A Novel Approach on the Joint De-Identification of Textual and Relational Data with a Modified Mondrian Algorithm

F. Singhofer
fabian.singhofer@uni-ulm.de
University of Ulm
Germany

A. Garifullina, M. Kern
{aygul.garifullina,mathias.kern}@bt.com
BT
United Kingdom

A. Scherp
ansgar.scherp@uni-ulm.de
University of Ulm
Germany

ABSTRACT

Traditional approaches for data anonymization consider relational data and textual data independently. We propose rx-anon, an anonymization approach for heterogeneous semi-structured documents composed of relational and textual attributes. We map sensitive terms extracted from the text to the structured data. This allows us to use concepts like k-anonymity to generate a joint, privacy-preserved version of the heterogeneous data input. We introduce the concept of redundant sensitive information to consistently anonymize the heterogeneous data. To control the influence of anonymization over unstructured textual data versus structured data attributes, we introduce a modified, parameterized Mondrian algorithm. We evaluate our approach with two real-world datasets using a Normalized Certainty Penalty score, adapted to the problem of jointly anonymizing relational and textual data. The results show that our approach is capable of reducing information loss by using the tuning parameter to control the Mondrian partitioning while guaranteeing k-anonymity. As rx-anon is a framework approach, it can be reused and extended by other anonymization algorithms, privacy models, and textual similarity metrics.

CCS CONCEPTS

• **Security and privacy** → **Data anonymization and sanitization**; • **Computing methodologies** → Information extraction.

KEYWORDS

data anonymization, heterogeneous data, k-anonymity

ACM Reference Format:
F. Singhofer, A. Garifullina, M. Kern, and A. Scherp. 2021. A Novel Approach on the Joint De-Identification of Textual and Relational Data with a Modified Mondrian Algorithm. In *ACM Symposium on Document Engineering 2021 (DocEng '21), August 24–27, 2021, Limerick, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3469096.3469871

1 INTRODUCTION

Researchers benefit from companies, hospitals, or other research institutions, who share and publish their data. It can be used for predictions, analytics, or visualizations. However, often data to be shared contains Personally Identifiable Information (PII) which does require measures in order to comply with privacy regulations. One possible measure to protect PII is to anonymize all personal identifiers. Prior work considered such personal data to be name, age, email address, gender, sex, ZIP, any other identifying numbers, among others [5, 6, 13, 15, 27]. Research in data mining and predictive models shows that a combination of structured and unstructured data leads to more valuable insights. Zhao and Zhou [34] showed that linking the mining results can provide valuable answers to complex questions related to SARS-CoV-2. Teinemaa et al. [29] developed a model for predictive process monitoring that benefits from adding unstructured data to structured data.

However, state of the art methods focus either on anonymizing structured relational data [10, 14, 16, 19, 28, 30] or anonymizing unstructured textual data [2, 6, 15, 21, 23, 25], but not jointly anonymizing on both. To the best of our knowledge, the only work that aimed to exploit synergies between anonymizing texts and structured data is by Gardner and Xiong [6]. The authors transferred textual attributes to structural attributes and subsequently applied a standard anonymization approach. However, there is no recoding of the original text, i. e., there is no transfer back of the anonymized sensitive terms. Thus, essentially Gardner and Xiong [6] only anonymize structured data. Furthermore, there is no concept of information redundancy, which is needed for a joint de-anonymization, and there is no weighting parameter to control the influence of relational versus textual attributes as splitting criterion for the data anonymization. Our experiments show that such a weighting is important as otherwise it may lead to a skewed splitting favoring to retain relational attributes over textual attributes.

To illustrate the problem of a joint anonymization of textual and structured data, we consider an example from a blog dataset [24]. As Table 1 indicates, a combined analysis relies on links between the structured and unstructured data. Therefore, it is important to generate a privacy-preserved, but also consistently anonymized release of heterogeneous datasets consisting of structured and unstructured data. Due to the nature of natural language, textual attributes might contain redundant information which is already available in a structured attribute. Anonymizing structured and unstructured parts individually neglects redundant information and leads to inconsistencies in data, since the same information might be anonymized differently. Moreover, for privacy-preserving releases, assumptions on the knowledge of an attacker are made. Privacy might be at risk if the anonymization tasks are conducted individually and without sharing all information about an individual.

We provide a formal problem definition and software framework *rx-anon* on a joint anonymization of relational data with free text fields. We experiment with two real-world datasets to demonstrate

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DocEng '21, August 24–27, 2021, Limerick, Ireland
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8596-1/21/08…$15.00
https://doi.org/10.1145/3469096.3469871

Table 1: Running example of a de-normalized dataset D with relational and textual attributes. A^* is an attribute directly identifying an individual. $A_1, ..., A_5$ are considered quasi-identifiers and do not directly reveal an individual. X is the textual attribute.

A^*	Relational Attributes $A_1, ..., A_5$					Textual Attribute X
id	**gender**	**age**	**topic**	**sign**	**date**	**text**
1	male	36	Education	Aries	2004-05-14	My name is Pedro, I'm a 36 years old engineer from Mexico.
1	male	36	Education	Aries	2004-05-15	A quick follow up: I will post updates about my education in more detail.
2	male	24	Student	Leo	2005-08-18	I will start working for a big tech company as an engineer.
3	male	37	Banking	Pisces	2004-05-27	During my last business trip to Canada I met my friend Ben from college.
4	female	24	Science	Aries	2004-01-13	As a scientist from the UK, you can be proud!
4	female	24	Science	Aries	2004-01-17	Four days ago, I started my blog. Stay tuned for more content.
4	female	24	Science	Aries	2004-01-19	2004 will be a great year for science and for my career as a biologist.
5	male	29	indUnk	Pisces	2004-05-15	Did you know that Pisces is the last constellation of the zodiac.
6	female	27	Science	Aries	2004-05-15	Rainy weather again here in the UK. I hope you all have a good day!

the benefits of the rx-anon framework. As a baseline, we consider a scenario where relational and textual attributes are anonymized alone. We show that we can reduce the information loss in texts under the k-anonymity model. Furthermore, we demonstrate the influence of the λ parameter that influences the weight between relational and textual information and optimize the trade-off between relational and textual information loss.

In summary, our work makes the following contributions:

- We formalize the problem of anonymizing relational and textual attributes under the k-anonymity model and introduce the concept of redundant information.
- We present an anonymization framework based on Mondrian [13] with an adapted partitioning strategy and recoding scheme for sensitive terms in textual data. To this end, we introduce the tuning parameter λ to control the share of information loss in relational and textual attributes.
- We evaluate our approach by measuring statistics on partitions and information loss on two real-world datasets. We adapt the Normalized Certainty Penalty score to the problem of a joint anonymization of relational and textual data.

