Repairing mappings under policy views
Angela Bonifati, Ugo Comignani, Efthymia Tsamoura

To cite this version:
Angela Bonifati, Ugo Comignani, Efthymia Tsamoura. Repairing mappings under policy views. Computing Research Repository, 2019, 10.48550/arXiv.1903.09242. hal-04370079v2

HAL Id: hal-04370079
https://hal.science/hal-04370079v2
Submitted on 3 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT
The problem of data exchange involves a source schema, a target schema and a set of mappings from transforming the data between the two schemas. We study the problem of data exchange in the presence of privacy restrictions on the source. The privacy restrictions are expressed as a set of policy views representing the information that is safe to expose over all instances of the source. We propose a protocol that provides formal privacy guarantees and is data-independent, i.e., if certain criteria are met, then the protocol guarantees that the mappings leak no sensitive information independently of the data that lies in the source. We also propose an algorithm for repairing an input mapping w.r.t. a set of policy views, in cases where the input mapping leaks sensitive information. The empirical evaluation of our work shows that the proposed algorithm is quite efficient, repairing sets of 300 s-t tgds in an average time of 5s on a commodity machine. To the best of our knowledge, our work is the first one that studies the problems of exchanging data and repairing mappings under such privacy restrictions. Furthermore, our work is the first to provide practical algorithms for a logical privacy-preservation paradigm, described as an open research challenge in previous work on this area.

CCS CONCEPTS
• Information systems → Data exchange.

KEYWORDS
privacy-preserving data integration, data exchange, mapping repairs

1 INTRODUCTION
We consider the problem of exchanging data between a source schema S and a target schema T via a set of source-to-target (s-t) dependencies Σ_{st} that usually come in the form of tuple-generating dependencies (tgds). This triple of a source schema, a target schema and a set of dependencies is called a mapping. The s-t dependencies specify how and what source data should appear in the target and are expressed as sentences in first-order logic [10].

Our work considers a privacy-aware variant of the data exchange problem, in which the source comes with a set of constraints, representing the data that is safe to expose to the target over all instances of the source. We also assume that all users, both the malicious and the non-malicious ones, might know the source and the target schema, the data in the target as well as the s-t tgds. Under these assumptions, our work will address the following issues: how could we represent privacy restrictions on the sources and what would it mean for a data exchange setting to be safe under the proposed privacy restrictions?; assuming that the privacy-preservation protocol is fixed, how could we assess the safety of a data exchange setting w.r.t. the privacy restrictions and provide strong guarantees of no privacy leak?; finally, in case of privacy violations, how could we repair the s-t tgds?

Regarding the first issue, we assume that the restrictions on the sources are expressed as a set of views, called policy views. Inspired by prior work on privacy-preservation [3, 13], we define a set of s-t tgds to be safe w.r.t. the policy views if every positive information that is kept secret by the policy views is also kept secret by the s-t tgds. As we will see in subsequent sections, the proposed privacy-preservation protocol is data-independent allowing us to provide strong privacy-preservation guarantees over all instances of the sources. The above addresses the second aforementioned issue, as well. Regarding the third issue, our work proposes a repairing algorithm for the proposed privacy-preservation protocol. The feature of the proposed repairing algorithm is that it can employ techniques for learning the user preferences during the repairing process. The empirical evaluation of our work over an existing benchmark shows that the proposed algorithm is quite efficient. Indeed it can repair a set of 300 s-t tgds in less than 5s on a commodity machine.

Our secure data exchange setting is exemplified in Figure 1 and in the following running example inspired by a real world scenario from an hospital in the UK1.

![Figure 1: A data exchange setting with mappings and policy views.](image)

**Example 1.** Consider the source schema S consisting of the following relations: P, HN, Hs, O and S. Relation P stores for each person registered with the NHS, his insurance number, his name, his ethnicity group and his county. Relations HN and Hs store for each patient who has been admitted to some hospital in the north or the south of UK, his insurance number and the reason for being admitted to the hospital. Relation O stores information related to patients in oncology departments and, in particular, their insurance numbers, their treatment

1https://www.nhs.uk/
and their progress. Finally, relation \( S \) stores for each student in UK, his insurance number, his name, his ethnicity group and his county.

Consider also the set \( V \) comprising the policy views \( V_1 \)–\( V_4 \). The policy views define the information that is safe to make available to public. View \( V_1 \) projects the ethnicity groups and the hospital admittance reasons for patients in the north of UK; \( V_2 \) projects the counties and the hospital admittance reasons for patients in the north of UK; \( V_3 \) projects the treatments and the progress of patients of oncology departments; \( V_4 \) projects the ethnicity groups of the school students. The policy views are safe w.r.t. the NSS privacy preservation protocol. Indeed, the NSS privacy preservation protocol considers as unsafe any non-evident piece of information that can potentially de-anonymize an individual. For example, views \( V_1 \) and \( V_2 \) do not leak any sensitive information, since the results concern patients from a very large geographical area and, thus, the probability of de-anonymizing a patient is significantly small. For similar reasons, views \( V_3 \) and \( V_4 \) are considered to be safe: the probability of de-anonymizing patients of the oncology department from \( V_3 \) is zero, since there is no way to link a patient to his treatment or his progress, while \( V_4 \) projects information which is already evident to public.

\[
P(i, n, e, c) \land H_S(i, d) \rightarrow V_1(e, d) \quad (1) \\
P(i, n, e, c) \land H_S(i, d) \rightarrow V_2(c, d) \quad (2) \\
O(i, t, p) \rightarrow V_3(t, p) \quad (3) \\
S(i, n, e, c) \rightarrow V_4(e) \quad (4)
\]

Finally, consider the following set of s-t dependencies \( \Sigma_{sd} \). The dependencies \( \mu_e \) and \( \mu_c \) project similar information with the views \( V_1 \) and \( V_2 \), respectively. They focus, however, on patients in the north of UK. Finally, the dependency \( \mu_s \) projects the ethnicity groups of students who have been in some oncology department.

\[
P(i, n, e, c) \land H_S(i, d) \rightarrow \text{EthDis}(e, d) \quad (\mu_e) \\
P(i, n, e, c) \land H_S(i, d) \rightarrow \text{CountyDis}(c, d) \quad (\mu_c) \\
S(i, n, e, c) \land O(i, t, p) \rightarrow \text{SO}(e) \quad (\mu_s)
\]

The questions addressed in our paper are the following ones: Are the s-t dependencies safe w.r.t. the policy views? Are there any formal guarantees for privacy preservation in the context of policy views? If the s-t dependencies are not safe w.r.t. the policy views, how could we repair them and provide formal privacy preservation guarantees?

Our technique is inherently data-independent thus bringing the advantage that both the safety test and the repairing operations are executed on the metadata provided through the mappings and not on the underlying data instances. The logical foundations of information disclosure in ontology-based data integration have been laid in [3] in the presence of boolean policies. Instead, we focus on non-boolean policies. Taking a step forward, we also propose an algorithm for repairing a set of unsafe s-t tgds w.r.t. our privacy preservation protocol. To the best of our knowledge, our work is the first to provide practical algorithms for a logical privacy-preservation paradigm, described as an open research challenge in [3, 13]. We leave out probabilistic approaches and anonymization techniques [12, 16], which involve modifications of the underlying data instances and are orthogonal to our approach. A careful treatment of related work is deferred to Section 2.

The source code and the experimental scenarios are publicly available at https://github.com/ucomignani/MapRepair.git.

The paper is organized as follows. Section 2 discusses the related work. Section 3 presents the basic concepts and notions. Section 4 lays our privacy preservation protocol. Section 5 presents our repairing algorithms and their properties. Section 6 outlines the experimental results, while Section 7 concludes our paper.

2 RELATED WORK

Privacy in data integration Safety of secret queries formulated against a global schema and adhering to the certain answers semantics has been tackled in previous theoretical work [13]. They define the optimal attack that characterizes a set of queries that an attacker can issue to which no further queries can be added to infer more information. They then define the privacy guarantees against the optimal attack by considering the static and the dynamic case, the latter corresponding to modifications of the schemas or the GLAV mappings. The same definition of secret queries and privacy setting is adopted in [3], which instead focuses on boolean conjunctive queries as policy views and on the notion of safety with respect to a given mapping. An ontology-based integration scenario is assumed in which the target instance is produced via a set of mappings starting from an underlying data source. Whereas they study the complexity of the view compliance problem in both data-dependent and data-independent setting, we focus on the latter and extend it to non-boolean conjunctive queries as policy views. We further consider multiple policy views altogether in the design practical algorithm for checking the safety of schema mappings and for repairing the mappings in order to resume safety in case of violations.

Privacy in data publishing Data publishing accounts for the settings in which a view exports or publishes the information of an underlying data source. Privacy and information disclosure in data publishing linger over the problem of avoiding the disclosure of the content of the view under a confidential query. A probabilistic formal analysis of the query-view security model has been presented in [12], where they offer a complete treatment of the multi-party collusion and the use of external adversarial knowledge. Access control policies using cryptography are used in [12] to enforce the authorization to an XML document. Our work differs from theirs on both the considered setting, as well as the adopted techniques and the adopted privacy protocol.

Directed Query Evaluation Directed Query Evaluation is a confidentiality enforcement framework introduced in [15] and refined in [7],[5] and [6], in which a policy declaratively specifies sensitive information and confidentiality is enforced by a censor. Provided a query as input, a censor verifies whether the query leads to a violation of the policy and in case of a violation it returns a distorted answer. It has been recently adopted in ontologies expressed with Datalog-like rules and in lightweight Description Logics [11]. They assume that the policies are only known to database administrators and not to ordinary users and that the data has protected access through a query interface. Our assumptions and setting are quite different, since our multiple policy views are accessible to every user and our goal is to render the s-t mappings safe with respect to a set of policies via repairing and rewriting.
3 PRELIMINARIES

Let Const, Nulls, and Vars be mutually disjoint, infinite sets of constant values, labeled nulls, and variables, respectively. A schema is a set of relation names (or just relations), each associated with a nonnegative integer called arity. A relational atom has the form \( R(\bar{t}) \) where \( R \) is an \( n \)-ary relation and \( \bar{t} \) is an \( n \)-tuple of terms, where a term is either a constant, a labelled null, or a variable. An equality atom has the form \( t_1 = t_2 \) where \( t_1 \) and \( t_2 \) are terms. An atom is called ground or fact, when it does not contain any variables. A variable is a substitution from variables into constants or labeled nulls. A conjunction of relational, null-free atoms. We usually omit the positional nulls.

