rec. No.	First Name	Last Name	DOB (mm/dd/yyyy)	SSN	Gender	Link (Y/N)
111	John	Gray	--D/--D-	--D--D---	M	<input type="checkbox"/> N
	Jon	Grey				
112	Alex	Parker II	T/X/----	-----	-	<input type="checkbox"/> Y
	Alex	Parker				
113	Donna	Balmer	--/------	-----TX-	-	<input type="checkbox"/> Y
	Ms. Donna	Palmer				
114	Timothy	Richards	--/TX/----	M	-	<input type="checkbox"/>
	Tom	Richards				
115	Anita	Gorge	--/--TX	--D--D---	D	<input type="checkbox"/>
	Anita	George				
116	Michael	Smith	--/------	-----	M	<input type="checkbox"/>
	Michael	S				

Table B: Encoded data displayed during the clerical review process. Here, the researcher has to make a decision about the matching of the records based on encoded fields. Researchers put 'Y' for records belonging to the same person and 'N' otherwise. For DOB, T/X/---- means the values for month and day are transposed [as in 5/2/1977 and 2/5/1977]. For SSN, --TX---- means two digits are transposed [as in 123-45-6789 and 123-54-6789]. The researcher is now making a decision about record 114.

Figure 2. Clerical Review

1. Alice is a researcher who will be responsible for resolving ambiguous record linkage regions via clerical review for a study on linking Cancer Registry data from two hospitals,  $L_A$  and  $L_B$ , located in NC. During the process, she will be shown a partial list of PII's from  $L_A$  and  $L_B$  to resolve the ambiguous regions, denoted as  $L_{A(PII)}$  (Alice) and  $L_{B(PII)}$  (Alice). The expected size of the partial list is a tiny fraction of the full lists.
2. The tables are decoupled as  $L_A = L_{A(PII)} + L_{A(SD)}$ ,  $L_B = L_{B(PII)} + L_{B(SD)}$ , where  $L_{A(PII)}$  is the PII from the hospital records in  $L_A$  and  $L_{A(SD)}$  is all columns except PII including all the sensitive data. Alice has no access to  $L_{A(SD)}$  and  $L_{B(SD)}$  which specify the type of cancer diagnosis in the registry along with other information.
3. Bob, who works for a health insurance company, bribes Alice to find out if Ian has cancer and gives Alice Ian's full PII denoted as  $Ian_{(PII)}$ . Note that there is no guarantee that  $Ian_{(PII)}$  is unique in the universe of real world entities, denoted as  $R_A$ , from which the data is collected. We denote the unique real world entity that Bob is interested in as  $Ian_{Bob}$ .
4. We assume that Alice knows that  $L_{A(PII)}$  and  $L_{B(PII)}$  are cancer registries.
5. During her clerical review process, Alice can combine her prior knowledge of  $Ian_{PII}$  with the partial lists of PII's given to her,  $L_{A(PII)}$  (Alice) and  $L_{B(PII)}$  (Alice), to make inferences.

Under simple uniform models,

Risk (Bob finding out that  $Ian_{Bob}$  has cancer |  $Ian_{PII}$  is on either  $L_{A(PII)}$  or  $L_{B(PII)}$ )

$$= \left[ \Pr(Ian_{PII} \in L_{A(PII)} (Alice)) + \Pr(Ian_{PII} \in L_{B(PII)} (Alice)) \right] * \Pr(Ian_{PII} == Ian_{Bob})$$

$$= \left[ \frac{size(L_{A(PII)} (Alice))}{size(L_{A(PII)})} + \frac{size(L_{B(PII)} (Alice))}{size(L_{B(PII)})} \right] * \left[ \frac{1}{Number\ of\ Ian_{PII}\ in\ R_A} \right]$$

= very small as clerical review should occur in only a tiny percent of the full lists

Thus the main threat of inferring that Ian has cancer results from

- having seen his PII in one of the partial lists
- knowing that  $L_{A(PII)}$  and  $L_{B(PII)}$  are cancer registries
- AND the rarity of  $Ian_{PII}$  in the universe  $R_A$