Below, we discuss related work on data anonymization. We provide a problem formalization in Section 3 and introduce our joint de-anonymization approach rx-anon in Section 4. The experimental apparatus is described in Section 5. We report our results in Section 6. We discuss the results in Section 7, before we conclude.

2 RELATED WORK

Anonymization of Structured Data. Early work of Sweeney [26] showed that individuals can be identified by using publicly available data sources. Such attempts to reveal individuals using available linkable data are called record linkage attacks. Her work introduced explicit identifiers and quasi-identifiers. The former category is also called direct identifier and poses information which directly reveals an identity. Attributes of the latter category do not reveal an identity directly, but can if used in combination with other attributes.

This observation led to extensive research on privacy frameworks. An important and very influential approach is k-anonymity, which prevents re-identification attacks relying on record linkage using additional data [28]. k-anonymity describes a privacy model where records are grouped and each group is transformed such that their quasi-identifiers are equal. To achieve k-anonymity, Samarati [22] studied suppression and generalization as efficient techniques to enforce privacy. Several algorithms have been developed to efficiently compute a k-anonymous version of a dataset while keeping the information loss minimal. Sweeney [27] proposed a greedy approach with tuple suppression to achieve k-anonymity. LeFevre et al. [13] suggested a top-down greedy algorithm Mondrian which implements multidimensional k-anonymity using local recoding models. Ghinita et al. [7] showed how optimal multidimensional k-anonymity can be achieved by reducing the problem to a one-dimensional problem which improves performance while reducing information loss. Machanavajjhala et al. [16] introduced the model of l-diversity to prevent homogeneity and background knowledge attacks on the k-anonymity model. l-diversity uses the concept of sensitive attributes to guarantee diversity of sensitive information within groups of records. Li et al. [14] introduced t-closeness, which extends the idea of diversity by guaranteeing that the distribution within groups does not differ more than a threshold t from the global distribution of sensitive attributes.

Nergiz et al. [19] investigated the problem of anonymizing multi-relational datasets. Gong et al. [9] introduced (k, l)-diversity as a privacy model which is capable of anonymizing 1:M datasets. Terrovitis et al. [30] applied k-anonymity to transactional data. Given a set of items within a transaction, they treated each item to be a quasi-identifier as well as a sensitive attribute simultaneously. Finally, Poulis et al. [20] showed how k-anonymity can be applied to data consisting of relational and transactional data and stated that a combined approach is necessary to ensure privacy.

Anonymization of Unstructured Data. Sánchez et al. [23] proposed an anonymization method which makes use of the Information Content (IC) of terms. The IC states the amount of information a term provides and can be calculated as the probability that a term appears in a corpus. The reasoning behind using the IC of terms to detect sensitive information is that terms which

provide high information tend to be also sensitive. Early work on Named Entity Recognition (NER) to identify sensitive terms was based on rules and dictionaries [21, 25]. Gardner and Xiong [6] introduced an integrated system which uses Conditional Random Fields (CRF) to identify PII. Dernoncourt et al. [2] implemented a de-identification system with Recurrent Neural Networks (RNNs) achieving high scores in the 2014 Informatics for Integrating Biology and the Bedside (i2b2) challenge. Liu et al. [15] proposed a hybrid automatic de-identification system which incorporates subsystems using rules as well as CRFs and Bidirectional Long Short-Term Memory (BiLSTM) networks. Johnson et al. [11] were first to propose a de-identification system using transformer models [32]. Their results indicate that transformers are competitive to modern baseline models for anonymization of free texts.

In addition to the detection of sensitive information using NER, important related work is also on replacement strategies for such information in text. Simple strategies involve suppressing sensitive terms with case-sensitive placeholders [21] or with their types [18]. More complex strategies use surrogates as consistent and grammatically acceptable replacements for sensitive terms [4, 31]. Sánchez et al. [23] used generalization to transform sensitive terms to a more general version in order to reduce the loss of utility.

Work Using Synergies Between Both Fields. There were few works using synergies between both fields. Chakaravarthy et al. [1] brought a replacement technique for structured data to the field of unstructured texts. They used properties from k-anonymity to determine the sensitive terms to be anonymized within a single document by investigating their contexts. Moreover, to the best of our knowledge, only Gardner and Xiong [6] studied the task of anonymizing heterogeneous datasets consisting of texts and structured data. They provided a conceptual framework including details on data linking, sensitive information extraction, and anonymization. However, their work has no concept of redundant information between structured and textual data, as we introduce in rx-anon. Furthermore, they have no weighting parameter to balance anonymization based on structural versus textual data like we do. Basically, Gardner and Xiong [6] transfer the problem of text anonymization to the structured world and then their approach "forgets" about where the attributes came from. They do not transfer back the anonymized sensitive terms to recode the original text. So the output of Gardner and Xiong [6]'s anonymization approach is just structured data, which lacks its original heterogeneous form.

3 PROBLEM FORMALIZATION

We aim to anonymize a dataset by hiding directly identifying attributes. To prevent classical record linkage attacks using quasi-identifying attributes, we use k-anonymity as our privacy model [28]. Within texts, we adapt k-anonymity to prevent explicit information leakage, while keeping the structure of the texts as best as possible intact to allow for text mining on implicit information. In other words, using our privacy model, an attacker shall not be able to identify an individual based on attributes, their values, or sensitive terms in texts. Table 2 provides an overview of the notations used.

Heterogeneous RX-dataset. Given a dataset D in form of n relations $R_1, ..., R_n$, containing both relational and textual attributes. D contains all data we want to anonymize. We pre-process D for the anonymization process following Nergiz et al. [19] by using the natural join, i. e., $D = R_1 \bowtie ... \bowtie R_n$, i. e., we "flatten" the relational structures. Table 1 shows an example of a dataset composed of two joined relations, where the first relation describes the individuals (*id*, *gender*, *age*, *topic*, *sign*), while the latter relation (*id*, *date*, *text*) contains the posts and links them to an individual with *id* being the foreign key. We call D an *RX*-dataset, if one attribute A^* directly identifies an individual, one or more traditional relational attributes[1] A_i contain single-valued data, and one textual attribute[2] X is in D. We call a row in D a tuple t. Relational attributes A_i are single-valued and can be categorized into being nominal, ordinal, or numerical (i. e., ratio or interval, which are treated equally in the anonymization process). A textual attribute X is any attribute, where its domain is some form of free text. Therefore, we can state that $t.X$ consists of an arbitrary sequence of tokens $T =< t_1, ..., t_j >$.

Table 2: Notation for a given RX-Dataset D.

