A Hybrid Approach to Private Record Linkage

Ali Inan [#1*], Murat Kantarcioglu [#2*], Elisa Bertino [†3*], Monica Scannapieco [‡4]

[#]*Department of Computer Science, The University of Texas at Dallas*
Richardson, TX 75083, USA
[1]inan@student.utdallas.edu
[2]muratk@utdallas.edu

[†]*Department of Computer Sciences, Purdue University*
West Lafayette, IN 47907, USA
[3]bertino@cs.purdue.edu

[‡]*Dipartimento di Informatica e Sistemistica, Universita di Roma "La Sapienza"*
Roma 00198, Italy
[4]monscan@dis.uniroma1.it

***Abstract*— Real-world entities are not always represented by the same set of features in different data sets. Therefore matching and linking records corresponding to the same real-world entity distributed across these data sets is a challenging task. If the data sets contain private information, the problem becomes even harder due to privacy concerns. Existing solutions of this problem mostly follow two approaches: sanitization techniques and cryptographic techniques. The former achieves privacy by perturbing sensitive data at the expense of degrading matching accuracy. The later, on the other hand, attains both privacy and high accuracy under heavy communication and computation costs. In this paper, we propose a method that combines these two approaches and enables users to trade off between privacy, accuracy and cost. Experiments conducted on real data sets show that our method has significantly lower costs than cryptographic techniques and yields much more accurate matching results compared to sanitization techniques, even when the data sets are perturbed extensively.**

I. Introduction

Integration of information maintained by different entities is critical for various applications. Consider the health care industry, where complete medical history of a patient is often not readily available to researchers at a single source but is distributed across many hospital databases. While medical researchers would try to experiment with as much data as possible, hospitals would not be willing to disclose private records of their patients. In this scenario, private record linkage is the first and possibly the most important step towards utilization of private information.

The process of identifying and linking different representations of the same real-world entity across multiple data sources is known as the *record linkage* problem. Since it is a key component of data integration methodologies, record linkage has been investigated extensively [1]. However especially after the introduction of powerful data mining techniques, privacy concerns related to sharing of individual information have pushed research towards the re-formulation of the record linkage problem and the development of new solutions [2], [3], [4], [5], [6].

*Authors have been partially supported by AFOSR grant FA9550-07-1-0041.

In order to prevent privacy concerns from hampering sharing of private information, two main approaches have been developed. These are sanitization methods that perturb private information to obscure individual identity [7], [8], [9], [10] and cryptographic methods that rely on Secure Multi-party Computation (SMC) protocols [11].

Sanitization techniques such as k-anonymization [8] or random noise addition [9], [12] usually involve privacy metrics that measure the amount of privacy protection. Higher levels of protection typically translate into further deviation from the original data and consequently less accurate results. Therefore sanitization techniques involve trade-off between accuracy and privacy.

Cryptographic techniques do not sacrifice accuracy to achieve privacy. The algorithms applied to private data are converted to series of functions with private inputs. Then, using SMC protocols, accurate results are obtained. Under reasonable assumptions regarding computational power of the adversary, SMC protocols guarantee that only the final result and any information that can be inferred from the final result is revealed [11]. SMC protocols generally have some security parameters (e.g. encryption key sizes) that allow users to trade off between cost and privacy [11].

Both those approaches are thus not able to provide a comprehensive solution addressing all relevant application requirements with respect to privacy, cost, and accuracy. The goal of our work is to address the limitations of such approaches. We propose a novel method to address the private record linkage problem that combines cryptographic techniques and anonymization methods, a branch of sanitization techniques. Unlike existing methods, trade-off in our solution is along three dimensions: privacy, cost, and accuracy. To the best of our knowledge, ours is the first systematic approach in this direction.

We assume three participants in our method. These are two data holders, with the data sets to be linked, and the querying party, who provides the classifier that determines matching record pairs. The basic idea is to utilize anonymized data sets to accurately match or mismatch a large portion of record pairs

so that the need for costly SMC protocols is minimized. We call this the *blocking* step, in reference to a similar technique employed in record linkage methods [13]. The *blocking* step provides cost savings proportional to the level of anonymity, set independently by each participant. Later on, the *blocking* step is followed by the *SMC* step, where unlabeled record pairs are labeled using cryptographic techniques.

If the input data sets are too large, we may have to label significant amounts of record pairs using cryptographic techniques. In such cases, since cost of the private record linkage process is not known in advance, data holder parties might be unwilling to participate. That's why, we consider limiting the costs of cryptographic techniques. This also allows us to analyze the cost-accuracy and cost-privacy relationships. When the upper bound imposed on SMC costs is too low, some record pairs might remain unlabeled at the end of the *blocking* step. In order not to reveal irrelevant pairs, we label them non-matched. While this precaution ensures 100% precision, it degrades recall since some of those unlabeled record pairs might actually be matching. Fortunately, anonymized data sets can help reduce the effects. Based on generalizations of records, pairs that are more likely to match can be given priority in the *SMC* step.

Our method has many advantages over existing methods. The main advantages can be summarized as follows:

- Costs are usually lower than, and at worst, equal to the costs of existing cryptographic techniques.
- Precision is always 100%, which implies that irrelevant record pairs are protected against disclosure.
- Recall varies with the upper bound on SMC costs, imposed by participants.
- Our method applies to any anonymization method and any cryptographic technique. Participants can choose different anonymization methods, anonymity levels, quasi-identifier attribute sets.

Rest of the paper is organized as follows. In Section II, we formally define the problem. An overview of the proposed solution is provided in Section III. Then we describe the blocking mechanisms in Section IV. Section V presents the SMC protocol and various selection heuristics. Experimental results are provided in Section VI. We review related work in the area in Section VII. In Section VIII, we conclude with the discussion on future research directions.

II. Problem Definition

Record linkage is the process of identifying record pairs, across two input data sets, that correspond to the same real-world entity. In essence, the problem consists of building a classifier that accurately classifies pairs of records as "match" or "mismatch" and applying this classifier to the input data sets efficiently. In the private record linkage problem, on the other hand, an accurate classifier is assumed to be available [5] (i.e. in our problem setting, the classifier is provided by the querying party). Therefore private record linkage methods focus on classifying all record pairs within the input data sets privately, accurately and efficiently.