Figure 3. Threat Model

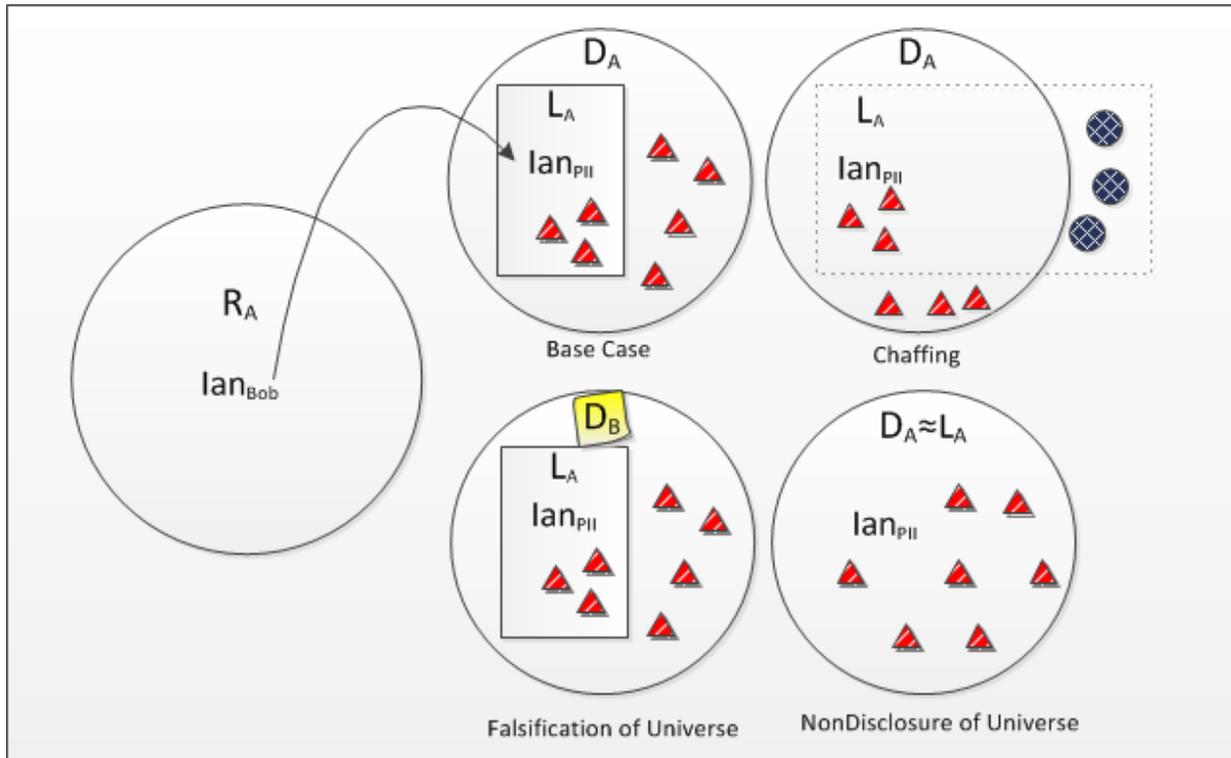


Figure 4.  $R_A$  represents the real-world entities from which the data is collected.  $D_A$  represents the universe of all cancer register patients in NC.  $L_A$  represents the subset of cancer patients at a specific hospital.  $Ian_{Bob}$  is the real-world entity represented by the data point  $Ian_{PII}$ . The triangles represent cancer patients, while the circles are those living in NC who are not cancer patients. For the purpose of falsification,  $D_B$  represents another universe of cancer patients other than those of NC but it actually contains the same data as those of NC

An orthogonal method to changing the nature of the list through chaffing is to confuse the identity inference. One method to block identity inference is to falsify the universe by presenting the list as if it came from a different data space. In our threat model, we might present the list to be from a hospital located across the country, say CA. Given that the researchers believe the information presented is accurate, such a presentation would make it probabilistically impossible for them to infer that the viewed name ( $Ian_{PII}$ ) represents the real person ( $Ian_{Bob}$ ) of interest who lives in NC. In other words, by presenting the list as if it came from a different data space ( $L_B \in D_B$ ), the  $\Pr(Ian_{PII} == Ian_{Bob}) = 0$  because  $Ian_{Bob} \in D_A$  and  $Ian_{PII} \in L_B \in D_B$  and  $D_A \cap D_B = \emptyset$ . Thus, if the researchers believes  $Ian_{PII} \in L_B$ , we effectively block any possible correct inferences on identity.

Sometimes it can be difficult to totally falsify the universe to a research team member responsible for clerical review. In such situations, not specifying the universe until after the clerical review can also be useful. When we provide the researchers with records from an undefined universe, the researcher can be led to similar uncertain conclusions about the identity of a person because they are led to assume the whole world as their universe. In order for the researchers to make a reasonably accurate inference from prior knowledge of the subject's name, they need two pieces of information: the existence of the same name on the list and the number of real people with the same name in the universe where the list came from. The probability that the spotted name on the list is the same real world entity of interest is  $\frac{1}{n(\text{same name in R})}$ . The rate of confidence is inversely correlated with the size of  $n$ . When the universe is undefined and the researcher must assume a huge universe of anyone living in the US, even for rare names it is difficult to assume  $n = 1$  with certainty. Thus, even if the researcher is able to recognize a particular name, there will be a constant degree of uncertainty in their judgment. Nondisclosure is quite similar to chaffing in that the universe becomes enlarged. This allows for an increase in probability of comparable real world entities to exist, an increase in  $n(\text{same name in R})$ , which reduces the confidence of the researcher about the identity.

