2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress

Identifying Data Exposure across Distributed High-Dimensional Health Data Silos through Bayesian Networks optimised by Multigrid and Manifold

Nikolai J. Podlesny and Anne V.D.M. Kayem and Christoph Meinel
Hasso-Plattner-Institute, University of Potsdam
Potsdam, Germany
Email: {Nikolai.Podlesny, Anne.Kayem, Christoph.Meinel}@hpi.de

***Abstract*—We present a novel, and use case agnostic method of identifying and circumventing private data exposure across distributed and high-dimensional data repositories. Examples of distributed high-dimensional data repositories include medical research and treatment data, where oftentimes more than 300 describing attributes appear. As such, providing strong guarantees of data anonymity in these repositories is a hard constraint in adhering to privacy legislation. Yet, when applied to distributed high-dimensional data, existing anonymisation algorithms incur high levels of information loss and do not guarantee privacy defeating the purpose of anonymisation. In this paper, we address this issue by using Bayesian networks to handle data transformation for anonymisation. By evaluating every attribute combination to determine the privacy exposure risk, the conditional probability linking attribute pairs is computed. Pairs with a high conditional probability expose the risk of de-anonymisation similar to quasi-identifiers and can be separated instead of deleted, as in previous algorithms. Attribute separation removes the risk of privacy exposure, and deletion avoidance results in a significant reduction in information loss. In other words, assimilating the conditional probability of outliers directly in the adjacency matrix in a greedy fashion is quick and thwarts de-anonymisation. Since identifying every privacy violating attribute combination is a W[2]-complete problem, we optimise the procedure with a multigrid solver method by evaluating the conditional probabilities between attribute pairs, and aggregating state space explosion of attribute pairs through manifold learning. Finally, incremental processing of new data is achieved through inexpensive, continuous (delta) learning.**

***Index Terms*—Data Anonymisation, Bayesian networks, multigrid, manifold**

I. Introduction

Growing data collections have increased the possibility of obtaining both direct and correlated data to describe users. Reports of de-anonymisation incidents have seen an increased news footprint resulting, not infrequently, in serious consequences [1] such as a class action lawsuit [2]. Such incidents are not limited to any industry and appear in almost any industry like telecommunication [3], transportation [4, 5] and financial transactions [6, 7]. The challenge of data anonymisation is compounded by the fact that state-of-the-art data gathering approaches create multi-attribute high-dimensional datasets. Anonymisation, however, is not only an ethical requirement but actually enforced by legislation in almost any country. In Europe, one of the strictest data privacy law has been recently introduced prohibiting the usage of data including personal identifiable information (PII) without obtaining explicit user consent. Therefore, many business cases and research projects fall back to anonymised data as a method of avoiding the impracticality of obtaining informed consent for all use-cases.

Problem Statement: Recent developments in the digital health field highlight the absence of strong solutions for identifying data exposure in distributed and high-dimensional data repositories (silos) properly [8, 9, 10]. Yet, the growing dimensions with often more than 200-400 describing attributes serve as an important knowledge-base for medical research and drug triage [11]. Simultaneously, strong guarantees of anonymity must be provided to abide to recent privacy regulation like GDPR [10]. While a plethora of existing solutions tackle the issue of anonymising individual data silos efficiently, to the best of our knowledge, efficiently generating anonymised distributed high-dimensional data silos that are privacy-preserving, still needs to be addressed.

Contributions: We propose a novel and use case agnostic method of identifying data exposure in multi-attribute, high-dimensional and distributed data repositories (silos). By learning probabilistic inferences between attribute values, risks of data exposure cannot only be identified but also be circumvented. Our contributions can be summarised as follows:

- We show that Bayesian networks can be used efficiently for anonymising high-dimensional data in distributed repositories. In addition, by using the properties of Bayesian network models, we can use inferential probabilities to identify risks of data exposure.
- Generating the inference graph happens by identifying all the possible attribute combinations that can result in inference and hence privacy exposure. In the Bayesian network, each node represents an attribute value, and each edge, the probability of inference if both attribute values appear in the same data record. To reduce the processing time of creating and searching large high-dimensional