D	original dataset, $D = R_1 \bowtie ... \bowtie R_n$
R_i	relation of D
A^*	attribute of D identifying an individual directly
A_i	attribute of a relation R_j
X	textual attribute
t	tuple in D
D^*	person centric view on D
r	record (tuple) in D^*
D'	anonymized dataset
X'	set of all non-redundant sensitive terms of X
T	some text in the form of a sequence of tokens
$\mathbf{F}$	set of aggregation functions, $\mathbf{F} = \{\mathbf{F_1}, ..., \mathbf{F_l}, \mathbf{F_{X'}}\}$
E	set of sensitive entity types
er	entity recognition function, $er : T \rightarrow E$
$emap$	mapping function, $emap : \{A_1, ..., A_p\} \rightarrow E$

Note, in the case of partial information, the approach would still work by grouping the tuples with missing attribute values. For example, if we are missing the attribute *age* for an individual tuple, we can group all people together which have age missing by assigning a placeholder (e. g., "na") to the missing age fields.

Sensitive Entity Types. We define E to be a set of entity types, where each value $e \in E$ represents a distinct entity type (e. g., person or location) and each entity type is critical for the anonymization task. We then define a recognition function er on texts as $er: T \rightarrow E$. The recognition function detects sensitive terms in the text T and assigns a sensitive entity type $e \in E$ to each token $t \in T$. Moreover, we define a mapping function $emap$ on the set of structural attributes as $emap: \{A_1, ..., A_p\} \rightarrow E$. The mapping function $emap$ maps attributes $A_1, ..., A_p$ to a sensitive entity type in E, which is used to match redundant sensitive information with the text.

Redundant Sensitive and Non-redundant Sensitive Terms. Some sensitive information might appear in a textual as well as in a relational attribute. In order to consistently deal with those occurrences, we introduce the *concept of redundant sensitive information.* Redundant sensitive information is any sensitive term $x \in t.X$ with $er(x) = e_j$ for which a relational value $v \in \{t.A_i \mid \forall t \in D\}$ with

[1] By traditional relational attributes we refer to numerical, date, or categorical attributes.
[2] For the ease of reading we explore only one textual attribute. However, our approach can be extended for multiple textual attributes $X_1, ..., X_m$.

$emap(A_i) = e_j$ exists and where $x = v$. In other words, redundant sensitive information is duplicated information, i. e., has the same value which appears under the same sensitive entity type e_j in a relational attribute $t.A_i$ and a sensitive term x in $t.X$.

We introduce the attribute X', which contains all *non-redundant* sensitive information of X. For the remainder of this work, attribute names with apostrophes indicate that these attributes contain the extracted sensitive entities with their types (see *text'* in Table 3). We model X' as a set-valued attribute since in texts of $t.X$, zero or more sensitive terms can appear. Therefore, we explicitly allow empty sets to appear in $t.X'$ if no sensitive information appears in $t.X$. We then replace X in D with X', so that the schema of D becomes $\{A^*, A_1, ..., A_p, X'\}$.

Person Centric view D^* on the Dataset D. If a dataset D is composed of multiple relations, there might be multiple tuples t which correspond to a single individual. In order to apply anonymization approaches on this dataset, we need to group the data in a person centric view similar to Gong et al. [9], where one record r (i. e., one row) corresponds to one individual. Therefore, we define D^* being a grouped and aggregated version of D. This means, that we can retrieve D^* from D as $D^* =_{A^*} G_{\mathbf{F_1}(A_1),...,\mathbf{F_l}(A_p),\mathbf{F}_{\mathbf{X'}}(\mathbf{X'})}(D)$, where A^* denotes a directly identifying attribute related to an individual used to group rows of individuals together, G concurrently applies a set of aggregation functions $\mathbf{F_i}$ and $\mathbf{F_{X'}}$ defined on relational attributes A_i as well as sensitive textual terms X'. For relational attributes A_i, we use **set** as a suitable aggregation function, where two or more distinct values in A_i for one individual result in a set containing all distinct values. For set-based attributes like X', we use the aggregation function **union**, which performs an element-wise union of all sets in X' related to one individual. Table 3 presents a person centric view of our initial example where each record r represents one individual. Dates as well as any non-redundant sensitive terms have been aggregated, as discussed.

k-anonymity in D^*. Based on the notion of equivalence classes [20] and the definition of equality of set-based attributes [10], an equivalence class for D^* can be defined as a partition of records P where for any two records $r, s \in P$ holds $(r.A_1, ..., r.A_p) = (s.A_1, ..., s.A_p)$ and $r.X' = s.X'$. Thus, within an equivalence class each record has the same values for the relational attributes and their sets of sensitive terms have the same values, too. Given our definition of equivalence classes, a person centric dataset D^* is said to be k-anonymous if all equivalence classes of D^* have at least the size k. We refer to the k-anonymous version of D^* as D'.

4 RX-ANON APPROACH

Using the definitions from Section 3, we present our anonymization approach rx-anon. We present how we preprocess our data to generate a person centric view. We show how Mondrian [13], a recursive greedy anonymization algorithm, can be used to anonymize *RX*-datasets. Mondrian transforms a dataset into a k-anonymous version by partitioning the dataset into partitions with sizes greater than k and afterwards recodes each partition individually. We introduce an alternative partitioning strategy called Global Document Frequency (GDF) as a baseline for partitioning a dataset with sensitive terms. We use the running example (Table 1) to show how an *RX*-dataset is transformed to a privacy-preserved version.

4.1 Pre-processing D to Person-centric View D^*

Using the running example from Table 1, we demonstrate the steps involved to create the person centric view shown in Table 3. First, we identify sensitive terms in the texts and assign sensitive entity types to them. In the remainder of this work, we will use subscripts to indicate the entity type assigned to a sensitive term. Given the first row of the example in Table 1, the text is "My name is Pedro, I'm a 36 years old engineer from Mexico". The sensitive terms are $Pedro_{\text{person}}$, $36\ years\ old_{\text{age}}$, $engineer_{\text{job}}$, and $Mexico_{\text{location}}$. This analysis of texts is executed for all tuples t in D, while there can be multiple sensitive terms from the same entity type within a text, or even no sensitive terms at all. In the next step, we find and mark redundant sensitive information using the results of the prior steps. Therefore, we perform row-wise analyses of relational values with sensitive terms to find links, which actually represent the same information. In our example above, the sensitive term 36 years old_{age} depicts the same information as the value *36* in the relational attribute *age*. Therefore, this sensitive term in the textual attribute is marked as redundant and is not considered as new sensitive information during the anonymization algorithm. Non-redundant sensitive information is stored in the attribute *text'*.

Finally, we build a person-centric view to have a condensed representation of all information available for each individual. Therefore, as described in Section 3, we group the data on a directly identifying attribute to get an aggregated dataset. In the example in Table 1, the directly identifying attribute A^* is *id*. We use **set** as the aggregation function for the relational attributes. Moreover, we collect all sensitive terms mentioned in texts of one individual by performing **union** on the sets of sensitive terms.

Table 3 shows the person-centric view D^* of our dataset D, which has been achieved by aggregating on the attribute *id*. Since the individuals with the *ids 1* and *4* have blogged more than once on different dates, multiple dates have been aggregated as sets. Moreover, since those people also have blogged different texts on different days, all sensitive terms across all blog posts have been collected in the attribute *text'*.