### III. EVALUATION RESULTS

#### A. Experiment

To better understand what kinds of PII should be displayed to the researcher during clerical review and to evaluate the effectiveness of chaffing and manipulating the universe, we did an experiment by conducting an online survey. The survey simulated the situation that the insider, Alice, would be in while performing the clerical review. Our goal was to test how well the different methods worked to reduce identity disclosure and attribute disclosure. In this experiment, we measured (1) the effect of chaffing, (2) the impact of the modification of the universe on identity inference, (3) the disclosure risk of different identifying information attributes; namely, the common name, the common name and DOB pair, and the rare name, and (4) how these attributes interact with each other when used in combination. We also tested the effects of missing and erroneous data on the different methods using variations on the common name and DOB pair.

The basic setup was to present the respondents with certain identifying information for a target student and an honor roll that included the same identifying information. We then asked the respondents to select their confidence level about the likelihood that the information given in the question and the honor roll referred to the same person. We used a seven-point Likert scale to measure the confidence level. The respondent could choose between three levels of yes, three levels of no, or *I Don't Know*. Figure 5 summarizes all the questions and the experimental setup. We used the high school honor roll scenario because it was a neutral list with no emotional or biased assumptions.

The survey had in total 18 questions; six scenarios, each with three questions. The six scenarios were ordered so that questions would build on each other. We started with the simple common name scenario, and then moved onto to common name and DOB. The next two scenarios were common name and missing DOB followed by common name and transposed DOB. The transpose was created by swapping month and date numbers in DOB (2/5 and 5/2). The survey questions for these cases had the

<p><b>Scenario 1. Common name : George Brown</b></p> <p><b>Q1.</b> Meet George Brown, a student at Meadowgreen High School. Using only the information on this screen, how likely is it that the George Brown introduced above is the same person listed on the honor roll provided here?</p> <ul style="list-style-type: none"> <li>• Highly likely to be the same person</li> <li>• Moderately likely to be the same person</li> <li>• Slightly likely to be the same person</li> <li>• I don't know if they are the same person or not</li> <li>• Slightly likely to be two different people</li> <li>• Moderately likely to be different people</li> <li>• Highly likely to be two different people</li> </ul> <p><i>* We highlighted the name of the high school in all the questions to bring attention to the change in the school name. * We also added, removed and/or shuffled the names in the list as per questions requirement.</i></p>		<p><b>Meadowgreen HS Honor Roll</b></p> <table border="1"> <thead> <tr> <th>Name</th> </tr> </thead> <tbody> <tr><td>Amanda Ward</td></tr> <tr><td>Edward Jones</td></tr> <tr><td>Hilary Ford</td></tr> <tr><td>George Brown</td></tr> <tr><td>Susan Miller</td></tr> <tr><td>David Green</td></tr> <tr><td>Alexander Parker</td></tr> <tr><td>Brian Richards</td></tr> <tr><td>Daniel Parker</td></tr> </tbody> </table>	Name	Amanda Ward	Edward Jones	Hilary Ford	George Brown	Susan Miller	David Green	Alexander Parker	Brian Richards	Daniel Parker
Name												
Amanda Ward												
Edward Jones												
Hilary Ford												
George Brown												
Susan Miller												
David Green												
Alexander Parker												
Brian Richards												
Daniel Parker												
<p><b>Q2.</b> Repeat Q1 but falsify where the honor roll list came from by titling the list - <b>Valley Mountain High School Honor Roll.</b></p>												
<p><b>Q3.</b> Repeat Q1 above but change the honor roll list to be for an undefined school, titled - <b>A High School Honor Roll.</b></p>												
<p><b>Scenario 2. Common name and DOB: Susan Miller (date of birth 4/17/1994)</b> <i>* Modified honor roll lists to include a column for date of birth.</i></p> <p><i>* We did not allow participants to go back to change their answers from this question onwards, giving them instruction – “We let you move back to answer the previous questions again, because we thought it might take a few questions for you to fully understand what we are asking. However, from this point onward, you cannot go back to questions answered.”</i></p>												
<p><b>Scenario 3. Common name and missing DOB: Amanda Ward (date of birth 10/2/1995)</b> <i>* Modified honor roll lists to include a column for DOB, but only six names had entries for DOB. The remaining four had missing DOB. Amanda Ward's DOB was missing in the list.</i></p>												
<p><b>Scenario 4. Common name and transposed dob : Alex Parker (date of birth 5/2/1997)</b> <i>* Modified honor roll lists to include a column for DOB. This time all 10 names had valid DOB, but Alex Parker's DOB was listed as (2/5/1997). To bring attention to the transposition, we highlighted the month and day of the birthdate in the question.</i></p> <p><i>* We introduced rare name scenario and provided instructions about specific changes in question format as - “Forget everything you have seen up to this point. Now we are going to give you another section of each honor roll containing less-common names. Also, we are going to ask you a slightly different question. The scales will also change to fit the new question. Answer the question using only the information given on the paper”</i></p>												
<p><b>Scenario 5. Rare name. Rahul Ghosh.</b></p> <p><b>Q1.</b> Meet Rahul Ghosh, a student at Meadowgreen High School. Using only the information provided here, how likely is it that the Rahul Ghosh introduced above, has made the honor roll at his school?</p> <ul style="list-style-type: none"> <li>• Highly likely to have made the honor roll</li> <li>• Moderately likely to have made the honor roll</li> <li>• Slightly likely to have made the honor roll</li> <li>• I don't know if he has made the honor roll</li> <li>• Slightly likely NOT to have made the honor roll</li> <li>• Moderately likely NOT to have made the honor roll</li> <li>• Highly likely NOT to have made the honor roll</li> </ul>		<p><b>Meadowgreen HS Honor Roll</b></p> <table border="1"> <thead> <tr> <th>Name</th> </tr> </thead> <tbody> <tr><td>Amanda Ward</td></tr> <tr><td>Edward Jones</td></tr> <tr><td>Soo Chien</td></tr> <tr><td>Shaniqua Parker</td></tr> <tr><td>David Green</td></tr> <tr><td>Diego Ramirez</td></tr> <tr><td>Hilary Ford</td></tr> <tr><td>Patrick Collard</td></tr> <tr><td>Rahul Ghosh</td></tr> </tbody> </table>	Name	Amanda Ward	Edward Jones	Soo Chien	Shaniqua Parker	David Green	Diego Ramirez	Hilary Ford	Patrick Collard	Rahul Ghosh
Name												
Amanda Ward												
Edward Jones												
Soo Chien												
Shaniqua Parker												
David Green												
Diego Ramirez												
Hilary Ford												
Patrick Collard												
Rahul Ghosh												
<p><b>Scenario 6. Rare name on a chaffed list.</b> We repeated the three identical questions for scenario 5, but before giving these last three questions, we gave the following instructions -</p> <p>“We did not tell you this before, but the tables used in the previous questions contained a few pieces of false data. Those tables included students who were <u>NOT</u> actually on the honor roll. It's too late to go back and change your answers for those questions, but we'll give you another chance here. <u>Knowing that these honor rolls are not fully correct</u>, answer the questions again as to how likely is it that the Rahul Ghosh made the honor roll at his school.”</p>												