978-1-7281-3024-8/19/$31.00 ©2019 IEEE
DOI 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00110

dataset representations for all possible combinations, we employ a multigrid solver method to aggregate nodes being in all probability.
- Since Bayesian networks are generated based on computed inferential probabilities of disclosure between attributes, there is the potential for a state-space explosion problem arising. We address this issue by aggregating the nodes to compress the inferential graph using manifold learning. This has the added benefit of reducing the graph's complexity.
- By leveraging on the reduced graph complexity, we are able to prevent homogeneity, background knowledge, and inferential attacks, by manipulating high risk conditional probabilities in the Bayesian network.
- In order to handle subsequent data changes, we employ continuous (delta) learning to lower data quality loss and ensure faster processing times for generating large anonymised distributed high dimensional datasets.

The rest of the paper is structured in the following manner: We discuss related work in Section 2. In Section 3, we introduce *Bayesian networks*. The concept including data interpretation, technical realisation, complexity reduction, identification of data exposure as well as continuous delta learning will be addressed in Section 4. This will be briefly evaluated in Section 5 and Section 6 summarises and concludes the contribution of this work.

II. Related Work

Anonymisation of health data has been studied extensively in a bid to find the answer to the problem of preventing disclosures of sensitive personal data. Given the NP-hard nature of established k-anonymity approaches and extensions [12] as well as the vulnerability of de-anonymisation based on exploiting the semantics of the data, the privacy community pivoted from syntactic data anonymisation towards semantic approaches like *differential privacy*. However, differential privacy and its statistical treatments are highly optimised for bulk data processing and predefined use cases [13, 14, 15], which knowledge might not be always available in advance. Also, postponing the anonymisation of data towards the runtime of a query leaves working surface for data leakage. Vulnerabilities regarding colluding users are also well known [16]. Leoni introduces "non-interactive" differential privacy by applying its statistical treatments a priori to a user query [17]. There are some publications that tackle the challenge of anonymising high-dimensional data sets. In the medical field, Kohlmayer et al. [18] present a flexible approach on top of k-anonymity, l-diversity and t-closeness as well as heuristic optimisation to anonymize distributed and separated data silos using a newly introduced Secure Multi-party Computing (SMC) protocol. Major challenges for distributed anonymisation in the healthcare industry are outlined by Mohammed et al. [19] who additionally propose *LKC-privacy* to achieve privacy in both centralized and distributed scenarios promising scalability for anonymising large datasets. *LKC-privacy* primarily leverages the thought that acquiring background knowledge is nontrivial and therefore limiting the length of quasi-identifier tuples to a predefined size L. In spite of the fact that this intention is fully comprehensible, it violates recent privacy regulations since anonymity cannot be guaranteed. Other work uses a MapReduce technique based on the Hadoop distributed file system (HDFS) to boost the computation capacities [20], this, however, does not resolve the exponential nature of anonymisation. Zhang et al. propose *PrivBayes* a data release technique leveraging Bayesian networks [21] and demonstrate impressively the accuracy of Bayesian theorem for anonymisation purposes. Yet, the *PrivBayes* leave space for improvements towards scalability through its NP-hard nature. Additionally, Zhang et al. outline future work for multi-database settings also in a non-distributed environment [21]. Using the advantages of Bayes' theorem is however still rare in the privacy field especially when it comes to vanquish high-dimensional datasets separated in several distributed data pools. In the data mining field, a variety of work exists to address privacy related constraints which includes but may not be limited to approaches including regression models [22], clustering [23] as well as classification through a naive Bayes approach [24]. These contributions typically strongly focus on data mining tasks in a specific application area with its related privacy models and constraints, particularly in merging the knowledge of distributed data sets while mining shared insights simultaneously ensuring privacy of each partition. Meng et al. present a random, projection-based method of using conditional probabilities in Bayesian networks for a set of linear equations given privacy-sensitive and distributed dataset across several parties [25]. A privacy-preserving protocol based on the K2 algorithm is introduced by Wright et al. [26] for learning the Bayesian network structure of distributed, heterogeneous data among different parties without revealing the underlying data. Zhang et al. investigated feature selection methods for high-dimensional micro-data with binary (categorical) features from a theoretical perspective [27].