4.2 Compute Anonymized Dataset D' from D^*

Given a person centric dataset D^*, we want to build a k-anonymous version D' by using the definitions of the previous section. In order to achieve anonymization, we adapt the two step anonymization algorithm of Mondrian by LeFevre et al. [13], which first decides on m partitions $P_1, ..., P_m$ (refer to Algorithm 1), and afterwards recodes the values of each partition to achieve k-anonymity. We use Global Document Frequency (GDF) partitioning as baseline partitioning algorithm (see Algorithm 2), which uses sensitive terms and their frequencies to create a greedy partitioning using presence and absence of sensitive terms.

Modified Mondrian Partitioning with Weight Parameter λ. The first step of the algorithm is to find partitions of records with a partition size of at least k. LeFevre et al. [13] introduced multi-dimensional strict top-down partitioning where non-overlapping partitions are found based on all relational attributes. Moreover, they introduced a greedy strict top-down partitioning algorithm Mondrian. Starting with the complete dataset D^* as an input, the partitioning algorithm chooses an attribute to split on and then

Table 3: Preprocessed version of the illustrative example. The attribute *date* has been aggregated as set. The attribute *text′* contains sensitive terms of the attribute *text* for all blog posts published by a single individual.

id	gender	age	topic	sign	date	text′
1	male	36	Education	Aries	2004-05-14, 2004-05-15	$\text{Pedro}_{\text{person}}$, $\text{engineer}_{\text{job}}$, $\text{Mexico}_{\text{location}}$
2	male	24	Student	Leo	2005-08-18	$\text{engineer}_{\text{job}}$
3	male	37	Banking	Pisces	2004-05-27	$\text{Ben}_{\text{person}}$, $\text{Canada}_{\text{location}}$
4	female	24	Science	Aries	2004-01-13, 2004-01-17, 2004-01-19	$\text{Four days ago}_{\text{date}}$, $\text{scientist}_{\text{job}}$, $\text{biologist}_{\text{job}}$, $\text{UK}_{\text{location}}$
5	male	29	indUnk	Pisces	2004-05-15	
6	female	27	Science	Aries	2004-05-15	$\text{UK}_{\text{location}}$

splits the partition by median-partitioning. The authors suggest using the attribute which provides the widest normalized range given a sub-partition. For numerical attributes, the normalized range is defined as minimum to maximum. For categorical attributes, the range is the number of distinct categories observed in a partition. Sensitive, textual terms are treated as categorical attributes.

In order to properly treat textual terms in this heuristic algorithm, we introduce a weight parameter λ to the modified Mondrian algorithm shown in Algorithm 1. It describes the priority to split partitions on relational attributes. $\lambda = 1$ means that the algorithm always favors to split on relational attributes. $\lambda = 0$ leads to splits only based on sensitive terms in textual attributes. $\lambda = 0.5$ does not influence the splitting decisions and therefore is considered as default. The partitioning algorithm stops if no allowable cut can be made such that the criteria of k-anonymity holds for both sub-partitions. Therefore, we can stop splitting partitions if $|P| < 2k$.

Algorithm 1: Modified Mondrian partitioning with weight parameter λ (adapted from LeFevre et al. [13]).

```
Input  : Partition P, weight λ
Output : Set of partitions with size of at least k
Function mondrian_partitioning(P, λ):
    if |P| < 2k then // no allowable cut
        return P
    end
    else
        A = next_attribute(λ)
        F = frequency_set(P, A)
        P_l = (r ∈ P | r.A < find_median(F))
        P_r = P \ P_l
        return mondrian_partitioning(P_l) ∪
          mondrian_partitioning(P_r)
    end
```

Global Document Frequency (GDF) Partitioning. Using the idea of a top-down strict partitioning algorithm, we propose with GDF a greedy partitioning algorithm using the presence and absence of sensitive terms. The main goal is to keep the same sensitive terms within the same partition. This is achieved by creating partitions with records which have sensitive terms in common. Algorithm 2 presents the GDF partitioning algorithm. Similar to Algorithm 1, we start with the whole dataset as a single partition. Instead of splitting the partition using the median of a relational attribute (Mondrian partitioning), *we split partitions on a chosen sensitive term.* While the first sub-partition contains only records, where the chosen sensitive term appears, the second sub-partition contains the remaining records. For choosing the next term to split on, multiple heuristics are possible. We propose to use the most frequently apparent sensitive term for the remaining texts in the partition as the term to split on. Taking the most frequent term allows us to keep the most frequently appearing term in a majority of texts while suppressing less frequently used terms. The term used to split is then removed and similar to Algorithm 1 the algorithm is recursively called using the first and second partition, respectively.

Algorithm 2: Top-down document-frequency-based (GDF) partitioning on sensitive terms in X'.

```
Input  : Partition P, terms with their frequencies F
Output : Set of partitions with size of at least k
Function gdf_partitioning(P, F):
    if |P| < 2k then // no allowable cut
        return P
    end
    else
        (x, f) = next_term(P, F)
        P_l = (r ∈ P | x ∈ r.X′)
        P_r = P \ P_l
        F = F \ {(x, f)} // remove considered term
        return gdf_partitioning (P_l, F) ∪ gdf_partitioning
          (P_r, F)
    end
```

Recoding. In the next step, each partition is transformed such that values of quasi-identifiers of records are indistinguishable. This process is called recoding. There are different recoding schemes for the different scales of the attribute. Nominal and ordinal values are usually recoded using Domain Generalization Hierarchies (DGHs) as introduced by Sweeney [27] and used in multiple other works [7, 19, 20, 33]. A DGH describes a hierarchy which is used to generalize distinct values to a more general form such that within a partition all values transform to a single value in the DGH. Alternatively, nominal and ordinal attributes can also be recoded as sets containing all distinct items of one partition. For numerical attributes, LeFevre et al. [13] propose to use either mean or range as a summary statistic. Additionally, numerical attributes can also be recoded using ranges from minimum to maximum. Moreover, for dates El Emam et al. [5] propose an automated hierarchical recoding based on suppressing some information of a date value shown

in Figure 1. The leaf nodes represent actual dates appearing in the dataset D (ref. to Table 1). Non-leaf nodes represent automatically generated values by suppressing information on each level.

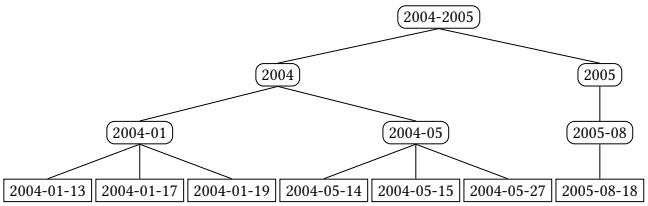


Figure 1: Domain Generalization Hierarchy for date attributes. Leaves depict values in the dataset D. First level of generalization suppresses day, followed by month, and year.