Figure 5 Basic survey question format

missing or transposed DOB in the honor roll for the target student. We left rare name and chaffing to the last two sections because we wanted the respondents to get used to inferring identity before giving them chaffed list experiment. In the fifth scenario for rare name, the target student had a rare name and the

honor rolls included other rare names such as Viswanath Sastry, Jie Lee, and Michelle Pham. In addition, we changed the question slightly to ask how likely it is that the target student had made the honor roll at his school. The scale wording was also adjusted to indicate the confidence of having made the honor roll. We changed the question slightly so that we could ask the respondents to answer the same question one more time, knowing that the honor roll included some false data of students who did not make the honor roll. This is the sixth scenario, wherein we tested a given rare name as the identifying information and a chaffed list. The exact instructions given just before the last scenario are shown in Figure 5. For the three questions in each of the six scenarios, respondents were given the honor roll from the same high school (Same HS), an honor roll from a different high school (Diff HS, falsified universe), and finally an honor roll from an unknown high school (No HS, undefined universe). We constructed the three honor rolls so that they shared some names including the target student but were sufficiently different from each other. The appendix has the basic scenario with the full three questions.

The full results are shown as 18 stacked bar charts in Figure 6. Each of the questions corresponds to one stacked bar. We tested the change in the confidence level of identity by comparing responses to different questions using the Wilcoxon signed rank test, a non-parametric T-test [10]. We mainly recruited from graduate students in various departments including public health and computer science who we anticipate will be doing the clerical reviews. We had 59 respondents. The basic demographics and data experiences of the respondents are given in our previous paper [7]. Although we made no attempt to recruit from students with experience in data analysis, we obtained a good mix of respondents with various experiences in data.

#### B. *Impact of Chaffing*

We found that by chaffing the list, the identity disclosure of rare names was reduced to similar or lower levels compared to common names. T-Test results ( $p$ -value  $< 0.005$ ) between common name scenario and rare name + chaffing scenario provide sufficient evidence to conclude that respondents had significantly less confidence about the identity of rare names in a chaffed list (Bar 1 vs. Bar 16). In other words, respondents were significantly less confident in the identity of rare names on a chaffed list compared to common names on an accurate list. The distributions of the responses given in Figure 6 clearly support this drop in the confidence level of identity. Hence, we suggest that lists revealed for clerical review should be chaffed. Most importantly, the researchers doing the review need to be well aware of the chaffing done so they can properly understand the nature of the list to prevent incorrect inferences.