This work uses the Bayesian theorem to identify data exposure in high-dimensional data silos. Unlike the environment of related work, those health data silos are distributed and often occur in multi-database settings. Significant processing optimisations through multigrid and manifold will be outlined to counteract Bayesian's NP-hard nature.

III. Bayesian Network

Bayesian networks are considered as a type of probabilistic (directed acyclic) graphical model and are primarily rest upon Bayes theorem [28]. Typically, Bayesian networks can be used to build models from existing data sources or domain experts and represent a set of random variables including their conditional dependencies through directed acyclic graphs. The Bayes theorem can also be described as the inter-dependency of two different events A and X, where A occurs knowingly that X already occurred. This corresponds to the conditional probability or dependency that A occurs when X has already taken place: $P(A|X)$. In most libraries, the graph is represented either with an adjacency matrix or an adjacency list. Adjacency matrices encode whether a combination of two nodes is adjacent

in the graph, meaning that there is a directed edge between said nodes, with a Boolean value in a two-dimensional array (see Table I). From a theoretical perspective, the Bayesian network requires the storage of $O(n^2)$ edges and doubles implementing the assigned probabilities, where n corresponds to the number of nodes. An adjacency list just uses the space to represent a node n times the number of edges e ($n \cdot e$ instead of n^2). Thus, adjacency lists are more efficient than adjacency matrices in terms of storage for sparse graphs (i.e. ones with few edges). However, one or the other is more efficient depending on the operations on the graph. Adjacency lists are better suited to return the neighbors of a node, where testing if nodes are adjacent is easier with adjacency matrices. Bayesian networks are not to be confused with naive Bayes classifiers nor Bayesian neuronal networks. A naive Bayes algorithm assumes a conditionally independent set of features enabling the usage of Bayes' theorem for probabilities while such an assumption is not applicable for Bayesian networks strictly requiring an a priori modelling of all dependencies.

IV. Bayesian Networks for Anonymisations

Introducing Bayesian networks for anonymisation processing promises several advantages. First, continuous, partial, use case agnostic and cheap updating of the underlying model is possible by just extending the model with the newly gathered datasets. Additionally, data exposure risks can be identified by interpreting the conditional probabilities within the network. Using the rounding error and actively intervene in the attribute value inferences as conditional probability, background knowledge as well as homogeneity are significantly exacerbated, and inferential attacks are scrambled. But it is also noteworthy, that better scalability in terms of storing size may be achieved, since only meta connections (probabilistic cases) are stored not the actual and duplicated ones. This comes with the disadvantage of time complexity. Since nodes in the context of Bayesian networks traditionally represent one single state, and a high-dimensional data set with many attributes holding many different states result in a huge growth of nodes, the NP-hardness of the exact [29] or even approximate [30] inference may be a bottleneck.

A. Data Interpretation

Bayesian Network represents occurrences of events, yet we are typically examining relational data. This requires us to interpreter those relational data in an event-like manner. Instead of seeing each data record itself in a common table representation (as in Figure 1a), we construe the probability of appearance of each data attribute value to each other. This can be depicted in a network-based way where each node represents on possible attribute value and the edges correspond to the likelihood of appearance in regards to the connected nodes (see Figure 1b). Exemplary, for both genders from the table, two nodes are drawn with edges to neighboring attribute values like their row-based linkage (ZIP, drug, disease). For reason of simplicity, non-existing attribute value combinations like $female$ and 14055 with a conditional probability of 0 are not linked with an edge in the network.

Correspondingly to the network topology, an adjacency matrix can be generated illustrated by Table I representing all edges between node pairs with their conditional probability. Such adjacency matrix is basically the technical representation of the described network with all node combinations of size 2. The conditional probability or rather attribute inference can be used now for identifying data exposure. For bidirectional edges, only half of the matrix may be filled, however, directed edges like in the present case requires a full matrix.