A Substitution \( \sigma \) is a mapping from variables into constants or labelled nulls. A Dependency describes the semantic relationship between relations. A Tuple Generating Dependency (tgd) is a formula of the form \( \forall \bar{x} \lambda(\bar{x}) \rightarrow \exists y \rho(\bar{x}, y) \), where \( \lambda(\bar{x}) \) and \( \rho(\bar{x}, y) \) are conjunctions of relational, null-free atoms. An Equality Generating Dependency (egd) is a formula of the form \( \forall x \lambda(x) \rightarrow x = y \), where \( \lambda(x) \) is a conjunction of relational, null-free atoms. We usually omit the quantification for brevity. We refer to the left-hand side of a tgd or an egd \( \delta \) as the body, denoted as \( \text{body}(\delta) \), and to the right-hand side as the head, denoted as \( \text{head}(\delta) \). An instance \( I \) satisfies a dependency \( \delta \), written \( I \models \delta \) if each homomorphism from \( \text{body}(\delta) \) into \( I \) can be extended to a homomorphism \( h^\prime \) from \( \text{head}(\delta) \) into \( I \). An instance \( I \) satisfies a set of dependencies \( \Sigma \), written as \( I \models \Sigma \), if \( I \models \delta \) holds, for each \( \delta \in \Sigma \). The Solutions of an instance \( I \) w.r.t. \( \Sigma \) is the set of all instances \( J \) such that \( J \supseteq I \) and \( J \models \Sigma \). A solution is called universal if it can be homomorphically embedded to each solution of \( I \) w.r.t. \( \Sigma \).

A conjunctive query (CQ) is a formula of the form \( \exists y \bigwedge_i A_i \), where \( A_i \) are relational, null-free atoms. A CQ is boolean if it does not contain any free variables. A substitution \( \sigma \) is an answer to a CQ \( Q \) on an instance \( I \) if the domain of \( \sigma \) is the free variables of \( Q \), and if \( \sigma \) can be extended to a homomorphism from \( \bigwedge_i A_i \) into \( I \). We denote by \( Q(I) \), the answers to \( Q \) on \( I \).

Given an instance \( I \) and a set of dependencies \( \Sigma \), the chase iteratively computes a universal solution of \( I \) w.r.t. \( \Sigma \) [4, 10]. Starting from \( I_0 = I \), at each iteration \( i \), it computes a new instance \( I_i \) by applying a tgd or an egd chase step:

**tgd chase step.** Consider an instance \( I_i \), a tgd \( \delta \) of the form \( \forall \bar{x} \lambda(\bar{x}) \rightarrow \exists y \rho(\bar{x}, y) \) and a homomorphism \( h \) from \( \lambda(\bar{x}) \) into \( I_i \), such that there does not exist an extension of \( h \) to a homomorphism from \( \rho(\bar{x}, y) \) into \( I_i \). Such homomorphisms are called active triggers. Applying the tgd chase step for \( \delta \) and \( h \) to \( I_i \) results in a new instance \( I_{i+1} = I_i \cup h'(\rho(\bar{x}, y)) \), where \( h' \) is a substitution such that \( h'(x_j) = h(x_j) \) for each variable \( x_j \in \bar{x} \) and \( h'(y_j) \), for each \( y_j \in y \), is a fresh labeled null not occurring in \( I_i \).

**egd chase step.** Consider an instance \( I_i \), an egd \( \delta \) of the form \( \forall x \lambda(x) \rightarrow x = y \) and a homomorphism \( h \) from \( \lambda(x) \) into \( I_i \), such that \( h(x_i) \neq h(x_j) \). Applying the egd chase step for \( \delta \) and \( h \) to \( I_i \) fails if \( (h(x_i), h(x_j)) \subseteq \text{Const} \), and otherwise it results in a new instance \( I_{i+1} = v(I_i) \), where \( v = \{ h(x_i) \mapsto h(x_j) \} \) if \( h(x_i) \in \text{Const} \), and \( v = \{ h(x_i) \mapsto h(x_j) \} \) if \( h(x_i) \notin \text{Const} \). We denote by chase(\( I, \Sigma \)), the chase of \( I \) w.r.t. \( \Sigma \).

Let \( S \) be a source schema and let \( T \) be a target schema. A mapping \( M \) from \( S \) to \( T \) is defined as a triple (\( S, T, \Sigma \)), where, generally, \( \Sigma = \Sigma_s \cup \Sigma_{st} \cup \Sigma_t \). \( \Sigma_s \) and \( \Sigma_t \) denote the source, s-t dependencies over \( S \) and \( T \), respectively. We usually refer to the dependencies in \( \Sigma_{st} \) as mappings. A variable \( x \) of a mapping \( \mu \in \Sigma_{st} \) is exported if it occurs both in the body and the head of \( \mu \). We denote by \( \text{exported}(\mu) \) the set of exported variables of \( \mu \). The inverse of set of s-t dependencies \( \Sigma_{st} \), denoted as \( \Sigma_{st}^{-1} \) is the set consisting, for each mapping \( \mu \in \Sigma_{st} \), of the form \( \lambda(\bar{x}) \rightarrow \rho(\bar{x}, y) \), a mapping \( \mu^{-1} \) of the form \( \rho(\bar{x}, y) \rightarrow \lambda(\bar{x}) \). In this paper, we will focus on the scenario, where \( \Sigma_s \) and \( \Sigma_t \) are empty, so \( \Sigma \) will be equal to \( \Sigma_{st} \). Furthermore, in this paper we focus on GLAV mappings, i.e., s-t dependencies corresponding to a set of views.

The certain answers of a CQ \( Q \) over \( T \) w.r.t. \( I \) and \( M \), denoted as certain(\( Q, I, M \)) are the intersection of all answers to \( Q \) over all solutions of \( I \) w.r.t. \( \Sigma \). Given a finite, null-free instance \( I \) of the source schema, the objective of data exchange is to compute a universal solution of \( I \) w.r.t. the dependencies \( \Sigma \) from \( M \).

4 PRIVACY PRESERVATION

In this section, we introduce our notion of privacy preservation. Let \( V \) be a set of policy views over \( S \). The policy views represent the information that is safe to expose for instances \( I \) of \( S \). We denote by \( M_V = (S, V, V') \) the mapping from \( S \) to \( V \), where \( V \) denotes the schema of the views occurring in \( V \). Our goal is to verify whether a user-defined mapping \( M = (S, T, \Sigma) \) is safe w.r.t. a view mapping \( M_V \). Below, we will introduce a notion for assessing the safety of a GAV mapping \( M \), with respect to a GAV mapping \( M_V \), when both make use of the same source schema \( S \). Below, let \( \Sigma_{st} = \Sigma_{st1} \), be the dependencies associated with \( M_V \).

4.1 A formal privacy-preservation protocol

Our notion of privacy preservation builds upon the protocol introduced in [3]. Below, we formalize the notion of privacy preservation from [3] and we extend it for non-boolean conjunctive queries. First we recapitulate the notion of indistinguishability of two source instances.

**Definition 1.** Two instances \( I \) and \( I' \) of a source schema \( S \) are indistinguishable with respect to a mapping \( M = (S, T, \Sigma) \), denoted as \( I \equiv_M I' \), if certain(\( Q, I, M \) = certain(\( Q, I', M \) for each CQ \( Q \) over \( T \).

Informally, Definition 1 tells us that two source instances are indistinguishable from each other if the instance targets have the same certain answers.
Definition 2. A mapping \( M = (S, T, \Sigma) \) does not disclose a CQ \( p \) over \( S \) on any instance of \( S \), if for each instance \( I \) of \( S \) there exists an instance \( I' \) such that \( I \equiv M I' \) and \( p(I') = 0 \).

The problem of checking whether a mapping \( M \) over \( S \) does not disclose a boolean and constants-free CQ \( p \) on any instance of \( S \) is decidable for GAV mappings consisting of CQ views [3]. In particular, \( M \) does not disclose \( p \) on any instance of \( S \) if and only if there does not exist a homomorphism from \( p \) into the unique instance computed by the visible chase \( \text{visChase}_S(\Sigma) \) of \( \Sigma \) under the critical instance \( Crt_S \) of \( S \). The visible chase computes a universal source instance — that is an instance, such that the visible part of any instance of \( S \) (i.e., the subinstance that becomes available through the mappings) can be mapped into it. The only constant occurring in the instance computed by \( \text{visChase}_S(\Sigma) \) is the critical constant * and it represents any other constant that can occur in the source instance.

For the purpose of repairing the mappings efficiently, we introduce our own variant of the visible chase, which organizes the facts derived during chasing into subinstances called bags. Algorithm 1 describes the steps of the proposed variant. Please note that Algorithm 1 derives the same set of facts with the algorithm from [3]. However, instead of keeping these facts in a single set, we keep them in separate bags. Before presenting Algorithm 1, we will introduce some new notions.

Definition 3. Consider an instance \( I \). Consider also a s-t tgd \( \delta \) and a homomorphism \( h \) from body(\( \delta \)) into \( I \), such that \( h(x) \in \text{nulls} \), for some \( x \in \text{exported}(\delta) \). Then, we say that the egd

\[ \text{body}(\delta) \rightarrow \bigwedge_{x \in \text{exported}(\delta)} x \equiv * \]  

is derived from \( \delta \) in \( I \). For an egd \( \epsilon \) that is derived from a s-t tgd \( \delta \) in \( I \), \( \text{tgд}(\epsilon) \) denotes \( \delta \). For a set of s-t tgd of \( \Sigma \) and an instance \( I \), \( \Sigma \) is the set comprising for each \( \delta \in \Sigma \), the egd that is derived from \( \delta \) in \( I \).