Since we use a strict-multidimensional partitioning scheme, we apply local recoding as suggested in Mondrian [13]. For numerical attributes, we use range as a summary statistic. For date attributes we use the automatically generated DGH by El Emam et al. [5] as shown in Figure 1. Moreover, since generalization hierarchies for gender, topic, and sign are flat, we recode nominal and ordinal values as sets of distinct values.

Example. After equivalence classes have been determined, relational attributes can be recoded. Table 4 shows how those recoding schemes are applied to the relational attributes of our running example. In addition, a k-anonymous representation of the text attribute X' has to be created. Terms, which are marked as redundant sensitive information, are replaced by the recoded value of its relational representatives. Using the anonymized version of our example in Table 4, the age appearing in the text of the first row is recoded using the value of the attribute *age* of the same row. Moreover, non-redundant sensitive information is recoded using suppression with its entity type. If a sensitive information appears within all records of an equivalence class, retaining this information complies with our definition of k-anonymity for set-valued attributes from Section 3. Therefore, it does not need to be suppressed (see sensitive term *engineer* in Table 4). However, if the same sensitive information is not appearing in every record within an equivalence class, this sensitive information (or the lack of it) violates our definition of k-anonymity and must be suppressed. An example for such a violation in Table 4 is *Mexico*, which appears in the first record, but in no other record of its equivalence class. The result is the k-anonymized dataset D^*.

5 EXPERIMENTAL APPARATUS

We evaluate our framework on two real-world datasets using the modified Mondrian partitioning algorithm with weighting parameter λ as well as the GDF partitioning baseline. We use λ to manipulate the splitting decisions in Mondrian as discussed in Section 4 and measure the resulting partitions as well as information loss.

5.1 Datasets

Blog Authorship Corpus. The Blog Authorship Corpus[3] was originally used to create profiles from authors [24] but has also been used in privacy research for author re-identification [12]. After cleaning the input data from unreadable characters and others, the corpus contains 681,260 blog posts from 19,319 bloggers, which have been written by a single individual on or before 2006 and published on blogger.com. A row in the corpus consists of the *id*, *gender*, *age*, *topic*, and *zodiac sign* of a blogger as well as the *date* and the *text* of the published blog entry. Each row corresponds to one blog post written by one individual, but one individual might have written multiple blog posts. On average, one blogger has published 35 blog posts. We treat *id* as a direct identifier, *gender*, *topic*, and *sign* as categorical attributes, while *age* is treated as a numerical attribute. The attribute *date* is treated as a special case of categorical attribute where we recode dates using the automatically generated DGH shown in Figure 1. The attribute *text* is used as the textual attribute. The attribute *topic* contains 40 different topics, including industry-unknown (indUnk). *Age* ranges from 13 to 48. *Gender* can be male or female. *Sign* can be one of the twelve astrological signs.

Hotel Reviews Dataset. We use the 515K Hotel Reviews Data in Europe dataset[4], called in the following briefly the *Hotel Reviews Dataset*, which contains 17 attributes, of which 15 attributes are relational and two attributes are textual. The textual attributes are *positive* and *negative reviews* of users. Among the relational attributes, we treat *hotel name* and *hotel address* as direct identifiers. *Negative* and *positive word count* as well as *tags* are ignored and therefore considered insensitive attributes. The remaining attributes are treated as quasi-identifiers, with seven numerical, one date, and two nominal attributes. After preparing the Hotel Reviews Dataset, we have 512,126 reviews for 1,475 hotels remaining.

5.2 Procedure

Similar to experiments conducted in prior work [7, 9, 20], we run our anonymization tool for different values of $k = 2, 3, 4, 5, 10, 20$, and 50. Regarding our new weighting parameter λ, we used values between 0.0 and 1.0 in steps of 0.1. Sensitive entity types in texts are those detected by spaCy's English models trained on the OntoNotes5 corpus[5]. We added rule-based detectors for the entities MAIL, URL, PHONE, and POSTCODE. We treat all sensitive terms appearing under those entity types as quasi-identifiers. For each value of k, we conduct experiments using different partitioning strategies and parameter settings. In particular, we vary the weight parameter λ to tune Mondrian. We use the same recoding scheme, namely local recoding, for all experiments to make partitioning results comparable. For the evaluation, we analyze the anonymized dataset with respect to the corresponding partitioning sizes and information loss.

5.3 Measures

Statistics on Partitions. We evaluate how partitions are created, based on relational attributes versus textual attributes, and how λ influences splitting decisions. In addition to the number of splits, we want to evaluate the size of the resulting partitions since they are closely related to information loss. By the nature of k-anonymity, all partitions need to be at least of size k. Relatively large partitions with respect to k will tend to produce more information loss.

[3] https://www.kaggle.com/rtatman/blog-authorship-corpus

[4] https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

[5] https://spacy.io/api/data-formats#named-entities

Table 4: Anonymized dataset D' for $k = 2$. Redundant information and remaining sensitive terms are marked bold.

id	gender	age	topic	sign	date	text
1	male	**[24-36]**	(Student,Education)	(Leo,Aries)	[2004-2005]	My name is **person**, I'm a **[24-36]** years old **engineer** from **location**.
1	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	A quick follow up: I will post updates about my education in more detail.
2	male	[24-36]	(Student,Education)	(Leo,Aries)	[2004-2005]	I will start working for a big tech company as an **engineer**.
3	male	[29-37]	(indUnk,Banking)	Pisces	2004-05	During my last business trip to **location** I met my friend **person** from college.
4	female	[24-27]	Science	Aries	2004	As a **job** from the **UK**, you can be proud!
4	female	[24-27]	Science	Aries	2004	**Date**, I started my blog. Stay tuned for more content.
4	female	[24-27]	**Science**	Aries	**2004**	**2004** will be a great year for **science** and for my career as a **job**.
5	male	[29-37]	(indUnk,Banking)	**Pisces**	2004-05	Did you know that **Pisces** is the last constellation of the zodiac.
6	female	[24-27]	Science	Aries	2004	Rainy weather again here in the **UK**. I hope you all have a good day!

Therefore, partition sizes closer to k will be favorable and increase utility.

Information Loss (Adapted to Heterogeneous Datasets). Measuring the information loss of an anonymized dataset is well-known practice for evaluating the amount of utility remaining for a published dataset. We use Normalized Certainty Penalty (NCP) [33] to determine how much information loss has been introduced by the anonymization process. We extend the definitions of NCP to the problem of anonymizing relational and textual data such that for one record r, the information loss is calculated as $NCP(r) = (w_R \cdot NCP_A(r) + w_X \cdot NCP_X(r))/(w_A + w_X)$, where w_A is the importance assigned to the relational attributes, and $NCP_A(r)$ denotes the information loss for relational attributes of record r. Analogously, we define w_X and $NCP_X(r)$ for the textual attribute. For our evaluation, we set w_A and w_X to 1, i. e., weigh the loss stemming from relational data and textual data equally.