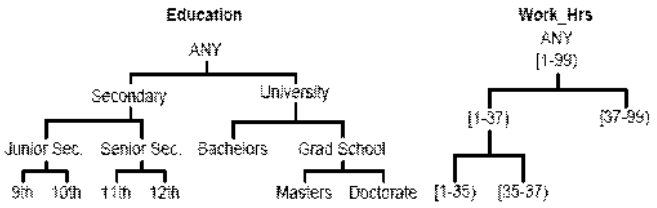

Fig. 1. VGHs for *Education* and *Work_Hrs* attributes

Without loss of generality, let the input data sets, R and S, be represented as relations. Let us also assume that these relations have the same schema, $R(A_1, ...A_n)$ and $S(A_1, ...A_n)$. If not, schemas of R and S can be matched using private schema matching techniques (e.g. the method described by Scannapieco et al. in [5]).

Given distance functions $d_i : Dom(R.A_i) \times Dom(S.A_i) \rightarrow \mathbb{R}^+$, defined over domains of corresponding attributes of R and S, and matching thresholds $\theta_i \geq 0$, record linkage can be expressed as a join operation over R and S. A record pair (r, s), where $r \in R$ and $s \in S$, is a matching record pair if $d_i(r.a_i, s.a_i) \leq \theta_i$ for all attributes $0 \leq i < n$. Then the join condition can be defined based on the following decision rule that returns $true$ for matching record pairs and $false$ for mismatching record pairs:

$$dr(r,s) = \left\{ \begin{array}{ll} \text{true} & \text{if } \forall 0 \leq i < n, d_i(r.a_i, s.a_i) \leq \theta_i \\ \text{false} & \text{otherwise} \end{array} \right\}.$$

Our task is to identify the join result $R \bowtie_{dr(r,s)} S$ in a privacy preserving manner such that the result will be available to the querying party and private records of the data holders, that do not satisfy the join condition are not disclosed.

III. Overview of the Solution

In this section, we provide an overview of the proposed solution. Our purpose is to exemplify the concepts before providing the formal definitions in Section IV and Section V.

Consider the relations R and S, with matching schemas (A_1, A_2), in Table I and Table II. Let R and S be the input relations that the querying party wants to join, such that the classifier provided by the querying party has parameters $\theta_1 = 0.5$, $\theta_2 = 0.2$ and d_1 is the Hamming distance, d_2 is the Euclidean distance. In this scenario, record pair (r, s) is labeled match if $r.a_1 = s.a_1$ and $\sqrt{(r.a_2 - s.a_2)^2} \leq 0.2 \times normFactor$, where $normFactor$ is the normalization factor for A_2. r and s are mismatched otherwise.

Let $A_1 \in Dom(Education)$ and $A_2 \in Dom(Work\_Hrs)$, where $Education$ and $Work\_Hrs$ are the attributes with the value generalization hierarchies (VGH) provided in Figure 1 [7]. Suppose that some anonymization method (e.g. k-anonymization [8]) outputs R' given R when $k = 3$ and S' given S when $k = 2$.

Notice that there are $|R| \times |S| = 6 \times 6 = 36$ record pairs across R and S. Based on anonymized relations R' and S', let us see if we can label any record pairs.

TABLE I

DATA SET R AND R'S 3-ANONYMOUS GENERALIZATION R'

R	A_1	A_2
r_1	Masters	35
r_2	Masters	36
r_3	Masters	36
r_4	9th	28
r_5	10th	22
r_6	12th	33

R'	A_1	A_2
r'_1	Masters	[35-37)
r'_2	Masters	[35-37)
r'_3	Masters	[35-37)
r'_4	Secondary	[1-35)
r'_5	Secondary	[1-35)
r'_6	Secondary	[1-35)

TABLE II

DATA SET S AND S'S 2-ANONYMOUS GENERALIZATION S'

S	A_1	A_2
s_1	Masters	36
s_2	Masters	35
s_3	Bachelors	27
s_4	11th	33
s_5	11th	22
s_6	12th	27

S'	A_1	A_2
s'_1	Masters	[35-37)
s'_2	Masters	[35-37)
s'_3	ANY	[1-35)
s'_4	ANY	[1-35)
s'_5	Senior Sec.	[1-35)
s'_6	Senior Sec.	[1-35)

Consider r'_1 and s'_5 with the generalized sequences $(Masters, [35-37))$ and $(SeniorSec., [1-35))$ respectively. $Masters$ is a leaf value in the $Education$ VGH. Therefore we know that $r_1.a_1 = Masters$. However, $SeniorSec.$ is not. Possible values of $s_5.a_1$ are $11th, 12th$, the set of leaf values to which $SeniorSec.$ can specialize. Obviously $d_1(r_1.a_1, s_5.a_1) = 1$ based on Hamming distance, because none of the specific values of $SeniorSec.$ equals $Masters$. Since $\theta_1 = 0.5$ and $d_1(r_1.a_1, s_5.a_1) = 1 > \theta_1$, we conclude that (r_1, s_5) is a mismatching record pair based on r'_1 and s'_5.

Without further ado, we can also mismatch the record pairs (r_1, s_6), (r_2, s_5), (r_2, s_6), (r_3, s_5), (r_3, s_6) because r_1, r_2, r_3 and s_5, s_6 are generalized to the same sequence. Notice that situation is the same with r_4, r_5, r_6 and s_1, s_2. Therefore, all record pairs in the Cartesian product $(r_4, r_5, r_6) \times (s_1, s_2)$ mismatch. We do not need to repeat the process for pairs generalized to the same sequences.

Now consider r'_1 and s'_1. Both r_1 and s_1 are not generalized on a_1 and share the value $Masters$. Therefore $d_1(r_1.a_1, s_1.a_1) = 0$ and θ_1 is not violated. Yet satisfying the matching threshold on A_1 is not sufficient to declare (r_1, s_1) a match, we should also make sure that this pair respects θ_2. According to Figure 1, the range of A_2 is $[1, 99)$. Since $\theta_2 = 0.2$, this implies $d_2(r_1.a_2, s_1.a_2) < 0.2 \times (99-1) = 19.6$ is the sufficient condition to match r_1 and s_1 ($normFactor = 98$ according to the VGH). We are given $r'_1.a_2 = s'_1.a_2 = [35-37)$. Definitely any two value chosen from the interval $[35-37)$ are less than 19.6 apart. That's why we match r_1 and s_1. Similarly, (r_1, s_2), (r_2, s_1), (r_2, s_2), (r_3, s_1), (r_3, s_1) can be matched based on the anonymized relations.