Definition 4. Consider an instance \( I \), whose facts are organized into the bags \( \beta_1, \ldots, \beta_m \). Consider also a derived egd \( \delta \) of the form (5) and an active trigger \( h \) for \( \delta \) in \( I \). A bag \( \beta_i \) is relevant for \( \delta \) and \( h \) in \( I \), where \( 1 \leq i \leq n \), if some fact \( F \in \text{body}(\delta) \) occurs in \( \beta_i \) and if some \( h(x) \) is a labeled null occurring in \( \beta_i \), where \( x \in \text{exported}(\delta) \).

Let \( \beta_{1i} \leq \beta_{2i} \leq \beta_{mi} \) be the set of bags that are relevant for \( \delta \) and \( h \) in \( I \). Let \( \nu = \{ h(x_j) \mapsto h(x_i) \} \) if \( h(x_i) \equiv * \), and \( \nu = \{ h(x_j) \mapsto h(x_i) \} \) if \( h(x_i) \equiv \) Const, where \( x_i, x_j \) are variables exported(\( \delta \)). Then, the derived bag \( \beta \) for \( \delta \) and \( h \) in \( I \) consists of the facts in \( \bigcup_{i=1}^{k} \nu(\beta_{hi}) \). The bags \( \beta_{1i}, \ldots, \beta_{ki} \) are called the predecessors of \( \beta \). We use \( \beta_{hi} \prec \beta \) to denote that \( \beta_{hi} \) is a predecessor of \( \beta \), for \( 1 \leq i \leq k \).

We are now ready to proceed with the description of Algorithm 1. Given a s-t mapping, Algorithm 1 computes a universal source instance whose facts are organized into bags. Algorithm 1 first computes the instance \( I_0 \) by chasing \( Crt_S \) using the s-t tgd, line 1. It then processes \( I_0 \) with the inverse s-t tgd \( \Sigma^{-1} \), line 2, and proceeds by chasing \( I_1 \) with the set of all derived egds \( \Sigma \), for each \( \delta \in \Sigma \) in \( I_1 \), line 4. Algorithm 1 computes a fresh bag at each chase step. In particular, for each active trigger \( h \) for \( \delta \) in \( I \), Algorithm 1 adds a fresh bag with facts \( h'(\text{head}(\delta)) \), if \( \delta \in \Sigma \cup \Sigma^{-1} \), line 9; otherwise, if \( \delta \in \Sigma \), then it adds the derived bag for \( \delta \) and \( h \) in \( I \), see Definition 4, line 20.

Note that, \( \Sigma \) aims at “disambiguating” as many labeled nulls occurring in \( I_1 \) as possible, by unifying them with the critical constant * . Since * represents the information that is “visible” to a third-party, chasing with \( \Sigma \) computes the maximal information from the source instance a third-party has access to. Note that Algorithm 1 always terminates [3]. Let \( B = \text{visChase}_S(\Sigma) \). We will denote by \( I_S(\Sigma) \), the instance \( \bigcup_{\beta \in B} \beta \).

\[ \text{Algorithm 1 visChase}_S(\Sigma) \]

1: \( B_0 = \text{bagChaseTGD}(\Sigma, Crt_S) \)
2: \( B_1 = \text{bagChaseTGD}(\Sigma^{-1}, (\bigcup_{\beta \in B} \beta) \setminus Crt_S) \)
3: Let \( \Sigma_0 \) be the set of all derived egds \( \Sigma_0 \), for each \( \delta \in \Sigma \) in \( I_1 \)
4: return \( \text{bagChaseEGR}(\Sigma_0, B_0 \cup B_1) \)
5: procedure \( \text{bagChaseTGD}(\Sigma, I) \)
6: \( B = \emptyset \)
7: for each \( \delta \in \Sigma \) do
8: for each active trigger \( h : \text{body}(\delta) \rightarrow I \) do
9: create a fresh bag \( \beta \) with facts \( h'(\text{head}(\delta)) \)
10: add \( \beta \) to \( B \)
11: return \( B \)
12: procedure \( \text{bagChaseEGR}(\Sigma_0, \beta) \)
13: \( i = 0 \)
14: \( I_1 = \bigcup_{\beta \in \beta} \beta \)
15: do
16: for each \( \delta \in \Sigma_0 \) of the form \( \delta \) do
17: for each active trigger \( h : \text{body}(\delta) \rightarrow I \) do
18: if \( h(x) \neq * \), for some \( x \in \text{exported}(\delta) \) then
19: Let \( \beta \) be the derived bag for \( \delta \) and \( h \) in \( I_{1-1} \)
20: add \( \beta \) to \( B \)
21: \( I_i = I_{i-1} \cup \beta \)
22: while \( I_{i-1} \neq I_i \)
23: return \( B \)

Example 2. We demonstrate the visible chase algorithm over the policy views and the s-t dependencies from Example 1.

We first present the computation of \( I_S(\mathcal{V}) = \bigcup_{\beta \in \text{visChase}_S(\mathcal{V})} \beta \).

The critical instance \( Crt_S \) of \( S \) consists of the facts shown in Eq. (6).

\[ P(*, *, *, *) \quad H_b(*, *) \quad H_s(*, *, *) \quad O(*, *, *) \quad S(*, *, *, *) \]

where * is the critical constant.

The instance \( I_1 \) computed by chasing the output of line 1 using \( \mathcal{V}^{-1} \) will consist of the facts

\[ P(n_i, n_n, n_c) \quad H_b(n_n, *) \quad O(n_c^*, *, *) \quad L_1 \]

\[ P(n_i^*, n_n^*, n_c^*) \quad H_s(n_c^*, *) \quad S(n_c^*, n_r^*, n_c^*, n_r^*) \]

where the constants prefixed by \( n \) are labeled nulls created while chasing \( Crt_S \) with the inverse mappings. Since there exists no homomorphism from the body of any s-t tgd into \( I_1 \) mapping an exported variable into a labeled null, \( \Sigma_0 \) will be empty, see Definition 3. Thus, \( I_S(\mathcal{V}) = I_1 \).

We next present the computation of \( I_S(\mathcal{V}^{\ast}) = \bigcup_{\beta \in \text{visChase}_S(\mathcal{V}^{\ast})} \beta \).

The instance \( I_1' \) computed by chasing the output of line 1 by \( \Sigma_0 \) will consist of the facts

\[ P(n_i, n_n, n_c) \quad H_b(n_n, *) \quad S(n_c^*, n_r^*, n_c^*, n_r^*) \quad L_1' \]

\[ P(n_i^*, n_n^*, n_c^*) \quad H_s(n_c^*, *) \quad O(n_c^*, n_r^*, n_c^*) \]

Since there exists a homomorphism from the body of \( \mu_e \) into \( I_1' \), mapping the exported variable \( e \) into the labeled null \( n_e \), and since there

A. Bonifati, U. Comignani, and E. Tsamoura
exists another homomorphism from the body of $\mu_c$ into $l'_1$, mapping the exported variable $c$ into the labeled null $n_e$. Let $\Sigma_n$ will comprise the edge $e_1$ and $e_2$ shown below:

$P(i, n, e, c) \land H_n(i, d) \rightarrow e = *$ (e₁)

$P(i, n, e, c) \land H_n(i, d) \rightarrow c = *$ (e₂)

The last step of the possible chase involves chasing $l'_2$ using $\Sigma_n$. WLOG, assume that the chase considers first $e_1$ and then $e_2$. During the first step of the chase, there exists a homomorphism from body($e_1$) into $l'_1$. Hence, $n_e = *$. During the second step of the chase, there exists a homomorphism from body($e_2$) into $l'_1$ and, hence, $n_c = *$. The instance computed at the end of the second round of the chase will consist of the facts:

$P(n_i, n_e, *, e) \land H_n(n_i, *) H_n(n'_i, *)$ (7)

$S(n''_i, n''_e, *, n''_c) \land O(n''_i, n''_e, n''_c)$

Since there exists no active trigger for $e_1$ or $e_2$ in the instance of Eq. (7), the chase will terminate.

The facts in $I_3(\Sigma_{st})$ will be organized into the following bags $P_1\ldots P_7$ (one bag per line)

SO(e) $\xrightarrow{(\mu_i', h_i)} S(n''_i, n''_e, *, n''_c), O(n''_i, n''_e, n''_c)$

CountyDis(c, d) $\xrightarrow{(\mu_i', h_i)} P(n_i, n_e, *, n_c), H(n_i, *)$

EthDis(e, d) $\xrightarrow{(\mu_i', h_i)} P(n_i, n_e, *, n_c), H(n_i, *)$

$P(n_i, n_e, *, n_c), H(n_i, *) \xrightarrow{(\epsilon_1, h_i)} P(n_i, n_e, *, n_c), H(n_i, *)$

The contents of the bags correspond to the right-hand side of the arrows. However, for presentation purposes, we also show the related dependency $\delta$ and the homomorphism $h$ that lead to the derivation of each bag (shown at the top of each arrow), as well as the facts, in $h(body(\delta))$ (left-hand side of each arrow).

### 4.2 Preserving the privacy of policy views

We consider a mapping consisting of CQ views $M = (S, T, \Sigma)$ to be safe w.r.t. a view mapping consisting of CQ views $M_V = (S, V, \Sigma_V)$, if $M$ does not disclose the information that is also not disclosed by $M_V$. Definition 5 and Theorem 1 presented below formalize our notion of privacy preservation and show that there exists a simple process for verifying whether $M$ is safe w.r.t. $M_V$.