TABLE I: Adjacency matrix representation of relational data

	m	f	14055	12160	14052	Ibuprofen	Aspirin	Flu
m	0	0	0.5	0	0.5	0.6	0.3	1
F	0	0	0	1	0	0	1	1
14055	1	0	0	0	0	1	0	1
12160	0	1	0	0	0	0	1	1
14052	1	0	0	0	0	0.5	0.5	1
Ibuprofen	1	0	0.5	0	0.5	0	0	1
Aspirin	0	1	0	0.5	0.5	0	0	1
Flu	0.75	0.25	0.25	0.25	0.5	0.5	0.5	0

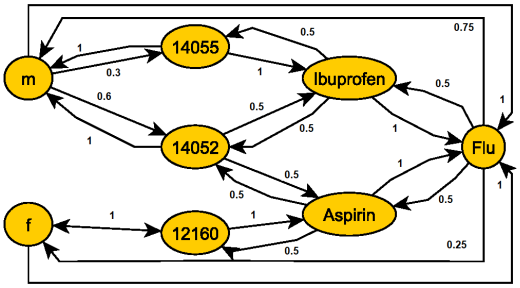
B. Adjacency Matrix: Implementation

The creation of an underlying adjacency matrix can be created in either an exact or approximate manner. While its well known that exact Bayesian model learning is NP-hard, Tsamardinos et al. proposed the approximation technique "hill climbing" to very efficiently implements a structure learning algorithm [31]. To minimise information loss and since approximate approaches are well explored, we focus on optimising exact learning techniques. Also, since only an exact approach can ensure anonymity and adhere to recent privacy legislation.

The most efficient exact processing we found so far to generate the adjacency matrix is illustrated below. Algorithm 1 pre-generates all column permutations $P(n,r) = \frac{n!}{(n-r)!}$ with the tuples size $n = 2$ (see line 5). Most programming languages expose a native function for generating permutation with a given length. Such list of column permutations may be processed in parallel and chunked accordingly to minimise inefficient process spawning and overheat in process management (see line 7). The results of the parallel execution are first collected as list and finally concatenated as bulk to minimise overheat once again (see line 8 and 10). Algorithm 2 delineates the actual dispatching of each adjacency matrix entry. Here, in detail, all present attribute value combinations for each attribute permutation from the original dataset like $female$ for gender and 12160 for ZIP are identified through a group by statement (line 4).. The actual probabilistic calculation is trivial as the count of appearance over the total number of attribute values (column size) in respect to the evaluated column tuple (line 7). This gives us already the conditional probability for

Gender	ZIP	DRUG	Disease
M	14055	Ibuprofen	Flu
F	12160	Aspirin	Flu
M	14052	Aspirin	Flu
M	14052	Ibuprofen	Flu

(a) Table representation of relational data



(b) Net representation of relational data

Fig. 1: Data transformation across different representation types

each attribute value tuple needed to build up an adjacency matrix. For assembling an equivalent matrix, we pre-populate both axis with all describing attributes values and inject at the intersection the corresponding conditional probability we have calculated earlier. Each of these probabilities serve as edge value directed by the column tuple setting, where the first element marks the edge's start and the latter marks the edge's end. The following example will illustrate the exact determination of the conditional probabilities:

Example 1: Given the data in Figure 1a, there are 12 attribute permutations under consideration: (gender, ZIP), (ZIP, drug), (drug, disease), (gender, drug), (gender, disease), (ZIP, disease) as well as their inverse parties. For each tuple, we calculate the likelihood of appearance for each value combination given the first tuple element value $(M, 14052) = 0.66$, $(M, 14055) = 0.33$, $(F, 12160) = 1$. In other words: How likely will the ZIP code be 14052 if gender equals M.

By refurbishing the original dataset in such way, we quickly generate and persist adjacency matrices like the sample one in Table I which serves as model for Bayesian networks like in Figure 1b. In the following, we will elaborate on exact optimization approaches.