#### C. *Impact of Falsifying or Undefined the Universe*

We found that manipulating the label of the displayed list was quite effective in reducing confidence in identity in all cases. In the base case for common names (George Brown from Meadowgreen HS), when a list was presented as being from the same high school, all but 11 answered *Highly or Moderately Likely the Same Person* indicating fairly high confidence in identity. There was a dramatic shift in responses to the second question when a list was presented under a different universe (Valley Mountain HS) and even went on to assert that the person on the list was a different person. Only 6 answered *Highly Likely the Same Person* while the number of respondents answering *Highly Likely Different People* shot up from 0 to 29. The median for the falsified universe is on the other end of the spectrum at *Moderately Likely Different People*. The results from the third question, the list with no high school defined, were quite different from the first two questions. The vast majority of the respondents fell in the middle of the spectrum with both the mode and median response being *I Don't Know*. The combined response to *Slightly Likely Same or Different People* and *I Don't Know* was 56%. We conclude that the respondents were confused and could not confidently assert a response on whether the presented person was the same person or not. The finding supports the hypothesis that an undefined universe will effectively introduce uncertainty such that people will not be able to make an affirmative conclusion about a name they find on the list.

Although not as strong an effect, we see a similar trend in the impact of falsified universe and undefined universe for scenarios where we presented the common name and DOB pair and a rare name.

Given a common name and DOB pair, the median response increased to *Slightly Likely the Same Person* for a different universe and *Moderately Likely the Same Person* for an undefined universe. In comparison, given a rare name, the median response was *I Don't Know* for a different universe but again *Moderately Likely the Same Person* for an undefined universe. Not surprisingly, the DOB and the rarity of a name serve to increase the confidence of the respondent compared to only a common name. Nonetheless the modification of the universe still had the impact of reducing the confidence levels as indicated by the T-tests which confirmed statistically different medians in all cases when the universe was manipulated. Most importantly, when used in combination with chaffing, the impact of manipulating the universe seems to have an additive effect (Bars 17 and 18). The median response for rare names with the universe undefined dropped further to the ideal level of *I Don't Know* when using the chaffed list (Bar 18). This is an important finding because it confirms that using a combination of chaffing and nondisclosure of the universe; we can reveal all names, including rare names, with minimum risk of attribute disclosure during the clerical review.

It is quite interesting to note that given a list from a different high school, in all questions, there were some respondents who selected either *Highly or Moderately Likely the Same Person* as the target student. The  $\Pr(\text{target student}_A \in L_B) = 0$  given that  $\text{target student}_A \in D_A$  and  $D_A \cap D_B = \emptyset$  regardless of how rare the identity might be. Logically it would be highly unlikely that a George Brown at Meadowgreen HS is the same George Brown at the Valley Mountain HS. However, 15% still answered that they were *Highly or Moderately Likely the Same Person*. For common name and DOB pair it was as high as 42% and for rare name it was 19%. In the pretest of the survey, we asked informally about such responses. Some respondents thought that the student may have transferred from one high school to the other or that the student was simultaneously enrolled in both schools. The respondents were unconsciously adding new dimensions, such as time, into the situation to accommodate their personal belief that two people with the same name and DOB are highly likely to be the same person (Susan Miller, DOB=4/17/1994). Given that Susan Miller is a fairly common name, we believe such responses strongly suggest the possibility of researchers making incorrect inferences and jumping to wrong conclusions. It could in fact be true that  $Ian_{PII} \neq Ian_{Bob}$ , but when researchers make incorrect assumptions that  $Ian_{PII} == Ian_{Bob}$ , harm can still occur. Thus, it seems that this result points to the need for good training before researchers are allowed to do clerical review.

#### D. Missing and Erroneous Data

Of all six scenarios, respondents were most confident about the identity of the target student when presented with the common name and DOB pair where 50 out of 59 respondents answered *Highly Likely the Same Person*, with 7 more respondents answering *Moderately Likely the Same Person*. These results indicate that people are highly confident about a person's identity given a pair of name and DOB. This finding is supported empirically by Weber et al. who show that the name and DOB pair was surprisingly effective for record linkage under certain circumstances [5]. Thus, we recommend that the raw DOB should not be included in the PII list during the clerical review process. Instead, information related to DOB should be displayed as differences as shown in figure 2 and discussed in the appendix.

However, this strong confidence in identity is quickly vanished in the scenarios with missing DOB and transposed DOB. In this experiment, we tested the effectiveness of the different methods in the presence of missing and erroneous data, because such data is common in real-world databases.

We found that both missing and erroneous data do not interfere with the loss of confidence, when the universe is manipulated. In case of missing DOB, respondents seemed to ignore the missing DOB all together, treating it as if the list had only names in all 3 questions. Bars 1 to 3 were not statistically different with the corresponding Bars 7 to 9. The transposed DOB added a confusion factor resulting in significant loss of confidence compared to the name only scenario (Bar 2 and Bar 10). Even when information resulting from erroneous data confused the respondents, we observed the exact same impact when the universe was manipulated. When the universe was falsified (Bar 11), even more people lost confidence in the identity of the target student compared to the correct universe (Bar 10). When the universe was undefined (Bar 12), more respondents showed loss in the confidence level, but not as much

as when the universe was falsified. The T-test confirmed that there is a statistical difference between responses of all three questions (Bars 10, 11, 12) where the universe was manipulated.