Can we always decide the label of a pair of records given corresponding records' generalizations? Let us try to label (r_1, s_3) given (r'_1, s'_3). As discussed before, we know that $r_1.a_1 = Masters$. However, $r_1.a_2$ can assume any value in the interval $[35-37)$. Similarly, $s_3.a_1$ can be any value within $Dom(Education)$ (since $s'_3.a_1 = ANY$) and $s_3.a_2$ can be any value in the interval $[1-35)$. Suppose that $s_3 = (Masters, 34)$ and $r_1 = (Masters, 35)$, both valid values. If this were the case, (r_1, s_3) would match (since $35-34 < 19.6$, normalized threshold). Now consider another pair of valid values: $s_3 = (11th, 32)$ and $r_1 = (Masters, 35)$. If this were the case, (r_1, s_3) would mismatch (since $Masters \neq 11th$). We conclude that a clear decision can not be made.

Repeating the process for all generalization sequences, 12 record pairs can be mismatched and 6 record pairs can be matched through the anonymized relations. Labels of the 18 remaining record pairs are unknown/undecided. For a formal discussion, please refer to Section IV.

In our method, all record pairs that cannot be labeled based on their generalization sequences are input to the SMC protocols. Given R' and S', there are 18 such pairs. Now, suppose that due to high costs, the participants can endure comparing at most 10 of these pairs with SMC protocols. We consider the problem of selecting these 10 pairs in Section V-C and labeling remaining 8 pairs in Section V-B. Details of the SMC protocols are provided in Section V-A.

Some extreme scenarios, regarding privacy requirements of the participants and cost limitations give excellent insights on the nature of privacy, accuracy and cost trade-offs in our method. Let us next discuss two of them.

(1) $k = 1$: privacy is minimum since anonymized relation is actually the original relation ($R' = R$ in the example above). Yet, since anonymized relation is very precise (i.e. no generalization values), all record pairs can be labeled based on the anonymized relation. This implies no costs due to SMC protocols.

(2) $k = |R|$: privacy is maximum since anonymized relations (most probably) consist of records generalized to the root value of VGHs on all attributes. Therefore the costs would be similar to pure SMC methods.

IV. BLOCKING STEP

Quasi-identifier attribute values of anonymized records are imprecise due to generalization but are always accurate in the sense that generalization values are chosen in accordance with the original value. Given generalization $gen(r)$ of a record r, the value of the i^{th} field of $gen(r)$, denoted by $gen(r).a_i$, determines the set of values that $r.a_i$ can assume. We call this set a *specialization set* and denote it with $specSet(.)$. If the i^{th} field is not a quasi-identifier, then $gen(r).a_i = \{r.a_i\}$ and therefore $specSet(gen(r).a_i) = r.a_i$. For discrete-valued quasi-identifiers, $specSet(.)$ consists of all leaf nodes of the value generalization hierarchy (VGH) of attribute i, into which $gen(r).a_i$ might specialize. For continuous attributes, $specSet(.)$ represents an interval. Consider $s'_5 = gen(s_5)$ in Table I, where $specSet(gen(s_5).a_1) = specSet(s'_5.a_1) = specSet(SeniorSec.) = \{11th, 12th\}$ according to Figure 1.

Anonymized data sets can help matching some record pairs and reducing SMC costs, as we have shown in Section III. However, we cannot use $dr(r, s)$ for this purpose, because

instead of r and s, only $gen(r)$ and $gen(s)$ are available. Therefore distance functions d_i do not apply anymore. In addition, now there are not two but three labels: match (M), mismatch (N), and unlabeled/unknown (U), since we cannot always decide whether given anonymized record pair matches or mismatches.

Before formalizing the blocking process, we would like to highlight the similarity with the probabilistic record matching problem discussed in [14]. In this version of the problem, the matching algorithm is not forced to label every record pair as match (M) or non-match (N). Instead, a third label, possible-match (P) is allowed. Whenever the decision between M and N is not obvious, the algorithm classifies it as P and delegates the decision to domain experts, who are accurate but expensive to hire. In our case, the SMC circuit corresponds to these domain experts since it attains high accuracy at high cost. The new decision mechanism corresponds to the probabilistic decision rule that identifies the sets M, N and P. The most important difference is that anonymized data is not dirty but imprecise, which is the reason why precision is 100% in our method.

Given two generalized values $v = gen(r).a_i$ and $w = gen(s).a_i$, we know for sure that $(r.a_i, s.a_i) \in specSet(v) \times specSet(w)$. The greatest lower bound on the distance between any pair of elements within $specSet(v) \times specSet(w)$ is defined as the infimum of the distance values. If the infimum distance for anonymized records $gen(r).a_i$ and $gen(s).a_i$ is larger than θ_i, then we can determine (accurately) that $d_i(r.a_i, s.a_i) > \theta_i$. Slack distance function that returns infimum distance sd_i^l is defined as

$$sd_i^l(v, w) = inf_{\substack{p \in specSet(v) \\ q \in specSet(w)}}(d_i(p, q)).$$

Supremum distance can be defined similarly as the least upper bound on the distance between any pair of elements in the corresponding specialization sets. If the supremum distance for anonymized records $gen(r).a_i$ and $gen(s).a_i$ is smaller than θ_i, then we can determine (accurately) that $d_i(r.a_i, s.a_i) \leq \theta_i$. Slack distance function that returns supremum distance is formally defined as

$$sd_i^s(v, w) = sup_{\substack{p \in specSet(v) \\ q \in specSet(w)}}(d_i(p, q)).$$

According to the decision rule dr, a pair of records (r, s) should agree on all attributes to match. Our method initially tries to make the decision through the anonymized data sets by computing sd_i^s and sd_i^l for each attribute. The slack decision rule, $sdr(v, w))$ can be expressed as

$$sdr(v, w) = \left\{ \begin{array}{cc} \text{N} & \text{if } \exists 0 \leq i < n, sd_i^l(v.a_i, w.a_i) > \theta_i \\ \text{M} & \text{if } \forall 0 \leq i < n, sd_i^s(v.a_i, w.a_i) \leq \theta_i \\ \text{U} & \text{otherwise} \end{array} \right\}.$$

V. Secure Multi-Party Computation Step

In classical SMC protocols, using some cryptographic assumptions, it can be proved that only the final results and anything that could be inferred by looking at the final results are revealed. In our case, we implicitly assume that the disclosure of the anonymized data is not a privacy violation. Therefore, our security guarantees are slightly different than the security guarantees provided by the generic SMC protocols [11]. In other words, our goal is to *only* reveal the final record linkage result, the anonymized data sets and anything that can be inferred by looking at the final result and the anonymized data sets. Since blocking step only depends on the anonymized data sets, it satisfies the goal stated above. (i.e. anything revealed during the *blocking* step could be inferred using the anonymized data sets.)