C. Vertex (Node) Aggregation

Previously, we have mentioned the NP-hardness of the exact [29] and even approximate [30] inference calculation. Since our motivation is to process large and high-dimensional datasets, we looked into measures of reducing the given complexity. Multigrid solver, which incarnate the concept of coarsening the state space, may be transferable to this problem space. Multigrid methods are originated from numerical analysis and form a family of efficient algorithms to approximate solutions of equation systems derived from the discretisation of partial differential equations [32, 33]. One of the main differences is, that an adjacency matrix could be symmetric using multigrid (unidirectional arcs instead of directed arcs). Our multigrid solver method currently works in an exact way by aggregating nodes of the Bayesian network representation being in all probability defined by the conditional probability for each edge or correspondingly the inferences in the adjacency matrix.

Algorithm 1: Calculate probabilistic linkages

```
Calculate probabilistic linkages (table, settings);
  Input  : Table table containing the dataset
           Object settings with setting values
  Output : Array result including all attribute value tuples
           and their probability
result = initialize a table with columns
  ["value1","value2","likelihood"]
partial_tables = initialize an empty list
// Prepare the dataset for parallelize its execution;
colcom = create a list of permutations of table's columns
  with length of 2
create a process pool:
    split colcom in chunks and execute
      build_probability_for_column for each chunk in the
      process pool
        append the outcome from each process to the
          partial_tables
close the process pool
result = concatenate the partial_tables
```

The aggregation is purely based on the conditional probability along the edges emerging across all attribute values and only reflect node of disjunctive attribute subsets. As a result for such aggregation, a new node is created and replaces the former existing ones with new edges. This *exact approach* when being in all probability may be weakened to an *approximated one* for accelerate computation time but involves higher information loss. In the latter case, nodes are aggregated given some threshold for high likelihoods of collective appearances. First experiments support a threshold for a specific dataset as >95% conditional probability. The determination of such threshold should be realised in a greedy way to ensure anonymity overall. Leveraging such methodology dramatically increases the nodes under aggregation but might result in loosing 5% information as well especially the rare outliers. Respectively, by altering any threshold this balance between computation time and information loss is equilibrated. However, in this work we focus on leveraging the exact multigrid solver approach. The following example briefly illustrates the approach of

Algorithm 2: Building probabilities for attribute value tuples

build probability for column $(coltuple, table)$;
Input : Array $coltuple$ as list of two column names, Table $table$ containing the dataset
Output : Array $result$ including all attribute value tuples and their probability
// Transform $table$ to adjacency format through GROUP BY method;
group = initialize a table with columns ["value1","value2","likelihood"]
group = group $table$ by $coltuple$ and count group appearance
column_size = count the total column length
// Calculate their conditional probability;
group['likelihood'] = divide each group size by $column\_size$
return group

aggregating nodes:

Example 2: Picking up the previous example from Figure 1a, in case of $f \rightarrow 12160$ being in all probability, f_12160 is derived as a new state and simultaneously conflating the original states and transitions. Figure 2a illustrates the transformation where the two nodes f and 12160 are condensed. Similar to $Ibuprofin \rightarrow Flu$ being in all probability, $Ibuprofin\_Flu$ is introduced as new node (see Figure 2b).

This way, the exploding complexity in large and high-dimensional datasets can be captured. By reducing the number of nodes successively in an exponential space, the runtime for sampling and analysing decreases significant since fewer node combinations and therefore edges must be evaluated. First evaluations show promising results. To better understand the effectiveness of a multigrid techniques, we refer to the security research field. Here, multigrid approaches proved to solves the Poisson-Equation for elliptic curves in $O(n)$ instead of $O(n^{2.5})$ [34, 35].