**Definition 5.** A mapping $M_2 = (S, T_2, \Sigma_2)$ preserves the privacy of a mapping $M_1 = (S, T_1, \Sigma_1)$ on all instances of $S$, if for each constants-free CQ $p$ over $S$, if $p$ is not disclosed by $M_1$ on any instance of $S$, then $p$ is not disclosed by $M_2$ on any instance of $S$.

**Theorem 1.** A mapping $M_2 = (S, T_2, \Sigma_2)$ preserves the privacy of a mapping $M_1 = (S, T_1, \Sigma_1)$ on all instances of $S$, if and only if there exists a homomorphism $h$ from $I_3(\Sigma_2)$ into $I_3(\Sigma_1)$, such that $h(\epsilon) = *$.

**Proof.** (Sketch) First we show that the following holds

**Lemma 1.** A mapping $M = (S, T, \Sigma)$ does not disclose a constants-free CQ $p$ over $S$ on any instance of $S$, iff $\epsilon \notin p(I)$, where $J = I_3(\Sigma_{st})$.

**Proof.** By adapting the proof technique of Theorem 16 from [3], we can show that $J = I_3(\Sigma_{st})$ is a universal source instance $I_3(\Sigma)$ satisfying the following property: for each pair of source instances $I$ and $I'$ such that $I'$ is indistinguishable from $I$ w.r.t. the mapping $M$, there exists a homomorphism $h$ from $I'$ into $I_3(\Sigma)$ mapping each schema constant into the critical constant $\epsilon$. Due to the existence of a homomorphism $h$ from $I'$ into $I_3(\Sigma)$, for each pair of indistinguishable source instances $I$ and $I'$, we can see that if $\epsilon \notin p(J)$ for a constants-free CQ $p$, then $p(I') = \emptyset$. Due to the above and due to Definition 2, it follows that $M = (S, T, \Sigma)$ does not disclose a constants-free CQ $p$ over $S$ on any instance of $S$.

**Lemma 1** states that, in order to check if a constants-free CQ $p$ is safe according to Definition 2, we need to check if the critical tuple is among the answers to $p$ over the instance computed by visChases($\Sigma$). Next, we show the following lemma.

**Lemma 2.** Given two instances $I_1$ and $I_2$, the following are equivalent:

1. For each CQ $p$, if $\epsilon \notin p(I_1)$, then $\epsilon \notin p(I_2)$, where $\epsilon$ is a vector of constants.

2. There exists a homomorphism from $I_1$ to $I_2$ preserving the constants of $I_1$.

**Proof of Lemma 2.** (2)⇒(1). Suppose that there exists a homomorphism $h$ from $I_1$ to $I_2$ preserving the constants of $I_1$. Suppose also that $\epsilon \notin p(I_1)$, with $p$ being a CQ. This means that there exists a homomorphism $h_1$ from $p$ into $I_1$ mapping each free variable $x_i$ of $p$ into $u_i$, for each $1 \leq i \leq n$, where $n$ is the number of free variables of $p$. Since the composition of two homomorphisms is a homomorphism and since $h$ preserves the constants of $I_1$ due to the base assumptions, this means that $h \circ h_1$ is a homomorphism from $p$ into $I_2$ mapping each free variable $x_i$ of $p$ into $u_i$, for each $1 \leq i \leq n$. This completes this part of the proof.

(1)⇒(2). Let $p_1$ be a CQ formed by creating a non-ground atom $R(y_1, \ldots, y_n)$ for each ground atom $R(u_1, \ldots, u_n) \in I_1$, by taking the conjunction of these non-ground atoms and by converting into an existentially quantified variable every variable created out of some labelled null. Let $\bar{x}$ denote the free variables of $p_1$ and let $n = |\bar{x}|$. From the above, it follows that there exists a homomorphism $h_1$ from $p_1$ into $I_1$ mapping each $x_i$ into some constant occurring in $I_1$. Let $\epsilon \notin p(I_1)$, and hence, there exists a homomorphism $h_2$ from $p_1$ into $I_2$ mapping each $x_i$ into $u_i$, for each $1 \leq i \leq n$. Since $h_1$ ranges over all constants of $I_1$ and since $h_1(x_i) = h_2(x_i)$ holds for each $1 \leq i \leq n$, it follows that there exists a homomorphism from $I_1$ to $I_2$ preserving the constants of $I_1$. This completes the second part of the proof.
Lemma 2 can be restated as follows

**Lemma 3.** Given two instances $I_1$ and $I_2$, the following are equivalent
(1) for each CQ $p$, if $\bar{t} \notin p(I_1)$, then $\bar{t} \notin p(I_2)$
(2) there exists a homomorphism from $I_1$ to $I_2$

We are now ready to return to the main part of the proof. Given a CQ $p$ over a source schema $S$, and a mapping $M$ defined as the triple $(S, T, \Sigma)$, where $T$ is a target schema and $\Sigma$ is a set of $s$-$t$ dependencies, we know from Proposition 1 that if $M$ discloses $p$ on some instance of $S$, then there exists a homomorphism of $p$ into $\text{visChase}_S(\Sigma)$ mapping the free variables of $p$ into the critical constant $\ast$.

From the above, we know that $M_2$ does not preserve the privacy of $M_1$ if there exists a CQ $p$ over $S$, such that $\bar{t} \notin p(I_1)$ and $\bar{t} \in p(I_2)$, where $J_1 = I_3(\Sigma_1)$ and $J_2 = I_3(\Sigma_2)$. We will now prove that $M_2$ preserves the privacy of $M_1$ iff there exists a homomorphism from $J_2$ into $J_1$ that preserves the critical constant $\ast$. This will be referred to as conjecture (C).

$(\Rightarrow)$ If $M_2$ preserves the privacy of $M_1$, then for each CQ $p$, if $\bar{t} \notin p(I_1)$, then $\bar{t} \notin p(I_2)$. From the above and from Lemma 3, it follows that there exists a homomorphism $\phi : J_2 \rightarrow J_1$, such that $\phi(\ast) = \ast$.

$(\Leftarrow)$ The proof proceeds by contradiction. Assume that there exists a homomorphism $h$ from $J_2$ into $J_1$ preserving $\ast$, but $M_2$ does not preserve the privacy of $M_1$. We will refer to this assumption as assumption (A1). From assumption (A1) and the discussion above it follows that there exists a CQ $p$ over $S$ such that $\bar{t} \notin p(I_1)$ and $\bar{t} \in p(I_2)$. Let $h_2$ be the homomorphism from $p$ into $J_2$ mapping its free variables into $\ast$. Since the composition of two homomorphisms is a homomorphism, this means that $h \circ h_2$ is a homomorphism from $p$ into $J_1$ mapping its free variables into $\ast$, i.e., $\bar{t} \in p(I_1)$. This contradicts our original assumption and hence concludes the proof of conjecture (C). Conjecture (C) witnesses the decidability of the instance-independent privacy preservation problem: in order to verify whether $M_2$ preserves the privacy of $M_1$ we only need to check if there exists a homomorphism $\phi : I_3(\Sigma_2) \rightarrow I_3(\Sigma_1)$, such that $\phi(\ast) = \ast$. □

Theorem 1 states that in order to verify whether $M_2$ is safe w.r.t. $M_1$, we need to compute $I_3(\Sigma_1)$ and $I_3(\Sigma_2)$ and check if there exists a homomorphism from the second instance into the first one that maps $\ast$ into itself. If there exists such a homomorphism, we say that $I_3(\Sigma_1)$ is safe w.r.t. $I_3(\Sigma_2)$, or simply safe, and we say that it is unsafe otherwise.

**Example 3.** Continuing with Example 1, we can see that the $s$-$t$ tgds are not safe w.r.t. the policy views according to Theorem 1, since there does not exist a homomorphism from the instance $I_3(\Sigma_{st})$ into the instance $I_3(\Sigma_1)$. This means that there exists information which is disclosed by $\Sigma_{st}$, in some instance that satisfies $\Sigma_{st}$, but it is not disclosed by $\Sigma$. Indeed, from $S(n''_1, n'_2, \ast, n''_3)$ and $O(n''_1, n''_2, n''_3)$, we can see that we can potentially leak the identity of a student who has been to an oncology department. This can happen if there exists only one student in the school coming from a specific ethnicity group and this ethnicity group is returned by $\mu_e$. Please note that the policy views are safe w.r.t. this leak. Indeed, it is impossible to derive this information through reasoning over the returned tuples under the input instance and the views $V_3$ and $V_4$.

Furthermore, by looking at the facts $P(n_1, n_2, \ast, \ast)$ and $H_{n_2}(n_1, \ast)$, we can see that we can potentially leak the identity and the disease of a patient who has been admitted to some hospital in the north of UK. This can happen if there exists only one patient who relates to the county and the ethnicity group returned by $\mu_e$ and $\mu_c$. Note that the policy views $V_1$ and $V_2$ do not leak this information, since it is impossible to obtain the county and the ethnicity group of an NHS patient at the same time.

5 REPAIRING UNSAFE MAPPINGS

In Section 4 we presented our privacy preservation protocol and a technique for verifying whether a mapping is safe w.r.t. another one, over all source instances. This section presents an algorithm for repairing an unsafe mapping $M$ w.r.t. a set of policy views $\mathcal{V}$.

Algorithm 2 summarizes the steps of the proposed algorithm. The inputs to it are, apart from $\Sigma$ and $\mathcal{V}$, a positive integer $n$ which will be used during the second step of the repairing process and a preference mechanism $\text{prf}$ for ranking the possible repairs. In the simplest scenario, the preference mechanism implements a fixed function for ranking the different repairs. However, it can also employ supervised learning techniques in order to progressively learn the user preferences by looking at his prior decisions.