For *relational attributes* $A = \{A_1, ..., A_p\}$ we define the information loss $NCP_A(r) = (\sum_{A_i \in A} NCP_{A_i}(r))/|A|$, where $|A|$ denotes the number of relational attributes. NCP_{A_i} is the information loss for a single attribute and depends on the type of attribute. It can be calculated either using NCP_{num} for numerical attributes or NCP_{cat} for categorical attributes. NCP_{num} for numerical values is defined as $NCP_{num}(r) = (z_i - y_i)/|A_i|$, with z_i being the upper and y_i being the lower bound of the recoded numerical interval and $|A_i| = max_{r \in D^*}(r.A_i) - min_{r \in D^*}(r.A_i)$. For categorical values, NCP_{cat} is defined as $NCP_{cat}(r) = \begin{cases} 0 & |u| = 1 \\ \frac{|u|}{|A_i|} & \text{otherwise} \end{cases}$, where $|u|$ denotes the number of distinct values which the recoded value u describes. For categorical values other than dates, $|u|$ will be the number of distinct values appearing in the recoded set. For date attributes, $|u|$ denotes the number of leaves of the subtree below the recoded value (see Figure 1).

For *textual attributes*, we define $NCP_X(r) = (\sum_{x \in r.X'} NCP_x(x))/|r.X'|$, where for each sensitive information x, we calculate the individual information loss $NCP_x(x)$ and normalize it by the number of sensitive terms $|r.X'|$. We define the individual information loss for one sensitive term as $NCP_x(x) = 1$ if x is suppressed, and 0 otherwise.

Finally, we can calculate the total information loss for an entire RX-Dataset D^* as $NCP(D^*) = (\sum_{r \in D^*} NCP(r))/|D^*|$, where for each record r the information loss $NCP(r)$ is calculated and divided by the number of records $|D^*|$.

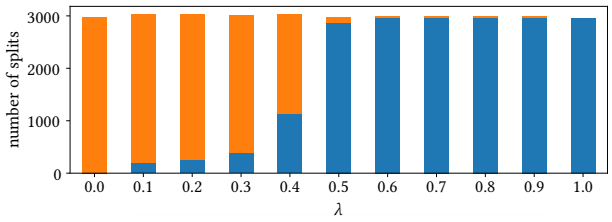
6 RESULTS

We present the results regarding partition statistics and information loss. For detailed experimental results with plots and tables for all parameter values, we refer to the supplementary material.

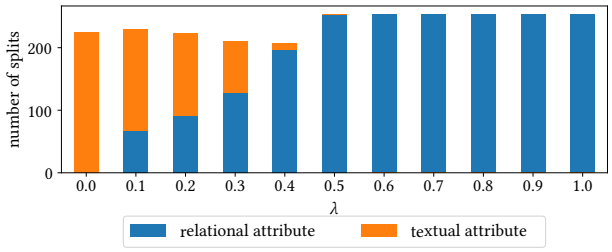
6.1 Partitions Splits, Counts, and Size

Partition Splits. Figure 2a shows the distribution of splitting decisions for experiments run on the Blog Authorship Corpus for $k = 5$ and $\lambda = 0.0$ to 1.0. As our results show, an unbiased run of Mondrian with $\lambda = 0.5$ causes partitions to be split mostly on relational attributes. Since the span of relational attributes is lower compared to sensitive terms, relational attributes provide the widest normalized span and are therefore favored to split on. For $\lambda > 0.5$, the majority of the weight for splitting is given to the relational attributes. However, for $\lambda < 0.5$, we observe that an increasing number of splits are made based on textual attributes. For $\lambda = 0.4$, already more than half of the splits are based on textual terms. For the Hotel Reviews Dataset, shown in Figure 2b, the number of splits is generally lower (see also partition sizes, below), since it contains fewer records. Also the splitting on textual attributes is less likely for hotel reviews compared to blog posts.

Partition Count and Size. Regarding the number of partitions and their size, Table 5 provides statistics on partitions using Mondrian partitioning with varying λ as well as GDF partitioning for the Blog Authorship Corpus. The Mondrian partitioning algorithm produces the same partitioning layout for λ between 0.6 and 0.9. This observation matches statistics on partition splits, since for



(a) Textual vs. relational splits on the Blog Authorship Corpus



(b) Textual vs. relational splits on the Hotel Reviews Dataset

Figure 2: Number of splits based on textual attributes (orange) versus relational attributes (blue) using Mondrian partitioning ($k = 5$) with varying weights λ.

these values of λ the Mondrian algorithm decides to use the same attributes to split on. Furthermore, GDF partitioning is not able to generate partition sizes close to k, compared to Mondrian partitioning. The results for the Hotel Reviews Dataset are shown in Table 6. We make the same observations as for the Blog Authorship Dataset. However, due to the lower number of records in the Hotel Reviews Dataset, the total count of partitions is comparatively smaller.

Table 5: Statistics on partitions for Blog Authorship Corpus. Count refers to the number of partitions found, while size refers to the average number of records per partition.

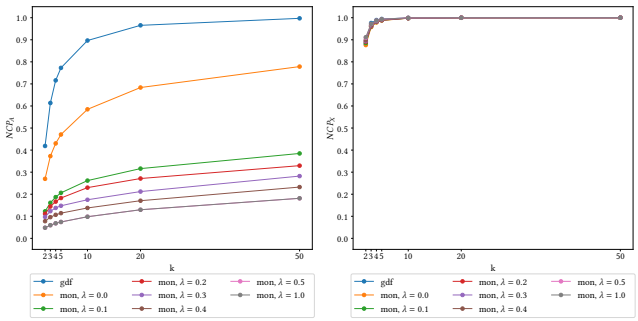
	λ	k	3	4	5	10	20
GDF	-	count	2479	1512	1078	352	92
		size	7.79	12.78	17.92	54.88	209.99
Mondrian	0	count	5162	3795	2971	1412	692
		size	3.74	5.09	6.50	13.68	27.92
	0.3	count	5236	3841	3023	1462	707
		size	3.69	5.03	6.39	13.21	27.33
	0.5	count	5198	3800	2979	1441	711
		size	3.72	5.08	6.49	13.41	27.17
	0.6 - 0.9	count	5180	3781	2987	1441	711
		size	3.73	5.11	6.47	13.41	27.17
	1	count	5128	3749	2964	1431	703
		size	3.77	5.15	6.52	13.50	27.48

6.2 Information Loss

Figure 3a provides an overview on relational information loss NCP_A (y-axis) for different values of k between 2 and 50 (x-axis) for the Blog Authorship Corpus. Figure 3b shows the textual information loss NCP_X. Figures 4a and 4b provide the information loss for experiments run on the Hotel Reviews dataset.