#### IV. CONCLUSION AND FUTURE WORK

Identity disclosure without sensitive attribute disclosure has little potential for harm. Recognizing that the desire for privacy protection is for the sensitive data rather than the identifying data, we introduce a simple but powerful decoupled data information system for data integration with privacy protection. We then analyze the possible insider threat model resulting from the clerical review of PII data and present ways to reduce the risk by controlling what information is displayed and how it is displayed during clerical review. We evaluated our approach by simulating the insider's point of view through a survey to determine identity disclosure and attribute disclosure in a variety of situations. We confirmed that only name should be displayed during review and found that chaffing and falsifying or not defining the universe around the data were effective in introducing uncertainty. Used in combination, we found that even rare names can be displayed with minimum risk of attribute disclosure.

It is very important that researchers are aware of the chaffing done. Thus, we recommend a short online reminder training for clerical review before the researcher starts the review for each project in the form of simple warm up exercises. The training should cover how the list has been chaffed, that a chaffed list effectively has an intractable universe, and thus the researcher cannot make any inferences with reasonable certainty. In addition, it should remind them about the prevalence of missing and erroneous data as well as the dangers of incorrectly jumping to conclusions.

#### ACKNOWLEDGMENT

We thank everyone who participated in the survey. We also thank Mike Reiter and Fred Brooks for their insightful comments, and Ian Sang-Jun Kim and Ren Bauer for their assistance with the experiment. This research was supported in part by funding from the NC Department of Health and Human Services and by NSF award no. CNS-0915364. The authors gratefully acknowledge their support..

#### REFERENCES

1. Elmagarmid K, Panagiotis GI, Verykios SV. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.* 2007;**19(1)**:1-16.
2. Sauleau EA, Paumier J, Buemi A. Medical record linkage in health information systems by approximate string matching and clustering, *BMC Health Services Research* 2007;**7**:154
3. Newcombe H, Kennedy J, Axford S, et al. Automatic linkage of vital records. *Science* 1959;**130**:954-59.
4. Fellegi P, Sunter AB. A theory for record linkage. In *JASA* 1969;**64(328)**:1183-210.
5. Weber SC, Lowe H, Das A, et al. A simple heuristic for blindfolded record linkage. *J Am Med Inform Assoc.* 2012.
6. Boscoe FP, Schrag D, Chen K, et al. Building capacity to assess cancer care in the Medicaid population in New York State. *Health Services Research* 2011;**46(3)**: 805-20.
7. Kum, H.C., Ahalt, S, Pathak, D. Privacy Preserving Data Integration Using Decoupled Data. *Security and Privacy in Social Network*, by Y. Elovici, Y. Altshuler, A. Cremers, N. Aharony, A. Pentland (Eds), Springer 2012;In print.
8. Fienberg SE. Confidentiality, privacy and disclosure limitation, *Encyclopedia of Social Measurement*, Academic Press 2005;**1**:463-9.
9. Rivest R, Shamir A, Adleman L. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM* 1978;**21(2)**:120-6.
10. Krenzke T, Hubble D. Toward Quantifying Disclosure Risk for Area-Level Tables When Public Microdata Exists. *Section on Survey Research Methods. JSM* 2009.
11. McCune M, Parno B, Perrig A, et al. Flicker: an execution infrastructure for tcb minimization. *EuroSys* 2008:315-28
12. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Jamia* 2011;**18(1)**:3-10.
13. Sweeney L. Simple Demographics Often Identify People Uniquely. *Data Privacy Working Paper 3*, 2000;1-34
14. Corder GW, Foreman DI. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*, New Jersey: Wiley 2009.