Since a mapping $M$ is safe w.r.t. $\mathcal{V}$ if the instance $I_3(\Sigma)$ is safe according to Theorem 1, Algorithm 2 rewrites the tgds in $M$, such that the derived visible chase instances are safe. The rewriting takes place in two steps. The first step rewrites $\Sigma$ into a partially-safe set of $s$-$t$ dependencies $\Sigma_1$, while the second step rewrites the output of the first one into a new set of $s$-$t$ dependencies $\Sigma_2$, such that $I_3(\Sigma_2)$ is safe. As we will explain later on, partial-safety ensures that the intermediate instance $I_3$ produced by $\text{visChase}_S(\Sigma_1)$ at line 2 of Algorithm 1 is safe, but it does not provide strong privacy guarantees. The benefit of this two-step approach is that it allows repairing one or a small set of dependencies at a time.

5.1 Computing partially-safe mappings

Since the problem of safety is reduced to the problem of checking for a homomorphism from $I_3(\Sigma)$ into $I_3(\mathcal{V})$, a first test towards checking for such a homomorphism is to look if the mappings in $\Sigma$ would lead to such a homomorphism or not. For instance, by looking at $\mu_e$ in Example 1 it is easy to see that it leaks sensitive information, since it involves a join between students and oncology departments, which does not occur in $I_3(\mathcal{V})$.

**Definition 6.** A mapping $M = (S, T, \Sigma)$ is partially-safe w.r.t. $\mathcal{M}_V = (S, \mathcal{V})$ on all instances of $S$, if there exists a homomorphism from $\text{chase}(\Sigma^{-1}, \text{Crt}_{\text{fg}}) \setminus \text{Crt}_{\text{fg}}$ into $I_3(\mathcal{V})$. □

From Algorithm 1, it follows that $\Sigma$ is partially-safe iff the intermediate instance $I_1$ computed by $\text{visChase}_S(\Sigma)$ is safe.

**Proposition 1.** A mapping $M = (S, T, \Sigma)$ is partially-safe w.r.t. $\mathcal{M}_V = (S, \mathcal{V})$ on all instances of $S$, if for each $\mu \in \Sigma$, there exists a

\begin{algorithm}
\caption{repair($\Sigma, \mathcal{V}, \text{prf}, n$)}
1: $I_1 := \text{frepair}(\Sigma, \mathcal{V}, \text{prf})$
2: $I_2 := \text{repair}(I_1, \mathcal{V}, \text{prf}, n)$
3: return $I_2$
\end{algorithm}
Repairing mappings under policy views

homomorphism from body(μ) into I_S(V) mapping each x ∈ exported(μ) into the critical constant ∗.

Note that according to Proposition 1, in our running example Σ_{st} would be partially-safe, if μ_s ∈ Σ_{st}, then since there exist homomorphisms from the bodies of μ_s and μ_e into I_S(V), mapping their exported variables into ∗. It is also easy to show the following

Remark 1. A mapping M = (S, T, Σ) is safe w.r.t. M_V = (S, V, V) on all instances of S, only if it is partially-safe w.r.t. M_V on all instances of S.

Proposition 1 presents a quite convenient, yet somewhat expected, finding: in order to obtain a partially-safe mapping, it suffices to repair each s-t dependency independently of the others. Furthermore, the repair of each μ ∈ Σ involves breaking joins and hiding exported variables, such that the repaired dependency μ_r satisfies the criterion in Proposition 1.

We make use of the result of Proposition 1 in Algorithm 3. Algorithm 3 obtains, for each μ ∈ Σ, a set of rewritings R_μ, out of which we will choose the best rewriting according to prf. The set R_μ consists of all rewritings that differ from μ w.r.t. the variable repetitions in the bodies of the rules and the exported variables. For performance reasons, we do not examine rewritings that introduce atoms in the bodies of the rules. However, this does not compromise the completeness of Algorithm 2 as we show at end of this section. Below, we present the steps of Algorithm 3.

For each s-t tdg μ and for each atom B ∈ body(μ), Algorithm 3 constructs a fresh atom C and adds C to a set Σ. The set of atoms C provides us with the means to identify all repairs of μ that involve breaking joins and hiding exported variables. In particular, each homomorphism ξ from C into I_S(V) corresponds to one repair of μ. In lines 12–25, Algorithm 3 modifies each atom B ∈ body(μ) by taking into account prior body atom modifications. The prior modifications are accumulated in the relation ρ and the mapping ψ. The relation ρ keeps for each variable x from body(μ), the fresh variables that were used to replace x during prior steps of the repairing process, while ψ is a substitution from the partially repaired body into I_S(V). In particular, at the end of the i-th iteration of the loop in line 12, ψ holds the substitution from the first repaired i atoms from body(μ) into I_S(V). We adopt this approach instead of replacing variable x in position ρ always by a fresh variable, in order to minimize the number of the joins we break.

Below, we describe how Algorithm 3 modifies each body atom of μ, w.r.t. a homomorphism ξ, lines 9–27. Let C = v(B) be the fresh body atom that was constructed out of B in line 5. For each atom B ∈ body(μ) and for each p ∈ pos(B), if the variable y in position p of C is not mapped to the critical constant ∗ via ξ and B_{yp} is a exported variable, this means that the variable sitting in position p of B should not be exported (see first condition in line 16). Similarly, if the variable sitting in position p of B is mapped to a different constant than the one that y maps via ξ, then this means that the variable sitting in position p of B introduces an unsafe join (see second condition in line 16). In the presence of these violations, we must replace variable x in position p of B, either by a variable that was used in a prior step of the repairing process, line 17–18), or by a fresh variable, lines 19–23. Otherwise, if there is no violation so far, then we add the mapping \{x \mapsto ξ(y)\} to ψ, if it is not already there, lines 24–25. Finally, the algorithm chooses the best repair according to the preference function, lines 28–31.

Example 4. We demonstrate an example of Algorithm 3.

Since Algorithm 3 focuses on I_S(V) overlooking the actual views in V, we will not explicitly define V. Instead, we will only assume that the visible chase computes the instance

I_S(V) = \{R_1(<*, n_1, n_2), S_1(n_1, n_2, n_3), S_1(n_1, n_3, *,), S_1(n_1, *, *)\}

where n_1–n_3 are labeled nulls. Consider also the mapping M consisting of the following s-t dependency

\[R_3(x, y, z) \land S_1(y, z, z) \rightarrow T_1(x, z)\ (μ_1)\]

Note that M is not partially-safe. Algorithm 3 computes two repairs for μ_1 by applying the steps described below. First, it computes the atoms \(R_1(x_1, x_2, x_3)\) S_1(x_4, x_5, x_6) and adds them to C, lines 3–7. Then, it identifies the following three homomorphisms from C into I_S(V):

\(ξ_1 = \{x_1 \mapsto *, x_2 \mapsto n_1, x_3 \mapsto n_2, x_4 \mapsto n_1, x_5 \mapsto n_2, x_6 \mapsto n_2\}\)
\(ξ_2 = \{x_1 \mapsto *, x_2 \mapsto n_1, x_3 \mapsto n_2, x_4 \mapsto n_1, x_5 \mapsto n_3, x_6 \mapsto *\}\)
\(ξ_3 = \{x_1 \mapsto *, x_2 \mapsto n_1, x_3 \mapsto n_2, x_4 \mapsto n_1, x_5 \mapsto *, x_6 \mapsto *\}\)

From ξ_1, we can see that the joins in the body of μ_1 are safe; however, it is unsafe to export z. From ξ_2, we can see that it is safe to reveal the third position of S_1, however, it is unsafe to join the second and the third position of S_1. Algorithm 3 then iterates over ξ_1 and ξ_2, line 9. When B = R_1(x_1, y, z) and p < 3, Algorithm 3 computes ψ to [x \mapsto *, y \mapsto n_1], since there is no violation according to line 16. When B = R_1(x, y, z) and p = 3, however, a violation is detected. This is due to the facts that z is an exported variable and ξ_1(x_3) = n_1.

Algorithm 3 tackles this violation by creating a fresh variable z1. Then, it adds the relation (z, z_1) to ρ, replaces z in B_3 by z_1 and adds the mapping \{z_1 \mapsto n_2\} to ψ, lines 19–23. Algorithm 3 then
The claim \( B = \psi \) due to \( \theta \) proof of the following claim, for each iteration let

\[
\text{Algorithm 3, then proceeds by repairing } \mu_1 \text{ based on } \xi_2. \text{ When } B = R_1(x, y, z), \text{ Algorithm 3 proceeds as described above and computes } \\
\psi \rightarrow \{ x \mapsto *, y \mapsto n_1, z_1 \mapsto n_2 \}. \text{ When } B = S_1(y, z, z) \text{ and } p = 1, \text{ then no violation is encountered since } \psi(y) = \xi_1(x_1), \text{ while when } \\
B = S_1(y, z, z) \text{ and } p = 2, \text{ there is a violation. Since the condition in line } \text{18 is not met, Algorithm 3 creates a fresh variable } z_2 \text{ and adds the mapping } (z_2 \mapsto n_3) \rightarrow \psi. \text{ When } B = S_1(y, z, z) \text{ and } p = 3, \text{ then no violation is met, since } z \in \text{exported(}x\text{)} \text{ and } \xi_2(x_6) = *. \text{ Hence, the second repair of } \mu_1 \text{ is}
\]

\[
R_1(x, y, z_1) \land S_1(y, z_2, z \rightarrow T_1(x, z) \tag{r1}
\]

Finally, we can see that the repair for \( \mu_1 \) w.r.t. \( \xi_3 \) is
\[
R_1(x, y, z_1) \land S_1(y, z_2, z \rightarrow T_1(x, z) \tag{r2}
\]

Proposition 2. For any \( M = (S, T, \Sigma) \), any \( M_V = (S, V, \Sigma) \) and any preference function \( prf \), Algorithm 3repair returns a mapping \( M_V = (S, T, \Sigma') \) that is partially-safe w.r.t. \( M_V \) on all instances of \( S \).