In this section, we describe methods for relabeling the record pairs that were labeled unknown (U) in the blocking step without revealing anything that could not be inferred by looking at the final result and the anonymized data sets. Also we assume that due to budget constraints of the participants, we can match at most $SMC\_allowance$ many record pairs using SMC protocol invocations. If $|U| \leq SMC\_allowance$, then obviously all unlabeled record pairs can be relabeled using the SMC protocols. Otherwise, there are two problems to be addressed: (1) How are the pairs, not selected for the *SMC* step, relabeled? (2) How are the pairs that will be relabeled by the SMC protocols selected? Before delving into details of these issues, we discuss the possible SMC protocols.

A. SMC Protocols for Record Linkage

For each pair of records that is not blocked, we need to securely learn whether such a pair actually matches or not. In other words, for each possibly matching record pair and for each attribute, we need to securely calculate whether $d_i(r.a_i, s.a_i) \leq \theta_i$ is satisfied or not. Such a secure calculation is possible using generic SMC circuit evaluation techniques [11]. Also recently many protocols have been proposed using special encryption functions such as commutative encryption [15] and homomorphic encryption [16]. For example, homomorphic public key encryption can be used to securely compute Euclidean distances.

Let $E_{pk}(.)$ denote the encryption function with public key pk and $D_{pr}(.)$ denote the decryption function with private key pr. A secure public key cryptosystem is called homomorphic if it satisfies the following requirements: (1) Given the encryption of m_1 and m_2, $E_{pk}(m_1)$ and $E_{pk}(m_2)$, there exists an efficient algorithm to compute the public key encryption of $m_1 + m_2$, denoted $E_{pk}(m_1 + m_2) := E_{pk}(m_1) +_h E_{pk}(m_2)$. (2) Given a constant k and the encryption of m_1, $E_{pk}(m_1)$, there exists an efficient algorithm to compute the public key encryption of km_1, denoted $E_{pk}(km_1) := k \times_h E_{pk}(m_1)$.

Using homomorphic encryption, $d_i(r.a_i, s.a_i) = (r.a_i - s.a_i)^2 = (r.a_i)^2 - 2 \times r.a_i \times s.a_i + (s.a_i)^2$ can be securely calculated as follows: Querying party creates a homomorphic public/private key pair, and sends the public key to data holders (say Alice and Bob). Later on, using the public key, Alice can compute $E_{pk}((r.a_i)^2), E_{pk}(-2 \times r.a_i)$ and send it to Bob. Now Bob can calculate $E_{pk}((r.a_i)^2) +_h (E_{pk}(-2 \times r.a_i) \times_h s.a_i) +_h E_{pk}((s.a_i)^2)$ which is equal to $E_{pk}((r.a_i - s.a_i)^2)$ and send the result back to querying site. Using the private

key, querying site can decrypt the received message to learn the distance result. Such secure distance evaluation could be combined with secure comparison to not to reveal even the distance result.

Although homomorphic encryption based SMC protocols could be used in our case, we would like to stress that in our SMC step, *any* SMC technique that can securely compute $d(r, s)$ could be used.

B. Labeling Remaining Unlabeled Pairs

All record pairs labeled U in the *blocking* step might not be relabeled in the *SMC* step due to cost constraints. We next analyze 3 strategies for handling such record pairs. Notice that the method of selecting pairs for the *SMC* step, discussed in Section V-C, depends on the strategy chosen here.

(1) Maximizing precision: Based on anonymized data sets, a classifier c_1 that selects probably-matching record pairs is built. Selected record pairs are labeled by the SMC protocols. All record pairs that were not chosen by the classifier in the *SMC* step are labeled mismatch. Since SMC protocols are accurate, there won't be any false-positives. However, recall might be low since some remaining unlabeled pairs can match.

(2) Maximizing recall: Based on anonymized data sets, a classifier c_2 that selects probably-mismatching record pairs is built. After the *SMC* step, all remaining unlabeled record pairs are considered matching pairs. Since SMC protocols are accurate, all matching pairs will be labeled correctly. Therefore recall is high, but precision might be low.

(3) Maximizing precision and recall: Record pairs for the *SMC* step are selected at random. Generalizations of selected pairs, together with the labels collected from the SMC protocol are used as training data to build a classifier c_3 that labels generalized record pairs. Notice that if the set of matching attributes ($\theta_i < 1$) is equal to the set of quasi-identifiers, then c_3 can not discriminate record pairs that have the same generalization. Considering the k-anonymization method [8], there are at least k records generalized to the same sequence. This implies that groups of record pairs, with at least k^2 cardinality, will be classified similarly. Due to low data quality resulting from anonymization, we conclude intuitively that c_3 can not attain high precision or recall.

The first strategy would be advantageous to the data holder parties because high precision prevents disclosure of irrelevant record pairs to the party issuing the join operation. On the other hand, the second strategy would be advantageous to the query issuer, because recall is high. Yet, possibly low precision in the second strategy would violate privacy of individuals. Since privacy is our primary concern, we choose to follow the first strategy.

C. Selection Heuristics for Circuit Evaluation

Since we label all remaining unlabeled pairs as match, selection heuristics should aim at finding possibly matching record pairs. For this purpose, we define a third distance function $dExp(gen(r).a_i, gen(s).a_i))$ that returns the expected distance between generalized record values, $gen(r).a_i$ and $gen(s).a_i$.

Let $V = specSet(gen(r).a_i)$ and $W = specSet(gen(s).a_i)$ be the random variables that represent possible values of $r.a_i$ and $s.a_i$. Then $dExp(gen(r).a_i), gen(s).a_i)) = E(d_i(V, W))$. To simplify this equation, we need to specify the distance function d_i and the probability distributions of V and W.

Without loss of generality, we assume that d_i is Euclidean distance for continuous attributes and Hamming distance for discrete (categorical) attributes.

Notice that participants would not (and should not) release any statistics on the distribution of original values within generalizations. Therefore, due to absence of such feedback, the best option is assuming $R.A_i$ is uniformly distributed for each attribute i. In this case, any value $v \in V$ is equally likely. For discrete attributes, this implies $Pr(r.a_i = v) = 1/|V|$.