In addition to multigrid, manifold learning could be used to aggregate those Bayesian network nodes as well. Manifolds as a topological space originates in mathematical physics where it has been used to reduce complicated geometric structures into simpler topological properties by resembling Euclidean space near the targeted point of observation. By formulating the dimensionality reduction problem as a classical problem in Riemannian geometry [36, 37], the number of nodes can be efficiently aggregated to reduce its processing complexity (see Figure 3). Various optimisation like the adaptive manifold learning proposed by Wang et al. [38] are used in different geometric research fields and could serve as basis for further state aggregation in Bayesian networks. Luke Olson and Jacob Schroder developed an algebraic multigrid (AMG) solver library [39] which implements a "multilevel technique for solving large-scale linear systems with optimal or near-optimal efficiency". Unlike geometric multigrid, AMG requires little or no geometric information about the underlying problem and develops a sequence of coarser grids directly from the input matrix. This feature is especially important for problems discretised on unstructured meshes and irregular grids.

Exactly those algebraic solver serve as basis to combine and subsume nodes and corresponding edges within our network to reduce its complexity especially during sampling.

D. Identification of Data Exposure & Ensuring Anonymity

The probabilistic linkage of attribute representatives across the dataset is characterised through the inference stated in an adjacency matrix. Such adjacency can serve as underlying model for a Bayesian network. Besides learning and sampling from this model, one is able to use the same inferences to identify risks of data exposure. Therefore, by iterating over the created adjacency matrix and summing all inferences for each cycle (row) the *summed cycle inference* (scf) metric is created:

$$scf = \sum_{i=1}^{m} cf_i \tag{1}$$

for m answering to the length (columns) of the adjacency matrix and t to the number of non-null values of the correspdoning row. Same applies to the row-based *mean cycle inference* (mcf)

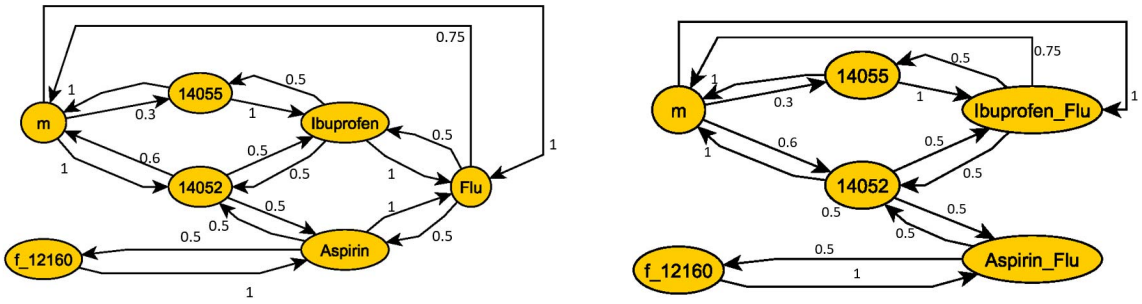
$$mcf = \frac{\sum_{i=1}^{m} cf_i}{t} \tag{2}$$

Both metrics grant insights on the likelihood of appearance for the corresponding attribute combinations. One may derive: **The higher the summed and mean cycle inference, the higher the risk of data exposure**. The following example will serve as illustration:

Example 3: For example in Table 1a, its adjacency matrix in Table I clearly addresses high inferences between $female \rightarrow 12160$, $female \rightarrow Aspirin$ and $female \rightarrow Flu$. This inference is marked through the conditional probability of 1 meaning in 100% of the cases those attribute values go along with each other. In fact, the weighted summed and mean inference is the highest in the entire matrix. When cross checking such injection function with the original example, the reader will quickly notice the uniqueness of the mentioned row.

Using the mean of the summed cycle inference and the corresponding deviation proofed to be efficient threshold to identify data exposure in an adjacency matrix. The higher the threshold, the more likely we miss data exposure risks. Consequently, the lower the threshold the more records fall into the observation and depending on the action may increase time complexity or information loss. However, the selection of such threshold can be implemented in a greedy way. As a result of a successful identification of a risky inference, one may gently alter the appropriate conditional probability in a way, that again the mean summed cycle inference approximate average values for eliminated any data exposure.