Proof. (Sketch) From Proposition 1, a mapping \( M = (S, T, \Sigma) \) is partially-safe w.r.t. \( M_V = (S, V, \Sigma) \) on all instances of \( S \), if for each \( \mu \in \Sigma \), there exists a homomorphism from body(\( \mu \)) into \( I_0(V) \) mapping each \( x \in \text{exported(} \mu \text{)} \) into the critical constant *. Since for each \( \mu \in \Sigma \), prepair computes a set of repaired tgdts \( \mathcal{R}_\mu \), it follows that Proposition 2 holds, if such a homomorphism exists for each repaired tgd in \( \mathcal{R}_\mu \). The proof proceeds as follows. Let \( \mu'_1 \) and \( \psi' \) denote the repaired s-t tgd and the homomorphism \( \psi \) computed at the end of each iteration \( i \) of the steps in lines 12–25 of Algorithm 3. Let also \( B^i \) denote the \( i \)-th atom in body(\( \mu_R \)). Since each \( C \in \Sigma \) is an atom of distinct fresh variables, since \( \xi \) is a homomorphism from \( C \) to \( I_0(V) \) and since \( \psi(B^i) = \mu_R \), it follows that in order to prove Proposition 1, we have to show that the following claim holds, for each \( i \geq 0 \):

- \( \phi, \psi' \) is a homomorphism from the first \( i \) atoms in the body of \( \mu_R \) into \( I_0(V) \) mapping each exported variable occurring in \( B^0, \ldots, B^i \) into the critical constant *.

For \( i = 0 \), \( \phi \) trivially holds. For \( i + 1 \) and assuming that \( \phi \) holds for \( i \), let \( C^{i+1} = \psi'(B^{i+1}) \), line 13. The proof of claim \( \phi \) depends upon the proof of the following claim, for each iteration \( p \geq 0 \) of the steps in lines 14–25:

- \( \theta, \psi'^{i+1}(B^{i+1}) \rightarrow \xi(y), \text{ where } y = C^{i+1} \).

The claim \( \theta \) trivially holds for \( p = 0 \), while for \( p > 0 \), it directly follows from the steps in lines 16–25. Since \( \phi \) holds for \( i \), since the steps in lines 16–25 do not modify the variable mappings in \( \psi' \) and due to \( \theta \), it follows that \( \phi \) holds for \( i + 1 \), concluding the proof of Proposition 1.

\section{5.2 Computing safe mappings}

Unifications of one or more labeled nulls occurring in \( I_1 \) with the critical constant *, might lead to unsafe instances. Consider, for instance, a simplified variant of Example 1, where \( \Sigma_{st} \) comprises only \( S_{st1} \) and \( S_{st2} \). Both \( S_{st1} \) and \( S_{st2} \) are partially-safe, as we have explained above. However, the unification of the labeled nulls \( n_0 \) and \( n_1 \) produces an unsafe instance. Algorithm 4 aims at repairing the output of the previous step, such that no unsafe unification of a labeled null with * takes place.

Consider again the simplified variant of \( \Sigma_{st} \) from above. Since \( \Sigma_{st} \) is partially-safe, it suffices to look for homomorphism violations in \( I_1 \), for \( i \geq 1 \). A first observation is that the homomorphism violations are “sitting” within the bags. This is due to the fact that each bag stores all the facts associated with the bodies of one or more s-t tgdts from \( \Sigma_{st} \). A second observation is that one way for preventing unsafe unifications is to hide exported variables. For example, let us focus on the unsafe unification of \( n_0 \) with *. This unification takes place due to \( \epsilon_1 \), which in turn has been created due to the fact that \( e \) is an exported variable in \( \mu_1 \). By hiding the exported variable \( e \) from \( \mu_ε \), we actually prevent the creation of \( \epsilon_1 \) and hence, we block the unsafe unification of \( e \) with *. Hiding exported variables is one way for preventing unsafe unifications with the critical constant. Another way for preventing unsafe unifications is to break joins in the bodies of the rules.

Example 5. This example demonstrates a second approach for preventing unsafe labeled null unifications.

Consider a set of policy views \( V \) leading to the following instance \( I_2(V) = \{ R_1(n_1, n_1, *), R_1(\epsilon, *, n_2), S_1(\epsilon) \} \), where \( n_1 \) and \( n_2 \) are labeled nulls. Consider also the mapping \( M \) consisting of the following s-t dependencies:

\[
R_1(x, y, z) \land S_1(y) \rightarrow T_1(y) \tag{m2}
\]
\[
R_1(x, y, z) \rightarrow T_2(x) \tag{m3}
\]

It is easy to see that \( M \) is partially-safe, but unsafe in overall. Indeed, \( I_2(\Sigma) \) will consist of the following bags (for presentation purposes, we adopt the notation from Example 2):

\[
T_1(\epsilon) \rightarrow R_1(n_3, n_5, *), S_1(\epsilon)
\]
\[
T_2(\epsilon) \rightarrow R_1(\epsilon, *, n_3)
\]
\[
R_1(n_3, n_5, *), S_1(\epsilon) \rightarrow R_1(\epsilon, *, n_3), S_1(\epsilon)
\]

where \( \epsilon_3 := R_1(x, y, z) \rightarrow x = *, \theta_1 = \{ y \mapsto e \}, \theta_2 = \{ x \mapsto e \} \), and \( \theta_3 = \{ x \mapsto n_3, y \mapsto e \} \). Note that \( \epsilon_3 \) has been created out of \( \mu_3 \), since there exists a homomorphism from body(\( \mu_3 \)) into \( R_1(n_3, n_5, *) \) mapping the exported variable \( x \) into \( n_3 \).

One approach for preventing the unsafe unification of \( n_0 \) with * is to hide the exported variable \( x \) from \( \mu_3 \). By doing this, we block the creation of \( \epsilon \) and hence the unsafe unification.

A second approach is to keep \( x \) as an exported variable in \( \mu_3 \), but modify the body of \( \mu_2 \) by breaking the join between the first and the second position of \( R_1 \):

\[
R_1(x, z, y) \land S_1(y) \rightarrow T_1(y) \tag{m2'}
\]

By doing this, we prevent the creation of \( \epsilon \), since the instance computed at line 2 of Algorithm 1 would consist of the facts \( R_1(n_3, n_5, *) \), \( R_1(n_5, n_5, *), S_1(\epsilon) \).
Repairing mappings under policy views

We use the following notation:

- **dependency depth**

with arguments (apart from $\xi_n$ modifies $\Sigma$ exported variables through $\tau$ dependencies, lines 4–25. Algorithm 4 terminates either when the homomorphism violations within each bag and preventing un-

safe unifications either by hiding exported variable, or by modifying the bodies of the s-t tgd.s. In brief, at each iteration $i \geq 0$, the algorithm repairs one or more dependencies from $\Sigma_i$, where $\Sigma_0 = \Sigma$, and incrementally computes the visible chase of the new set of dependancies, lines 4–25. Algorithm 4 terminates either when the dependencies are safe, or when the maximum number of iterations $n$ is reached, line 25, in which case it repairs all unsafe dependancies by hiding their exported variables. The algorithm starts by initializing $\Sigma_0$ to $\Sigma$, lines 1. Then, at each iteration $i$, it first identifies the lowest depth unsafe bag, line 7, and attempts to repair the dependencies from $\Sigma_i$ that lead to its creation, lines 7–22. If $i < n$, it proposes two different repairs for $\Sigma_i$, one based on hiding exported variables through hideExported (Algorithm 5), and the second based on eliminating joins through modifyBody (Algorithm 6), lines 10–19. Algorithm 4 applies the modifyBody if there exist two bags in the support of $\beta$ that are candidates for modifyBody. Informally, Algorithm 4 tries to apply modifyBody as early as possible (condition depth($\beta_1) = 1$, depth($\beta_2) = 2$) and when there are one or more repeated variables in the body of $\text{tgld(} \beta_i \text{)}$ (recall Example 5). Otherwise, if $i = n$, it either applies the function hideExported, or it eliminates the s-t tgd.s that are responsible for unsafe unifications.

**Example 6.** We demonstrate Algorithm 4 over a simplified version of the running example, where $\Sigma_{24}^e = \{\mu_e, \mu_c\}$. It is seen that $\text{visChase}_5(\Sigma_{24}^e)$ will consist of the bags $[\beta_2, \beta_1, \beta_3, \beta_4]$. We assume $n = \infty$. During the initialization of Algorithm 4, the lowest depth bag for which there exists a homomorphism violation is $\beta_4$. Since $i < n$, the algorithm tries to repair $\Sigma_{24}^e$ by calling hideExported and modifyBody with arguments (apart from $\text{V}$ and $\text{prf}$) $\beta_3$, $\beta_4$, and $\beta_1$, respectively.