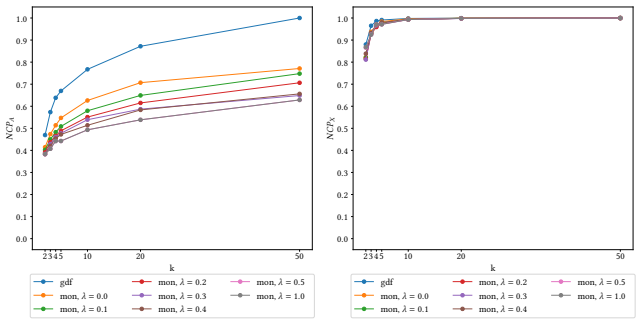
Table 6: Statistics on partitions for Hotel Reviews Dataset. Count refers to the number of partitions found, while size refers to the average number of records per partition.

	λ	k	3	4	5	10	20
GDF	-	count	272	163	127	43	16
		size	5.42	9.05	11.61	34.30	92.19
Mondrian	0	count	398	293	226	117	54
		size	3.71	5.03	6.53	12.61	27.31
	0.3	count	404	285	212	106	50
		size	3.65	5.18	6.96	13.92	29.50
	0.5	count	415	256	255	128	64
		size	3.55	5.76	5.78	11.52	23.05
	0.6 - 1	count	417	256	255	128	64
		size	3.54	5.76	5.78	11.52	23.05



(a) Relational information loss NCP_A (b) Textual information loss NCP_X

Figure 3: Information loss for relational attributes (a), and textual attributes (b) on the Blog Authorship Corpus.



(a) Relational information loss NCP_A (b) Textual information loss NCP_X

Figure 4: Information loss for relational attributes (a), and textual attributes (b) on the Hotel Reviews Dataset.

Relational Information Loss. The information loss increases with larger k throughout all experiments. Higher information loss is caused by having larger partitions and therefore higher efforts in recoding. Furthermore, we can state that information loss in the relational attributes increases if the tuning parameter λ decreases (see Figure 3a). This observation agrees with statistics on splitting decisions, since for lower values of λ, Mondrian more frequently decides to split on sensitive terms in textual attributes. This leads to more variations in relational values of partitions, which ultimately increases the relational information loss. Comparing with Figure 4a, we can state that relational information loss appears to be higher for the Hotel Reviews Dataset compared to the Blog Authorship Corpus. However, we still observe the same behavior where higher values of λ result in relatively lower relational information loss.

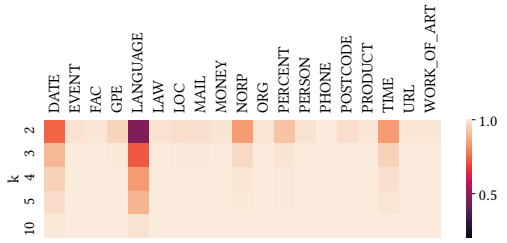
Textual Information Loss. Analyzing the information loss in the textual attribute, see Figure 3b, one observation is that for values of $k \geq 10$ the information loss in texts tends to become 1. This equals suppressing all sensitive terms in texts. Moreover, our modified Mondrian partitioning performs better compared to the naive partitioning strategy GDF. GDF partitioning results in partitions with unequal and larger sizes and therefore ends up with large partitions, which significantly increase information loss. We make the same observations on the Hotel Reviews Dataset plotted in Figure 4b. However, information loss for $k \leq 5$ tends to be slightly lower.

6.3 Attribute-level Textual Information Loss

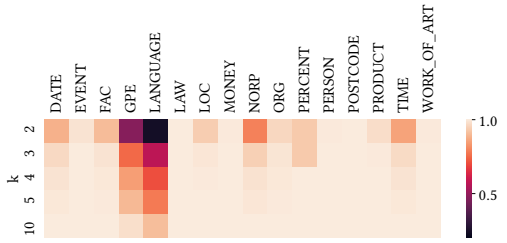
To get a deeper understanding of textual attributes on the anonymization process, we analyzed textual information loss on entity type level. Figure 5a provides an overview of information loss per different entity type extracted from text in the Blog Authorship Corpus for k is 2 to 50 and a fixed $\lambda = 0.2$. It shows that there is a high information loss for most attributes, even for small k. However, information loss of sensitive terms of type LANGUAGE may be reduced for values of $k \leq 5$. Since the number of distinct entities of type LANGUAGE is much lower compared to other entity types in the Blog Authorship Corpus like EVENT and PERSON, the entities (i. e., number of sensitive terms) of type LANGUAGE can be better preserved. We obtain similar results for Mondrian partitioning with other values of $\lambda \leq 0.4$. We make the same observations on the Hotel Reviews dataset for both textual attributes, the positive reviews and negative reviews (see Figures 5b and 5c). In addition to LANGUAGE entities, sensitive locations (GPE) can also be preserved for both textual attributes.

7 DISCUSSION

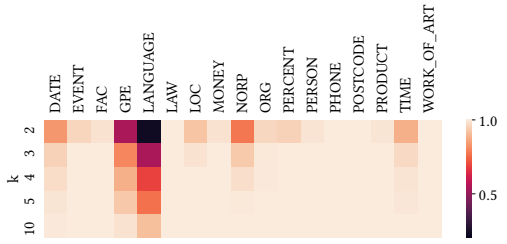
Due to heterogeneity of sensitive terms in texts, by default, they are less likely to be considered to split on. By introducing the tuning parameter λ in our framework, we were able to control the Mondrian algorithm to preserve more information in either relational or textual attributes. We observe that a value of λ between 0.4 and 0.5 results in balanced splits, i. e., about the same number of splits are based on relational attributes versus textual terms. Our anonymization approach allows us to reduce the information loss in texts under the k-anonymity privacy model. In contrast, in the related work [2, 15, 25] sensitive terms have been completely suppressed. Furthermore, our experiments show that for $k \leq 5$, not



(a) Entity-level textual loss on the Blog Authorship Corpus



(b) Entity-level textual loss on positive reviews of Hotel Review Dataset



(c) Entity-level textual loss on negative reviews of Hotel Review Dataset

Figure 5: Textual information loss per entity type with Mondrian ($\lambda = 0.2$). Not all entity types appear in all datasets.

all sensitive terms need to be suppressed. In case of entities of type LANGUAGE, our approach could preserve about 60% for $k = 2$ in the Blog Authorship Corpus (see Figure 5a) and up to 80% of terms for $k = 2$ in the Hotel Reviews Dataset (see Figures 5b and 5c). Generally, when applying k-anonymity on sensitive terms, it works better for texts from a specific domain (e. g., hotels) than cross-domain datasets (e. g., blogs), as the latter have a higher diversity.