For discrete attributes, assuming that V and W are independent, we first express the expected distance, $E_D = E(d_i(V, W))$, as a summation over the random variables and their probability distribution functions, as shown in Equation 1. Then we insert the values of $Pr(V = v)$ and $Pr(W = w)$ based on the uniform distribution assumption. Hamming distance $d_i(v, w)$ can be represented by an indicator, $I_{v \neq w}$. After replacing the indicator with $d_i(v, w)$, we arrive at Equation 3. Equation 4 inserts the summation result and Equation 5 simplifies Equation 4.

$$
\begin{aligned}
E_D &= \sum_{\substack{v \in V \\ w \in W}} d(v, w) \cdot Pr(V = v) \cdot Pr(W = w) && (1)\\
&= \sum_{\substack{v \in V \\ w \in W}} d(v, w) \cdot \frac{1}{|V|} \cdot \frac{1}{|W|} && (2)\\
&= \frac{1}{|V| \cdot |W|} \sum_{\substack{v \in V \\ w \in W}} I_{v \neq w} && (3)\\
&= \frac{1}{|V| \cdot |W|} \cdot (|V| \cdot |W| - |V \cap W|) && (4)\\
&= 1 - \frac{|V \cap W|}{|V| \cdot |W|}. && (5)
\end{aligned}
$$

For continuous attributes, instead of computing the expected Euclidean distance ($\sqrt{(x-y)^2}$), we prefer to compute $(x-y)^2$. This allows further simplifications. Let E_D denote the expected square distance. Our first step is inserting the value of this distance function as shown in Equation 6. Then, based on the assumption that V and W are independent variables, Equation 7 is derived. Finally, in Equation 8, we simplify each term by integrating over the intervals of the uniform random variables: $a_1 \leq v \leq b_1$ and $a_2 \leq w \leq b_2$.

$$
\begin{aligned}
E_D &= E(|V^2 - 2 \cdot V \cdot W + W^2|) && (6)\\
&= E(V^2) - 2 \cdot E(V) \cdot E(W) + E(W^2) && (7)\\
&= 1/3 \cdot (a_1^2 + b_1^2 + a_2^2 + b_2^2 + a_1 b_1 + a_2 b_2) && (8)\\
&\quad -1/2 \cdot (a_1 + b_1)(a_2 + b_2).
\end{aligned}
$$

Using expected distance functions, it is possible to devise many heuristics. We discuss some examples of such heuristics in Section VI.

VI. Experimental Results

We performed our experiments on the real-world *Adult* data set from the UC Irvine Machine Learning Repository [17]. This data set has been heavily used to evaluate different anonymization methods and its quasi-identifier attributes and corresponding value generalization hierarchies have been well established. We adopted value generalization hierarchies of all attributes, except the continuous *age* attribute, from [7] (see Section VI-A for details). The hierarchy that we used consists of 4 levels and equi-width leaf nodes cover 8-unit intervals.

In order to build two input data sets, we first removed all tuples with missing values. The remaining 30,162 records were randomly partitioned into three data sets, d_1, d_2 and d_3, each consisting of 10,054 records. Then, we merged d_1 and d_3 to build the first data set, D_1, and d_2 and d_3 to build the second data set, D_2. In this setting, regardless of the matching thresholds, θ_i, corresponding records in the non-empty intersection $D_1 \cap D_2 = (d_1 \cup d_3) \cap (d_2 \cup d_3) = d_3$ should match each other.

We implemented the SMC circuit described in Section V-A using Paillier homomorphic public key cryptosystem with 1024-bit key length [18]. Experiments conducted with a PC that attains 2.8GHz clock speed and has 2GB available memory indicate that, on average, computing the distance for a single continuous attribute takes 0.43 seconds. Including file I/O, the anonymization method described in Section VI-A takes 2.02 and 2.03 seconds to anonymize D_1 and D_2 respectively. Blocking step costs 1.35 seconds on average. Notice that, in total, the costs incurred for non-cryptographic operations are equivalent to the cost of comparing only $(2.02+2.03+1.35)/0.43 \approx 13$ continuous values (not records) with SMC. Yet we haven't even taken into account the cost of privately comparing these distance values with the matching thresholds. Obviously, cost of cryptographic operations dominates all other costs. Therefore, we restricted our cost model to the number of SMC protocol invocations. We represent the number of SMC protocol invocations, hereafter referred to as $SMC\_allowance$, as a percentage of the number of all record pairs, $|D_1| \times |D_2|$. If needed, translating this percentage into CPU time or network bandwidth is an easy task, given the key length of the secure circuit and data set sizes.

Our primary evaluation measure was accuracy. Please remember that in our method, any record pair that does not satisfy the join condition cannot be in the result set (i.e. no false-positives). Therefore, precision is always 100% and accuracy is actually determined by recall. In our context, recall is defined as the percentage of record pairs correctly labeled as match among all pairs satisfying the decision rule.

Recall measurements depend on the criteria for selecting the record pairs to be classified by SMC. We experimented 3 heuristics based on expected distance functions discussing in Section V-C. Record pairs with (1) $minFirst$: minimum attribute-wise expected distance first; (2) $maxLast$: maximum attribute-wise expected distance last; (3) $minAvgFirst$: minimum average attribute-wise expected distance first. Measurements corresponding to each heuristic are depicted as separate series in the figures.

We were also interested in the blocking efficiency of the underlying anonymization method, measured in terms of the percentage of record pairs that are permanently classified by the slack decision rule. For example, in Section III, among 36 record pairs, 18 pairs were blocked. Therefore the blocking efficiency would be 50%. Notice that blocking efficiency also indicates the sufficient $SMC\_allowance$ to achieve 100% recall. For instance, if 99% of all record pairs were classified in the blocking step, we would have to match the remaining 1% using the SMC protocol in order not to leave out any matching pair.

Both evaluation measures, that are, blocking efficiency and recall, vary with the matching thresholds, θ_i, as well as with the anonymization parameters, that are, anonymity requirement (k), set of quasi-identifiers. Recall also depends on the $SMC\_allowance$. Unless stated otherwise, default values of these parameters are as follows: k = 32, θ_i = 0.05 for all quasi-identifier attributes, $SMC\_allowance$ is 1.5% of all record pairs and the set of quasi-identifier attributes is {*age, work-class, education, marital status, occupation*}.