Algorithm 5 first computes $v = \{n' \mapsto x_1, n'' \mapsto x_2, n_e \mapsto x_3\}$, lines 3–4, and then computes all homomorphisms from $v(f)$ into the instance $I_4(\text{V})$, line 6. We can see that there exists only one such homomorphism $\xi = \{x_1 \mapsto n', x_2 \mapsto n'', x_3 \mapsto n_e\}$. We have $\text{tgld(} \beta_3 \text{)} = \mu_c$. The first two iterations of the loop in lines 8–12 have no effect, since despite that $\{x_1\} = n'$ and $\{x_2\} = n''$, the variables $i$ and $n$ from $\mu_c$ that are mapped to $n'$ and $n''$ via trigger($\beta_3) = h_2$ are not exported ones. During the last iteration, since $\{x_3\} = n_e$, since $h_2(e) = n_e$ and since $e$ is an exported variable, Algorithm 5 removes variable $e$ from the exported variables of $\mu_e$ and returns $\mu_e'$.

```
Algorithm 4 srepair($\Sigma, \text{V}, \text{prf, n}$)
1: $\Sigma_0 = \Sigma$
2: $\beta_1 = \text{visChase}_5(\Sigma)$
3: $i = 0$
4: do
5: $\Sigma_{i+1} = \Sigma_i$
6: $\text{cont} = \text{false}$
7: if $\exists$ unsafe $\beta \in B_i$, s.t. depth($\beta) \leq \text{depth(} \beta' \text{)}$, V unsafe bag $\beta'' \in B_i$ then
8: $\text{cont} = \text{true}$
9: if $i < n$ then
10: $r_1 = 0$, $r_2 = \text{hideExported(} \beta, \text{V, prf})$
11: if $\beta_1, \beta_2, \beta_3, \beta_4$ are candidates for modifyBody then
12: $r_1 = \text{modifyBody(} \text{tgld(} \beta_i \text{)}), \text{tgld(} \beta_1 \text{)}, \text{prf}$
13: if $r_1 = 0$ and it is preferred over $r_2$ w.r.t. prf then
14: remove $\text{tgld(} \beta_1 \text{)}$ from $\Sigma_{i+1}$
15: add $r_1$ to $\Sigma_{i+1}$
16: else
17: remove $\text{tgld(} \beta \text{)}$ from $\Sigma_{i+1}$
18: add $r_2$ to $\Sigma_{i+1}$
19: else
20: if $\beta' \text{, s.t. } \beta < \beta'' \in B_i$ then
21: remove $\text{tgld(} \beta \text{)}$ from $\Sigma_{i+1}$
22: else remove $\text{tgld(} \beta \text{)}$ from $\Sigma_{i+1}$
23: compute $J_{i+1}$ from $\Sigma_i, \Sigma_{i+1}$ and $\beta_1$
24: $i = i + 1$
25: while cont and $i < n$
26: return $\Sigma_n$
```

```
Algorithm 5 hideExported($\beta, \text{V, prf}$)
1: $J = \text{premise(} \beta \text{)}$
2: $v = \emptyset$
3: for each $n \in \text{Nulls occurring into } J$ do
4: add $(n' \mapsto x) \mapsto v$ where $x$ is a fresh variable
5: $R = \emptyset$
6: for each $\xi \in v(f) \to I_4(\text{V})$ do
7: $\mu = \text{tgld(} \beta \text{)}$
8: for each $\tau \in \text{dom(} \xi \text{)}$ do
9: if $\tau(x) \neq \emptyset$ then
10: for each $n' \in \text{exported}(\mu)$ do
11: if $\neg(\tau(y) = \text{v'}(x), \text{w.r.t. } \text{prf})$ then
12: add $\mu$ to $R$
13: else
14: choose the best repair $\mu_r$ of $\mu$ from $R$ based on prf
15: return $\mu_r$
```

```
Algorithm 6 modifyBody($\mu_1, \mu_2$)
1: $R = \emptyset$
2: if $\exists$ one or more repeated variables in $\text{body(} \mu_1 \text{)}$ then
3: for each $\xi \in \text{body(} \mu_2 \text{)} \to \text{body(} \mu_1 \text{)}$ mapping some $x_1 \in \text{exported(} \mu_1 \text{)}$ into some $x_2 \in \text{exported(} \mu_2 \text{)}$ do
4: let $B \subseteq \text{body(} \mu_1 \text{)}, \text{ s.t. } (\text{body(} \mu_2 \text{)}) = B$
5: let $V$ be the set of repeated variables from $B$
6: let $P$ be the set of positions from $B$, where all variables from $V$ occur
7: for each non-empty $S \subseteq P$ do
8: $\mu = \mu_1$
9: replace the variables in positions $S$ of $\mu$ by fresh variables
10: add $\mu$ to $R$
11: choose the best repair $\mu_r$ of $\mu$ from $R$ based on prf
12: return $\mu_r$
```

Algorithm 4 then calls modifyBody. The function does not return any repair, since there does not exist any variable repetition in the body of $\mu_e$. Hence, Algorithm 4 computes $\Sigma_{i+1} = \{\mu'_e, \mu_c\}$ and proceeds in the next iteration. The instance $I_5(\Sigma_{i+1})$ will consist of the following bags.

$h_2(e) = n_e$ and since $e$ is an exported variable, Algorithm 5 removes variable $e$ from the exported variables of $\mu_e$ and returns $\mu'_e$. The function $P(i, n, e, c) \land H_N(i, d) \implies \text{EthDis}(d)$ produces $\mu'_e$.
Proposition 2. For each mapping \( M = (S, T, \Sigma) \), each \( M_V = (S, V, V) \) and each preference function \( prf \) that always prefers the repairs computed by hideExported from the repairs computed by modifyBody, Algorithm 2 returns a non-empty mapping that is safe w.r.t. \( M_V \), if such a mapping exists.

Proof. (Sketch) From Algorithm 3, we can see that repair always computes a non-empty partially-safe mapping, if such a mapping exists. Note that a mapping, where no variable is exported and no repeated variables occur in the body of the s-t tgds is always partially-safe as long as, the predicates in the bodies of the s-t tgds are the same with the ones occurring in the policy views. Please also note that such a mapping is always considered by repair. The above argument, along with the fact that a partially-safe mapping can be transformed into a safe one by turning exported variables into non-exported ones by means of the function hideExported, show that Proposition 3 holds.

6 EXPERIMENTS

We investigate the efficiency of our repairing algorithm with the use of hard-coded preference function and with a preference function based on a learning approach. The source code and the experimental scenarios are publicly available at https://github.com/ucomignani/MapRepair.git.

We evaluated our algorithm using a set of 3,600 scenarios with each scenario consisting of a set of policy views and a set of s-t tgds. The source schemas and the policy views have been synthetically generated using iBench, the state-of-the-art data integration benchmark [1]. We considered relations of up to five attributes and we created GAV mappings using the iBench configuration recommended by [1]. We generated policy views by applying the iBench operators copy, merging, deletion of attributes and self-join ten times each. The characteristics of the scenarios are summarized in Table 1. In each scenario, we used a different number of s-t tgds \( n_{dep} \), a different number of body atoms \( n_{atoms} \) and a different number of exported variables \( n_{ars} \).

We implemented our algorithm in Java and we used the Weka library [9] that provides an off-the-shelf implementation of the k-NN algorithm. We ran our experiments on a laptop with one 2.6GHz 2-core processor, 16Gb of RAM, running Debian 9.

In the remainder, all data points have been computed as an average on five runs preceded by one discarded cold run.

6.1 Running time of repair

First, we study the impact of the number of s-t tgds and of the number of body atoms on the running time of repair. We adopt a fixed preference function that chooses the repair with the maximum number of exported variables, while, in case of ties, it chooses the repair with the maximum number of joins. We range the number of s-t tgds from 100 to 300 by steps of 50 and the number of body atoms from three to five. The results are shown in Figure (2a). Figure (2a)

Table 1: Properties of the generated iBench scenarios.

<table>
<thead>
<tr>
<th># s-t tgds per scenario (( n_{dep} ))</th>
<th>min</th>
<th>max</th>
<th>step</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>300</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

Algorithm 2 returns a non-empty mapping that is safe w.r.t. \( M_V \), if such a mapping exists.

Proof. (Sketch) From Algorithm 3, we can see that repair always computes a non-empty partially-safe mapping, if such a mapping exists. Note that a mapping, where no variable is exported and no repeated variables occur in the body of the s-t tgds is always partially-safe as long as, the predicates in the bodies of the s-t tgds are the same with the ones occurring in the policy views. Please also note that such a mapping is always considered by repair. The above argument, along with the fact that a partially-safe mapping can be transformed into a safe one by turning exported variables into non-exported ones by means of the function hideExported, show that Proposition 3 holds.

6 EXPERIMENTS

We investigate the efficiency of our repairing algorithm with the use of hard-coded preference function and with a preference function based on a learning approach. The source code and the experimental scenarios are publicly available at https://github.com/ucomignani/MapRepair.git.

We evaluated our algorithm using a set of 3,600 scenarios with each scenario consisting of a set of policy views and a set of s-t tgds. The source schemas and the policy views have been synthetically generated using iBench, the state-of-the-art data integration benchmark [1]. We considered relations of up to five attributes and we created GAV mappings using the iBench configuration recommended by [1]. We generated policy views by applying the iBench operators copy, merging, deletion of attributes and self-join ten times each. The characteristics of the scenarios are summarized in Table 1. In each scenario, we used a different number of s-t tgds \( n_{dep} \), a different number of body atoms \( n_{atoms} \) and a different number of exported variables \( n_{ars} \).

We implemented our algorithm in Java and we used the Weka library [9] that provides an off-the-shelf implementation of the k-NN algorithm. We ran our experiments on a laptop with one 2.6GHz 2-core processor, 16Gb of RAM, running Debian 9.

In the remainder, all data points have been computed as an average on five runs preceded by one discarded cold run.

6.1 Running time of repair

First, we study the impact of the number of s-t tgds and of the number of body atoms on the running time of repair. We adopt a fixed preference function that chooses the repair with the maximum number of exported variables, while, in case of ties, it chooses the repair with the maximum number of joins. We range the number of s-t tgds from 100 to 300 by steps of 50 and the number of body atoms from three to five. The results are shown in Figure (2a). Figure (2a)

Table 1: Properties of the generated iBench scenarios.

<table>
<thead>
<tr>
<th># s-t tgds per scenario (( n_{dep} ))</th>
<th>min</th>
<th>max</th>
<th>step</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>300</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

Algorithm 2 returns a non-empty mapping that is safe w.r.t. \( M_V \), if such a mapping exists.

Proof. (Sketch) From Algorithm 3, we can see that repair always computes a non-empty partially-safe mapping, if such a mapping exists. Note that a mapping, where no variable is exported and no repeated variables occur in the body of the s-t tgds is always partially-safe as long as, the predicates in the bodies of the s-t tgds are the same with the ones occurring in the policy views. Please also note that such a mapping is always considered by repair. The above argument, along with the fact that a partially-safe mapping can be transformed into a safe one by turning exported variables into non-exported ones by means of the function hideExported, show that Proposition 3 holds.