There may be a possible over-anonymization or under-anonymization in our rx-anon approach, influenced by the accuracy of the detected sensitive terms. Over-anonymization resembles the case where sensitive terms are falsely suppressed. It is caused by low precision and reduces utility of the anonymized data. Under-anonymization describes a case where sensitive terms are falsely kept. This case is generally considered more critical than falsely suppressing terms and is related to low recall. If entities which should have been anonymized are not detected at all, the information they provide will appear in the released dataset and might reveal information which should not have been disclosed. We address this thread of validity and use a state-of-the-art Natural Language Processing (NLP) library spaCy to extract named entities from text.

We use spaCy's recent transformer-based language model [3] for English (Version 3.0.0a0), which has an F1-score for NER tasks of 89.41. However, there are cases such as misspellings, jargon, foreign words, and others that we are yet missing. There can be different ways to express the same sensitive information which leads to over-anonymization. For example, the capital city of Germany may be referred to simply by its actual name "Berlin" or indirectly referred to as "Germany's capital". It is not possible for our current system to resolve such linkage. There may also be identical terms which actually have different semantics, which leads to under-anonymization. In example, consider the phrases "I live in Berlin" and "I love Berlin". Our approach would treat both appearances of "Berlin" the same way. In order to mitigate such cases, one can integrate more advanced text matching functions to our rx-anon framework, potentially depending on the requirements of a specific use case. For example, one could use contextualized word vectors [3, 17]. Note, in this work, we focus on showing that heterogeneous data can be anonymized using our rx-anon approach.

The idea of linking relational fields to attributes of other data types could be extended in order to retrieve a consistent, and privacy-preserved version of heterogeneous JSON or XML documents [8]. In addition, tuning the partitioning using a parameter like λ is not only relevant in the context of anonymizing heterogeneous data, but could also be adapted to an attribute level to favor distinct attributes over others.

8 CONCLUSION

We introduced rx-anon as a step towards a framework for anonymizing hybrid documents consisting of relational as well as textual attributes. We have formally defined the problem of joint anonymizing heterogeneous datasets. This is achieved by transferring sensitive terms in texts to an anonymization task of structured data, introducing the concept of redundant sensitive information, and establishing the tuning parameter λ to control and prioritize information loss in relational as well as textual attributes. We have demonstrated the usefulness of rx-anon at the example of two real-world datasets using the privacy model k-anonymity [28].

Data Availability and Reproducibility: The source code is available at https://github.com/Serpinex3/rx-anon as well as more technical details at https://arxiv.org/abs/2105.08842

REFERENCES

[1] Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K. Mohania. 2008. Efficient techniques for document sanitization. In *Int. Conf. on Information and Knowledge Mining (CIKM)*. ACM, 843–852.

[2] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *J. of the American Medical Informatics Association* 24, 3 (2017), 596–606. arXiv:1606.03475

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. ACL, 4171–4186.

[4] Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus. In *Int. Conf. on Recent Advances in Natural Language Processing*. Incoma Ltd., 259–269.

[5] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, and others. 2009. A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. *J. of the American Medical Informatics Association* 16, 5 (2009), 670–682.

[6] James Gardner and Li Xiong. 2008. HIDE: An Integrated System for Health Information DE-identification. In *Int. Symposium on Computer-Based Medical Systems*. IEEE, 254–259.

[7] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. 2007. Fast data anonymization with low information loss. In *VLDB*. ACM, 758–769.

[8] Olga Gkountouna and Manolis Terrovitis. 2015. Anonymizing Collections of Tree-Structured Data. *Trans. Knowl. Data Eng.* 27, 8 (2015), 2034–2048.

[9] Qiyuan Gong, Junzhou Luo, Ming Yang, Weiwei Ni, and Xiao Bai Li. 2017. Anonymizing 1:M microdata with high utility. *Knowledge-Based Systems* 115 (2017), 15–26.

[10] Yeye He and Jeffrey F. Naughton. 2009. Anonymization of Set-Valued Data via Top-Down, Local Generalization. *VLDB* 2, 1 (2009), 934–945.

[11] Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *ACM Conf. on Health, Inference, and Learning*. ACM, 214–221.

[12] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, Vol. 2006. ACM, 659–660.

[13] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. 2006. Mondrian Multidimensional K-Anonymity. In *ICDE*. IEEE, 25–25.

[14] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *ICDE*. IEEE, 106–115.

[15] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *J. of Biomedical Informatics* 75 (2017), S34–S42.

[16] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. L-diversity: privacy beyond k-anonymity. In *ICDE*. IEEE, 24–24.

[17] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119.

[18] Ishna Neamatullah, Margaret M. Douglass, Li-wei H Lehman, and others. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* 8 (2008), 32.

[19] Mehmet Ercan Nergiz, Christopher Clifton, and Ahmet Erhan Nergiz. 2007. MultiRelational k-Anonymity. In *ICDE*, Vol. 21. IEEE, 1417–1421.

[20] Giorgos Poulis, Grigorios Loukides, Aris Gkoulalas-Divanis, and Spiros Skiadopoulos. 2013. Anonymizing Data with Relational and Transaction Attributes. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 353–369.

[21] Patrick Ruch, Robert H. Baud, Anne-Marie Rassinoux, Pierrette Bouillon, and Gilbert Robert. 2000. Medical document anonymization with a semantic lexicon. In *American Medical Informatics Association Annual Symposium*. AMIA, 729–733.

[22] Pierangela Samarati. 2001. Protecting respondents identities in microdata release. *IEEE Trans. on Knowledge and Data Engineering* 13, 6 (2001), 1010–1027.

[23] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2013. Automatic general-purpose sanitization of textual documents. *IEEE Trans. on Information Forensics and Security* 8, 6 (2013), 853–862.

[24] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs*. AAAI, 199–205.

[25] Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system. In *Proceedings : a Conf. of the American Medical Informatics Association. AMIA Fall Symposium*. AMIA, 333–337.

[26] Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. In *Data Privacy Working Paper 3*. Carnegie Mellon U.

[27] Latanya Sweeney. 2002. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 571–588.

[28] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.

[29] Irene Teinemaa, Marlon Dumas, Fabrizio Maria Maggi, and Chiara Di Francescomarino. 2016. Predictive business process monitoring with structured and unstructured data. In *Business Process Management*, Vol. 9850. Springer, 401–417.

[30] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. 2008. Privacy-preserving anonymization of set-valued data. *VLDB* 1, 1 (2008), 115–125.

[31] Jan Trienes, Dolf Trieschnigg, Christin Seifert, and Djoerd Hiemstra. 2020. Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. In *Health Search and Data Mining*, Vol. 2551. CEUR, 3–11.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Conf. on Neural Information Processing Systems*. Curran, 5998–6008.

[33] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. 2006. Utility-based anonymization using local recoding. In *SIGKDD*. ACM, 785–790.

[34] Ying Zhao and Charles C. Zhou. 2020. Link Analysis to Discover Insights from Structured and Unstructured Data on COVID-19. In *Bioinformatics, Computational Biology and Health Informatics*. ACM, 1–8.