A. Anonymization Methods

Anonymization methods play a very crucial role in our method. Existing anonymization methods perform very poorly in terms of blocking efficiency. Apparently, the reason is the employed anonymization metrics. Later in this section, we will present an anonymization metric that favors the attribute with maximum entropy. The idea behind this metric is increasing the number of different generalization sequences within the anonymized data set. Let us now briefly discuss two of the existing anonymization methods.

$DataFly$ algorithm presented in [8] is one of the earliest anonymization methods. In this algorithm, records are generalized according to the attribute that has the most number of distinct values. When the anonymity requirement is met, or can be met by suppressing at most k records, the algorithm terminates. The attribute selection criteria is somehow similar to our entropy metric. But our method is not bottom-up, therefore we do not suppress any records.

In the top-down specialization method of [7], the purpose is maximizing the accuracy of some classifier trained on the anonymized data set. The algorithm first generalizes all records to the lowest granularity. Then, at each step, for each partition of specialized records, among the attributes that respect the k-anonymity requirement and that are beneficial for classification (i.e. information gain should not be 0), the one that maximizes information gain is selected. All records in the partition are specialized according to the selected attribute. This method has three major disadvantages for blocking purposes: (1) If a specialization is not beneficial, it is not performed. (2) Maximizing information gain implies minimizing class conditional

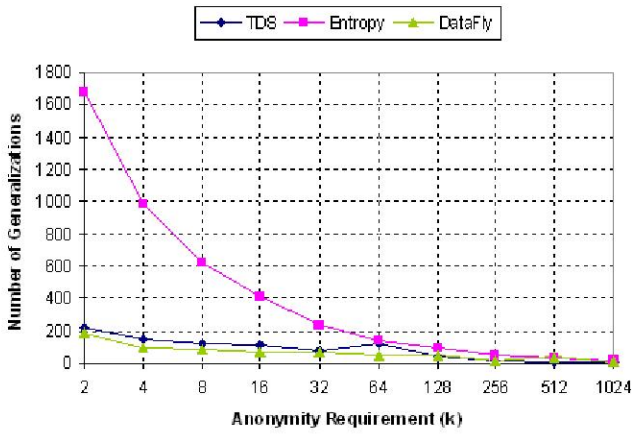

Fig. 2. Comparison of anonymization methods

entropy. Therefore, selected attribute might not maximize entropy. (3) Generalization hierarchies of continuous attributes, built on-the-fly based on entropy measures, tend to be shallow in depth due to wide leaf level intervals. As a result, continuous values are suppressed very quickly, preventing any record pairs from being blocked on these attributes.

We propose a new anonymization metric and apply this metric in a top-down fashion similar to [7]. As described above, the TDS algorithm [7] defines beneficial and valid attributes. In our method, every specialization is considered beneficial. However specialization might not be valid, based on the anonymity requirement k. Rather than minimizing class conditional entropy, at each step and for each partition, we choose the attribute that has maximum entropy. Therefore we make sure that partitions can withstand more specializations until the validity condition is violated. Consequently, the number of different generalizations is heuristically maximized. The advantage of more generalization sequences should be obvious: Since data set size is fixed, with more generalization sequences, every partition is smaller and more specific. This allows better blocking efficiency. Since our focus is not proposing anonymization methods, in order to save space, we are not providing the details. Please refer to [7] for details.

Figure 2 depicts the results on the *Adult* data set after removing records with missing values. As expected, the number of generalizations decreases as the anonymity requirement increases. Our anonymization method based on maximum entropy metric outperforms both $DataFly$ and TDS for lower values of k. However, as k increases (i.e. $k > 64$), our metric becomes less advantageous, due to over-generalization.

B. Anonymity Requirement: k

Anonymity requirement k is the most important parameter to adjust the amount of privacy protection and disclosure risk. Larger values of k imply more deviation from original data sets. Consequently, records are generalized to higher levels in the VGHs and thus the specialization sets grow larger. The result is reduced efficiency in the *blocking* step, as depicted in Figure 3.

For small values of k, due to high blocking efficiency, the *SMC* step can process all record pairs that are left unlabeled by the *blocking* step. However, as blocking efficiency decreases for larger values of k, the constant $SMC\_allowance$ level becomes insufficient. Due to the large number of unlabeled record pairs and limited $SMC\_allowance$, recall decreases significantly (see Figure 4). $MinAvgFirst$ performs much better than the other heuristics on over-perturbed data sets.

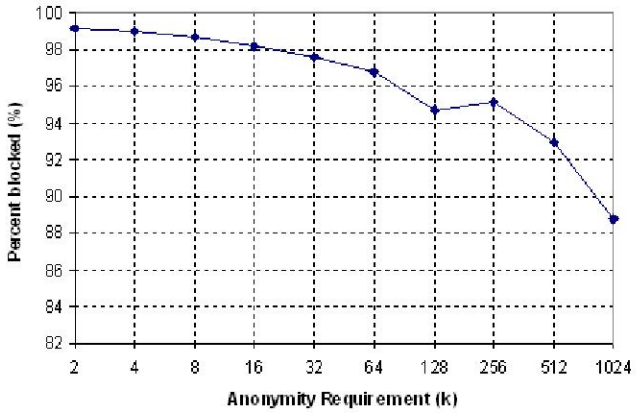

Fig. 3. Blocking efficiency vs. anonymity requirement, k

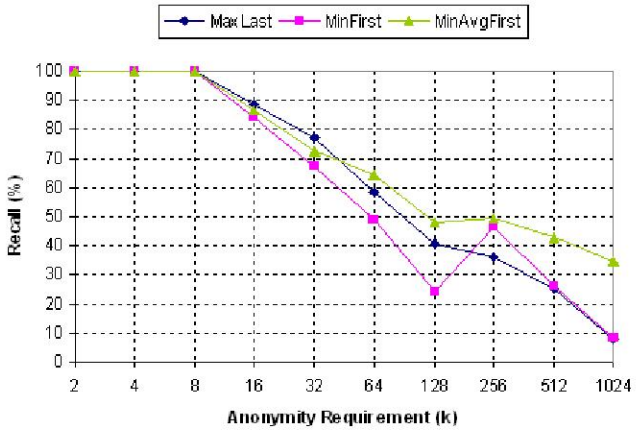

Fig. 4. Recall vs. anonymity requirement, k

C. Matching Thresholds: θ_i

Matching thresholds are among the most import parameters, since they determine the matching record pairs. Yet blocking efficiency does not change in response to varying matching thresholds. According to our analysis, this is because of the distance function applied to discrete attributes, the Hamming distance. When θ_i values are equal to 0.1, all blocked record pairs are blocked on discrete attributes. Since the Hamming distance returns either 0 or 1, reducing matching thresholds does not decrease blocking efficiency.