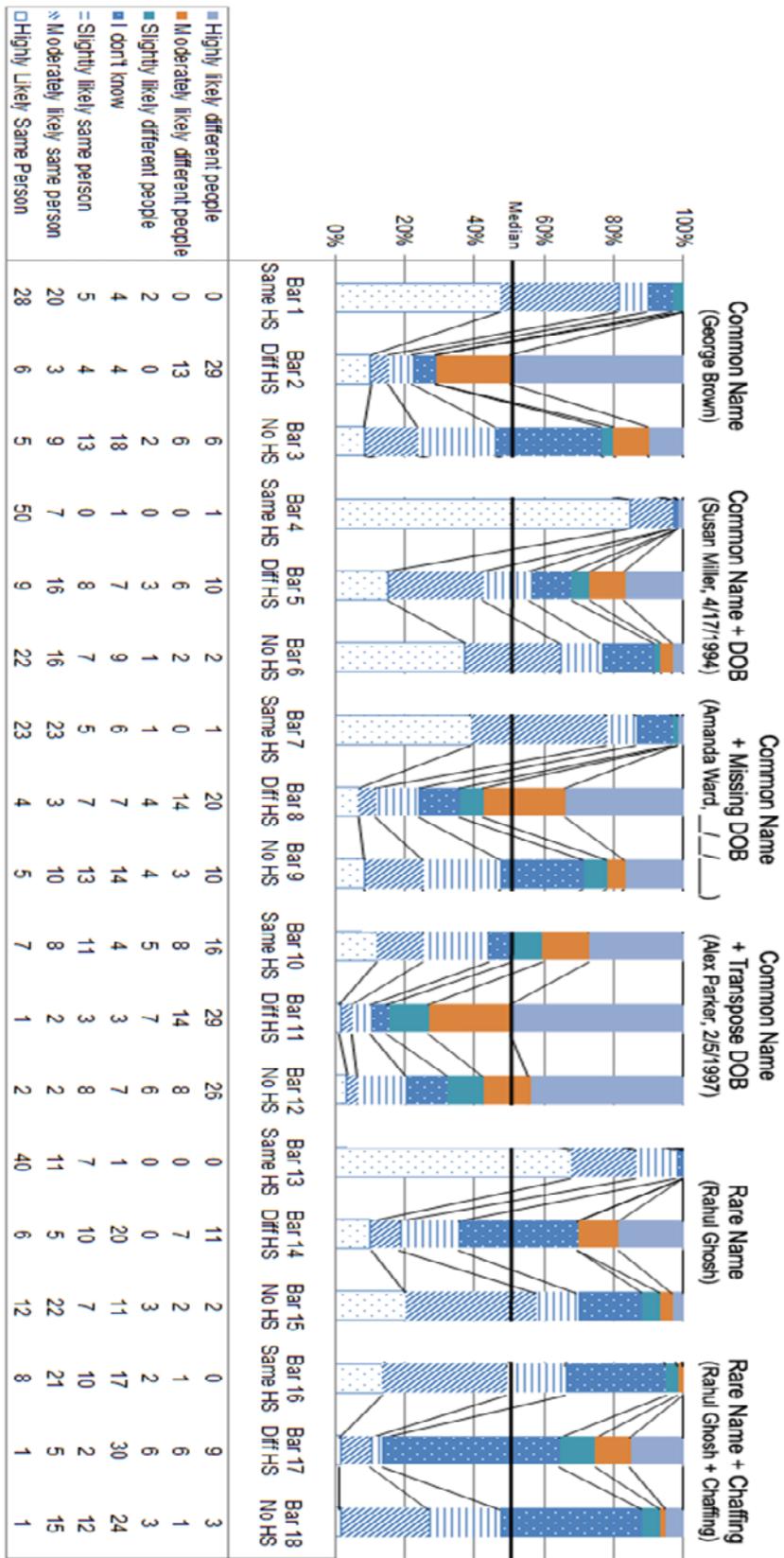


Figure 6. Results

## APPENDIX

### *A. Details for Encrypting Connection Information for Decoupling.*

Each decoupled table has its own private key to decrypt the row association. The private key for table association is held only by the system administrator of the decoupled information system while each private key for row association of particular tables are given out to the owners of each table. Thus, the full original table can only be reconstructed outside the system when both the system administrator and the owner of the table agree to do so and present the private key [9]. Such situations should rarely occur since the original table is never needed for research. During normal operations, this critical connection information is only required by the computerized third party, the decoupled data integration software, and never by any person. The computerized third party software needs access to both private keys to manage the record linkage process in order to produce the de-identified linked table (Figure 1). Thus, the private keys will be stored in the TPM (Trusted Platform Model)-based sealed storage. Flickr is an example of a TPM, which can execute security sensitive code in complete isolation using the new commodity processors from AMD and Intel. It can protect critical information, such as these private keys, even when the OS is compromised by utilizing the hardware support for late launch and attestation [11].

### *B. Name and DOB in Clerical Review*

During clerical review, revealing names without distortion is important for accurate entity resolution because it is too difficult to convey similarity information between the numerous variations of the same name. Common prefixes (Ms.) and postfixes (III) make the situation worse. We show that even when names are revealed, privacy can still be maintained due to the non-uniqueness of names and uncertainty in the universe around the data. There is no way to know how many people named 'John Smith' exist in the universe of the data, especially when the universe is unknown. Our experiments confirm that even with rare names, when the universe is effectively manipulated people are not able to confidently infer sensitive information. We discuss universe manipulation in section II.C.

Another useful PII often used for linkage is birthdate. Initially we considered revealing DOB during the review process because it would be too complex to display all possible permutations of data errors that needs to be considered. But in our experiment as well as in the literature, we found that revealing a pair of name and DOB had high potential for inferring identities [5]. Previous research show that combination of multiple characteristics may result in almost unique characteristics [11, 13]. Hence, we recommend DOB be dealt with similarly to SSN such that only the difference between two DOBs is shown. However, DOB comparisons should be made on an element to element basis for month, day, and year. In addition, transpose of month and day should be accounted for as well as transposes within one element. More research on common typos in dates is needed to determine exactly what are all the meaningful differences between two dates that should be taken into account during record linkage.