6 EXPERIMENTS

We investigate the efficiency of our repairing algorithm with the use of hard-coded preference function and with a preference function based on a learning approach. The source code and the experimental scenarios are publicly available at https://github.com/ucomignani/MapRepair.git.

We evaluated our algorithm using a set of 3,600 scenarios with each scenario consisting of a set of policy views and a set of s-t tgds. The source schemas and the policy views have been synthetically generated using iBench, the state-of-the-art data integration benchmark [1]. We considered relations of up to five attributes and we created GAV mappings using the iBench configuration recommended by [1]. We generated policy views by applying the iBench operators copy, merging, deletion of attributes and self-join ten times each. The characteristics of the scenarios are summarized in Table 1. In each scenario, we used a different number of s-t tgds \( n_{dep} \), a different number of body atoms \( n_{atoms} \) and a different number of exported variables \( n_{ars} \).

We implemented our algorithm in Java and we used the Weka library [9] that provides an off-the-shelf implementation of the k-NN algorithm. We ran our experiments on a laptop with one 2.6GHz 2-core processor, 16Gb of RAM, running Debian 9.

In the remainder, all data points have been computed as an average on five runs preceded by one discarded cold run.

6.1 Running time of repair

First, we study the impact of the number of s-t tgds and of the number of body atoms on the running time of repair. We adopt a fixed preference function that chooses the repair with the maximum number of exported variables, while, in case of ties, it chooses the repair with the maximum number of joins. We range the number of s-t tgds from 100 to 300 by steps of 50 and the number of body atoms from three to five. The results are shown in Figure (2a). Figure (2a)

Table 1: Properties of the generated iBench scenarios.

<table>
<thead>
<tr>
<th># s-t tgds per scenario (( n_{dep} ))</th>
<th>min</th>
<th>max</th>
<th>step</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>300</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

Algorithm 2 returns a non-empty mapping that is safe w.r.t. \( M_V \), if such a mapping exists.

Proof. (Sketch) From Algorithm 3, we can see that repair always computes a non-empty partially-safe mapping, if such a mapping exists. Note that a mapping, where no variable is exported and no repeated variables occur in the body of the s-t tgds is always partially-safe as long as, the predicates in the bodies of the s-t tgds are the same with the ones occurring in the policy views. Please also note that such a mapping is always considered by repair. The above argument, along with the fact that a partially-safe mapping can be transformed into a safe one by turning exported variables into non-exported ones by means of the function hideExported, show that Proposition 3 holds.

6 EXPERIMENTS

We investigate the efficiency of our repairing algorithm with the use of hard-coded preference function and with a preference function based on a learning approach. The source code and the experimental scenarios are publicly available at https://github.com/ucomignani/MapRepair.git.

We evaluated our algorithm using a set of 3,600 scenarios with each scenario consisting of a set of policy views and a set of s-t tgds. The source schemas and the policy views have been synthetically generated using iBench, the state-of-the-art data integration benchmark [1]. We considered relations of up to five attributes and we created GAV mappings using the iBench configuration recommended by [1]. We generated policy views by applying the iBench operators copy, merging, deletion of attributes and self-join ten times each. The characteristics of the scenarios are summarized in Table 1. In each scenario, we used a different number of s-t tgds \( n_{dep} \), a different number of body atoms \( n_{atoms} \) and a different number of exported variables \( n_{ars} \).

We implemented our algorithm in Java and we used the Weka library [9] that provides an off-the-shelf implementation of the k-NN algorithm. We ran our experiments on a laptop with one 2.6GHz 2-core processor, 16Gb of RAM, running Debian 9.

In the remainder, all data points have been computed as an average on five runs preceded by one discarded cold run.
shows that the performance of our algorithm is pretty high; the median repairing time is less than 1.5s, while for the most complex scenario containing up to five body atoms per s-t tgd, the median running time is less than 8s with 71s being the maximum.

Figure (2b) shows the time breakdown for repair. The first column shows the average running time to run the visible chase over the input s-t mappings, the second one shows the average running time for checking the safety of the computed bags and the third one shows the average running time for repairing the s-t tgds. The results show that the repairing time is 32 times greater than time to compute the visible chase and 40 times greater than the time to check the safety of the chase bags for scenarios with 300 s-t tgds. In the simplest scenarios, these numbers are reduced to five and nine, respectively. Overall, the absolute values of the rewriting times are kept low for these scenarios and gracefully scale while increasing the number of s-t tgds and the number of atoms in their bodies.

6.2 Time breakdown between repair and srepair

Figure (2c) shows the average running time for repair and srepair for the considered scenarios. We can see that srepair is the most time-consuming step of our algorithm. We can also see that the running time of srepair increases more in comparison to the running time of repair when increasing the number of the s-t tgds and the number of atoms in their bodies. This is due to overhead that is incurred during the incremental computation of the visual chase after repairing a s-t tgd (line 23 of Algorithm 4). Figure (2d) shows the correlation between the number of active triggers detected while incrementally computing the visual chase and the running time of srepair for scenarios with 100 s-t tgds using the ANOVA method (p-value < 2.2e-16). Figure (2d) shows that the most complex scenarios lead to the detection of more than 45,000 active triggers. Despite the high number of the detected active triggers, the running time of srepair is kept low thus validating its efficiency.

6.2.1 Evaluating learning accuracy and efficiency. We adopted the following steps in order to evaluate the performance of our learning approach. First, we defined the following two golden standard preference functions that we will try to learn:

- \( P_{max} \) which chooses the repair with the maximum number of exported variables (i.e., the first repair if \( \Delta_{FV} < 0 \), else the other repair) and in case of ties, it chooses the repair with the maximum number of joins (i.e., the first repair if \( \Delta_{J} < 0 \), else the second repair).
- \( P_{avg} \) which computes the average value \( \Delta = \frac{\Delta_{FV} + \Delta_{J}}{2} \) and chooses the first repair, if \( \Delta < 0 \); otherwise, it chooses the second repair.

For both preference functions, we created a training set of 10,000 measurements for the k-NN classifier by running the repairing algorithm on fresh scenarios of 50 s-t tgds and five body atoms per s-t tgd. For each input vector \( (\delta_{FV}, \delta_{J}) \) whose repair we wanted to predict, we computed the Euclidean distance between \( (\delta_{FV}, \delta_{J}) \) and the vectors of the training set. We also set the value of parameter \( k \) to 1. This parameter controls the number of neighbors used to predict the output. Higher values of this parameter led to comparable predictions and are omitted for space reasons. Finally, we used the trained k-NN classifier as a preference function in srepair, rerun the scenarios from Section 6.1 and compared the returned repairs with the ones returned when applying the golden standards \( P_{max} \) and \( P_{avg} \) as preference functions.

Learning \( P_{max} \). Table (2a) shows the confusion matrix associated to learning \( P_{max} \). The confusion matrix outlines the choices undertaken during the iterations of the k-NN algorithm. In our case, Table (2a) shows that \( P_{1} \) has been selected 230 times, while \( P_{2} \) has been chosen 395,680 times. We can thus see that \( P_{2} \) is chosen in the vast majority of the cases. Notice that \( P_{2} \) is also the default value in cases where the preference function weights equally \( P_{1} \) and \( P_{2} \).

Apart from the confusion matrix, we also measured the accuracy of learning the preference function, by weighing the closeness of the learned mapping to the golden standard mapping.

We used the Matthews Correlation Coefficient metric (MCC) [2] to compare the repairs returned by the trained k-NN classifier and the ones returned when applied \( P_{max} \). This is a classical measure that allows to evaluate the quality of ML classifiers when ranking is computed between two possible values (in our case, the choice between \( P_{1} \) and \( P_{2} \)). This measure is calculated using the following:

\[
MCC = \frac{N_{1,1} \times N_{2,2} - N_{1,2} \times N_{2,1}}{\sqrt{(N_{1,1} + N_{1,2})(N_{1,1} + N_{2,1})(N_{2,2} + N_{1,2})(N_{2,2} + N_{2,1})}}
\]

The results of MCC range from -1 for the cases where the model perfectly predicts the inverse of the expected values, to 1 for the cases where the model predicts the expected values. The value MCC = 0 means that there is no correlation between the predicted value and the expected one. By applying MCC to the learning of \( P_{max} \), we observed that the data are clearly discriminated, thus leading to a perfect fit of our prediction in this case (MCC = 1).

Learning \( P_{avg} \). Table (2b) shows the confusion matrix associated to learning \( P_{avg} \). We can see that the predictions are less accurate in this case. The data is not as clearly discriminated as before, leading to a fairly negligible error rate (< 0.02%). This error is still acceptable for the learning, since only < 0.02% of the predictions are erroneous. This is corroborated by a MCC value equal to 0.93, thus leading to a still acceptable fit of our preference function also in the case of \( P_{avg} \).

6.2.2 Running time of repair with a learned preference function. In the last experiment, we want to measure the impact of learning on the performance of our algorithm. To this end, we compare the running time of repair when adopting a hard-coded preference function (as in the results reported in Figure 2) and when adopting a learned preference function. Figure 3 shows the running times for the same scenarios used in Figure 2. We can easily observe that the runtimes are rather similar with and without learning and the difference amounts to a few milliseconds. This further corroborates the utility of learning the preference function and shows that the learning is robust and does not deteriorate the performances of our algorithm.

7 CONCLUSION

We have addressed the problem of safety checking w.r.t. a set of policy views in a data exchange scenario. We have also proposed efficient repairing algorithms that sanitize the mappings w.r.t. the
policy views, in cases where the former leak sensitive information. Our approach is inherently data-independent and leads to obtaining rewritings of the mappings guaranteeing privacy preservation at a schema level. As such, our approach is orthogonal to several data-dependent privacy-preservation methods, that can be used on the companion source and target instances to further corroborate the privacy guarantees. We envision several extensions of our work, such as the study of general GLAV mappings and the interplay between data-independent and data-dependent privacy methods.