Increasing θ_i values increases the number of relevant (matching) record pairs. Yet anonymized data sets remain the same and as a result of this selection heuristics feed the SMC protocols in the exact same order in each test case. Evidently, the same record pairs are matched by our method in all test cases. Since the number of relevant record pairs increases and the number of correctly matched record pairs remains constant, recall decreases. We report the experimental results in Figure 5. $MaxLast$ heuristic outperforms the others in this experiment. The average improvement is 4% over $MinAvgFirst$ and 10% over $MinFirst$.

We also conducted experiments for larger values of the matching thresholds. But we choose not to report the detailed results due to space constraints. The results were very predictable: for matching thresholds closer to 1 our method (as any method would) attains excellent recall because almost all record pairs match when thresholds are large.

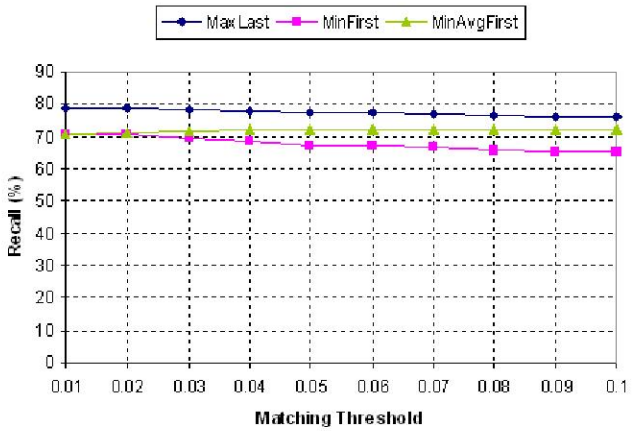

Fig. 5. Recall vs. matching threshold, θ

D. Number of Quasi-identifiers

The set of quasi-identifiers for the *Adult* data set is {*age, work class, education, marital status, occupation, race, sex, native country*}. For the experiment with q quasi-identifiers, we used top-q of the attributes in this set.

Figure 6 reports the blocking efficiency measurements for varying number of quasi-identifiers. Shrinking the quasi-identifier set increases the number of different generalization sequences within anonymized data sets. This implies that smaller groups of records generalized to the same sequence, since the data set sizes are fixed. As the size of the generalization groups decreases, the ratio of record pairs blocked in the *blocking* step decreases.

Figure 7 reports the recall versus the number of quasi-identifiers. As expected, recall increases as more record pairs are labeled in the *blocking* step. $MinFirst$ heuristic has the poorest performance. $MaxLast$ and $MinAvgFirst$ attains around the same recall on average.

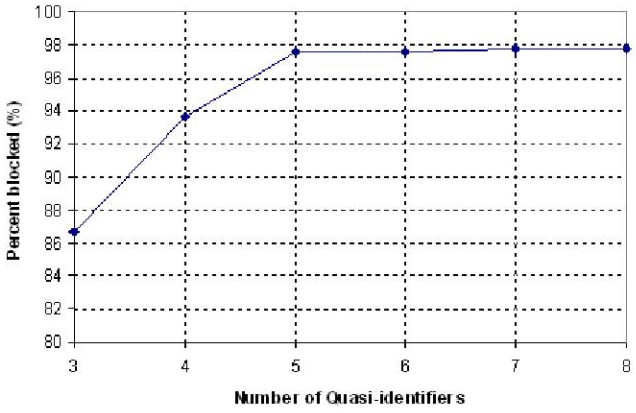

Fig. 6. Blocking efficiency vs. number of QIDs

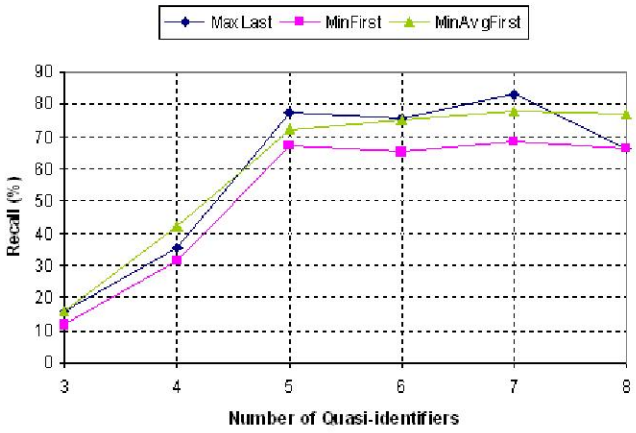

Fig. 7. Recall vs. number of QIDs

E. SMC Allowance

As noted before, $SMC\_allowance$ does not affect blocking efficiency. Around 97.57% of all possible pairs were blocked in all test cases. Notice that recall reaches 100% for $SMC\_allowance$ values larger than 2.33%, since all record pairs that were unlabeled at the end of the *blocking* step can be labeled in the *SMC* step. Figure 8 indicates that recall is very sensitive to $SMC\_allowance$; increasing the number of record pairs matched by SMC protocols improves recall drastically. In this test case none of the heuristics dominates the others.

VII. Related Work

The record linkage problem has been studied for more than five decades since mathematical foundations were established by Fellegi et al. [19] in 1969. Several contributions have been provided in each of the typical phases that compose a record matching process, mainly consisting of preprocessing, search space reduction, and matching decision (see the recent survey [1]). Also some toolkits for dynamically building

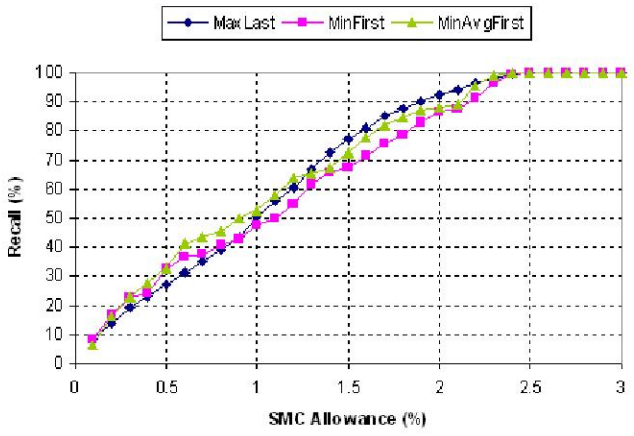

Fig. 8. Varying SMC allowance

record matching workflows have been recently proposed [13], [20].