*C. A Full Set of Three Questions for the Base Case.*

<p>Q1. Meet George Brown, a student at Meadowgreen High School.</p> <p>Using only the information on this screen, how likely is it that the George Brown introduced above is the same person listed on the honor roll provided here?</p> <ul style="list-style-type: none"> <li>• Highly likely to be the same person</li> <li>• Moderately likely to be the same person</li> <li>• Slightly likely to be the same person</li> <li>• I don't know if they are the same person or not</li> <li>• Slightly likely to be two different people</li> <li>• Moderately likely to be different people</li> <li>• Highly likely to be two different people</li> </ul>	<p><b>Meadowgreen HS Honor Roll</b></p> <table border="1"> <thead> <tr> <th>Name</th> </tr> </thead> <tbody> <tr><td>Amanda Ward</td></tr> <tr><td>Edward Jones</td></tr> <tr><td>Hilary Ford</td></tr> <tr><td>George Brown</td></tr> <tr><td>Susan Miller</td></tr> <tr><td>David Green</td></tr> <tr><td>Alexander Parker</td></tr> <tr><td>Brian Richards</td></tr> <tr><td>Daniel Parker</td></tr> <tr><td>Alex Parker</td></tr> </tbody> </table>	Name	Amanda Ward	Edward Jones	Hilary Ford	George Brown	Susan Miller	David Green	Alexander Parker	Brian Richards	Daniel Parker	Alex Parker
Name												
Amanda Ward												
Edward Jones												
Hilary Ford												
George Brown												
Susan Miller												
David Green												
Alexander Parker												
Brian Richards												
Daniel Parker												
Alex Parker												
<p>Q2. Meet George Brown, a student at Meadowgreen High School.</p> <p>Using only the information on this screen, how likely is it that the George Brown introduced above is the same person listed on the honor roll provided here?</p> <ul style="list-style-type: none"> <li>• Highly likely to be the same person</li> <li>• Moderately likely to be the same person</li> <li>• Slightly likely to be the same person</li> <li>• I don't know if they are the same person or not</li> <li>• Slightly likely to be two different people</li> <li>• Moderately likely to be different people</li> <li>• Highly likely to be two different people</li> </ul>	<p><b>Valley Mountain HS Honor Roll</b></p> <table border="1"> <thead> <tr> <th>Name</th> </tr> </thead> <tbody> <tr><td>Alexander Parker</td></tr> <tr><td>Mary Scott</td></tr> <tr><td>Thomas Wright</td></tr> <tr><td>Alex Parker</td></tr> <tr><td>Amanda Ward</td></tr> <tr><td>David Green</td></tr> <tr><td>Brian Richards</td></tr> <tr><td>Garrett Fox</td></tr> <tr><td>Susan Miller</td></tr> <tr><td>George Brown</td></tr> </tbody> </table>	Name	Alexander Parker	Mary Scott	Thomas Wright	Alex Parker	Amanda Ward	David Green	Brian Richards	Garrett Fox	Susan Miller	George Brown
Name												
Alexander Parker												
Mary Scott												
Thomas Wright												
Alex Parker												
Amanda Ward												
David Green												
Brian Richards												
Garrett Fox												
Susan Miller												
George Brown												
<p>Q3. Meet George Brown, a student at Meadowgreen High School.</p> <p>Using only the information on this screen, how likely is it that the George Brown introduced above is the same person listed on the honor roll provided here?</p> <ul style="list-style-type: none"> <li>• Highly likely to be the same person</li> <li>• Moderately likely to be the same person</li> <li>• Slightly likely to be the same person</li> <li>• I don't know if they are the same person or not</li> <li>• Slightly likely to be two different people</li> <li>• Moderately likely to be different people</li> <li>• Highly likely to be two different people</li> </ul>	<p><b>A High School Honor Roll</b></p> <table border="1"> <thead> <tr> <th>Name</th> </tr> </thead> <tbody> <tr><td>Susan Miller</td></tr> <tr><td>Daniel Parker</td></tr> <tr><td>Amanda Ward</td></tr> <tr><td>Alexander Parker</td></tr> <tr><td>George Brown</td></tr> <tr><td>Laura Baldwin</td></tr> <tr><td>Alex Parker</td></tr> <tr><td>Jennifer Wood</td></tr> <tr><td>David Green</td></tr> <tr><td>Brian Richards</td></tr> </tbody> </table>	Name	Susan Miller	Daniel Parker	Amanda Ward	Alexander Parker	George Brown	Laura Baldwin	Alex Parker	Jennifer Wood	David Green	Brian Richards
Name												
Susan Miller												
Daniel Parker												
Amanda Ward												
Alexander Parker												
George Brown												
Laura Baldwin												
Alex Parker												
Jennifer Wood												
David Green												
Brian Richards												

Table A.1 The set of three basic survey questions for the first scenario.