However, few methods for *private* record matching have been investigated. Some initial approaches are motivated by the strict privacy requirements of e-health applications [3], [4]. The work which is closely related to ours is by Al Lawati et al. [6], that proposes a secure blocking scheme to reduce costs. The approach has the disadvantage to work only for a specific comparison function. Also, as the focus is mainly on efficiency, the effectiveness of the approach has not been assessed. Conversely, our approach can be used with different comparison functions.

In addition to the above approaches, there are two major areas that are closely related to our work, even though no work in such areas addresses our exact problem: secure set intersection and private data sharing.

Several approaches have investigated the secure set intersection problem (see [16] for a survey). Secure set intersection methods deal with exact matching and are too expensive to be applied to large databases due their reliance on cryptography. Furthermore, these protocols deal with the intersection of sets of simple elements, and are not designed for exploiting the semantics behind database records.

Agrawal et al. [15] formalizes a general notion of private information sharing across databases that relies on commutative encryption techniques. This work has opened the way to many other related protocols [21], [22].

Cryptographic methods usually require extensive communication and computation between participants. An alternative that has recently become popular is anonymization.

Anonymization techniques rely on the fact that privacy of sensitive data is a concern only if the individuals related to this data can be identified. However, removing personal identifiers does not always protect individuals against disclosure of identity. The most popular solution to anonymity problem is k-anonymity, which requires that an individual should be indistinguishable from at least $(k-1)$ others in the anonymized data set [8].

The work by Sweeney employs generalization and suppression over a Value Generalization Hierarchy (VHG) in a bottom-up fashion [8]. Iyengar presents a solution using genetic algorithms for increasing the accuracy of classification models, trained on anonymized data sets [23]. Fung et al. proposes the reverse procedure of [7], starting from the most general case and specializing down the VGHs.

Recent work in the area extends the k-anonymity notion. In [24], LeFevre et al. propose multidimensional k-anonymity, where quasi-identifier attributes generalized to different levels of VGH appear together in the anonymized data set. The work in [10] extends k-anonymity to l-diversity, arguing that lack of diversity in sensitive attributes may leak identifying information if the attacker is equipped with background information.

VIII. Conclusion

In this paper, we proposed a novel approach that combines anonymization and cryptographic methods to solve the private record linkage problem. Our method allows participants to trade-off between accuracy, privacy and costs. To the best of our knowledge, ours is the first study in this direction.

As future work, we will extend our existing solution to handle alphanumeric attributes (e.g., address information) as well. We will solve this problem by addressing the following two challenges: distance functions are much more complex than Hamming distance (e.g. edit distance) and there are many possible generalization mechanisms to choose from.

Another promising area of future research might be extending the idea of *hybrid* approaches to other privacy preserving data mining tasks. We believe that the hybrid approach could provide substantial performance improvements for privacy preserving distributed data mining protocols.

References

[1] A. Elmagarmid, G. Panagiotis, and S. Verykios, "Duplicate record detection: A survey," *IEEE Transaction on Knowledge and Data Engineering*, vol. 19, no. 1, 2007.

[2] C. Clifton, M. Kantarcioglu, A. Foan, G. Schadow, J. Vaidya, and A. Elmagarmid, "Privacy-preserving data integration and sharing," in *Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2004.

[3] C. Quantin, H. Bouzelat, F. Allaert, A. Benhamiche, J. Faivre, and L. Dusserre, "How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure," *International Journal of Medical Informatics*, vol. 49, no. 1, 1998.

[4] T. Churces and P. Christen, "Some methods for blindfolded record linkage," *BMC Medical Informatics and Decision Making*, vol. 4, no. 9, 2004.

[5] M. Scannapieco, I. Figotin, E. Bertino, and A. Elmagarmid, "Privacy preserving schema and data matching," in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, 2007.

[6] A. Al-Lawati, D. Lee, and P. McDaniel, "Blocking-aware private record linkage," in *Proceedings of IQIS 2005*, Baltimore, Maryland, 2005.

[7] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *ICDE '05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 205–216.

[8] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557–570, 2002.

[9] R. Agrawal and R. Srikant, "Privacy preserving data mining," *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pp. 439–450, 2000.

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ICDE 2006*, p. 24, 2006.

[11] O. Goldreich, *The Foundations of Cryptography*. Cambridge University Press, 2004, vol. 2, ch. General Cryptographic Protocols. [Online]. Available: http://www.wisdom.weizmann.ac.il/ oded/PSBookFrag/prot.ps

[12] Q. W. H. Kargupta, S. Datta and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," *ICDM 2003*, pp. 96–106, 2003.

[13] M. G. Elfeky, A. K. Elmagarmid, and V. S. Verykios, "TAILOR: A record linkage tool box," in *Proceedings of the 18th International Conference on Data Engineering (ICDE-2002)*, 2002, pp. 17–28.

[14] S. Gomatam, R. Carter, M. Ariet, and G. Mitchell, "An empirical comparison of record linkage procedures," *Statistics in Medicine*, p. 14851496, 2002.

[15] R. Agrawal, A. Evfimievski, and R.Srikant, "Information sharing across private databases," in *Proceedings of ACM SIGMOD 2003*, San Diego, California, 2003.

[16] L. Kissner and D. Song, "Privacy-preserving set operations," in *Advances in Cryptology — CRYPTO 2005*, 2005. [Online]. Available: citeseer.ist.psu.edu/739924.html

[17] D. Newman, S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[18] P. Paillier, "Public key cryptosystems based on composite degree residuosity classes," in *Advances in Cryptology - Eurocrypt '99 Proceedings, LNCS 1592*. Springer-Verlag, 1999, pp. 223–238.

[19] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[20] M. Fortini, M. Scannapieco, L. Tosco, and T. Tuoto, "Towards and open source toolkit to build record linkage workflows," in *SIGMOD International Workshop on Information Quality in Information Systems (IQIS 2006)*, 2006.

[21] F. Emekci, D. Agrawal, A. El Abbadi, and A. Gulbeden, "Privacy preserving query processing using third parties," in *Proceedings of ICDE 2006*, Atlanta, GA, 2006.

[22] R. Agrawal, D. Asonov, M. Kantarcioglu, and Y. Li, "Sovereign joins," in *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*. Washington, DC, USA: IEEE Computer Society, 2006, p. 26.

[23] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2002, pp. 279–288.

[24] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*. Washington, DC, USA: IEEE Computer Society, 2006, p. 25.