Definition 1: Anonymity Assurances
Let $\mu(t)$ be a suitable measure, that quantifies the data exposure risk of any cycle $t = \{c_1, ..., c_n\}$ given by the involved columns



(a) Combining exact dependencies: f + 12160

(b) Combining high probabilistic dependencies: drug + disease

Fig. 2: Complex Reduction through manifolds and multigrid

c_i. The set of cycles with length n containing the column c is denoted by $T_n(c)$. Any cycle t that deviates strongly from the statistics of $T_n(c)$, $c \in t$, i.e.

$$|E\left[\mu(T_n(c))\right] - \mu(t)| \leq \theta\, \sigma(\mu(T_n(c))), \tag{3}$$

where θ is a threshold parameter, E the expected value and σ the standard deviation, is likely to be a quasi-identifier and thus poses a potential anonymity risk.

Only, when no unique column combination remains, and therefore no individual record may be linked back to its original owner, anonymity is ensured. For a trustworthy comparison standard, unique columns (combination) answering to the maximal partial minimal unique column combinations (mpmUCC) in data profiling [40] may be identified to embrace all potential quasi-identifiers of all length and combination in a data set without separation of sensitive and non-sensitive. Afterwards, the high-risk cycle inferences and the list of mpmUCC can be compare to assess the overlap. Consequently, any candidates must be discriminated to avert de-anonymisation. Such treatment may occur through the deviation of punctiform and selective conditional probabilities in between nodes or by decreasing the overall summed cycle inference. Since these discrimination activities are simple subtraction operations, its realisation is rather simple. Finally, a Bayes model remains without the possibility to recreate any linkable unique column combinations. Without the feasibility of re-identifying individuals with the given data, the outcome may be considered anonymous.

Example 4: In the previous example of the adjacency matrix in Table I, high inferences between $female \rightarrow 12160$, $female \rightarrow Aspirin$ and $female \rightarrow Flu$ are identified. A quick sight on the original Table 1a confirms the unique column combination. By successively decreasing the affected inferences in a greedy manner towards the mean cycle inference corresponding to the average conditional probability in the matrix, risks of exposure can be exacerbated.

E. Continuous (delta) Learning

When working with large and high-dimensional data sets, the necessity of reprocessing the entire dataset is often a time and resource consuming activity. Therefore, approaches are desired where only partial differences must be reprocessed instead of the entire data set. Such partial processing in the machine learning field is known as delta learning, where a model is update on partial or newly data records and can be either a change or an evolution like in a differential equation. Bayesian networks may rely on adjacency matrices as model basis which can be selectively retrained, or simply extended with new records. Such extension is done by updating only selective conditional probabilities in the adjacency matrix instead of creating the entire matrix from scratch. This principle can be leveraged as delta learning especially across distributed data repositories (silos) by sharing the matrices itself rather than the raw data. Then, in each data silo, the model can be selectively extended with the given data and in case of new records partially retrained. For its accomplishment, the steps described in Algorithm 1 and Algorithm 2 are executed only on the newly introduced data records. As a consequence, only the corresponding conditional probabilities are updated or in case of new describing attributes added. Also, the multigrid solver as complexity reduction can be run for each of the distributed data repositories (silos) separately without a dependency to the other ones. It speaks for itself that higher aggregation and therefore better complexity reduction can be achieved for the solver if more context and therefore not only a partial model is available.

Periodically, a complete retraining is advisable since the more values exist in the original dataset, the more accurate the probabilistic linkages are. To ensure anonymity across each data silos, the data exposure risk by its summed cycle inference should be already identified and thwarted before passing on the model.

V. Evaluation

It is paramount to gain insights and assess the benefits and disadvantages of introducing Bayesian networks to identifying data exposure across high-dimensional health data silos. The evaluation will take place on a machine equipped with ten cores (Intel Xeon Gold 6140) and 32GB RAM using a semi-synthetic health data set with 109 attributes and 1M rows from multiple real world data sources and enriched with fake profile

data in close adjustment with real world data distributions. To understand the impact of Bayesian networks, Figure 3a delineates the time complexity in a high-dimensional context. It becomes clear that the execution time for creating the exact adjacency matrix growths proportional to the number of edges and permutations. The increase, however, stays relatively reasonable in contrast to the exponential explosion of edges. Since the trustworthiness of the data exposure identification is desired, Figure 3d compares the approach of summed cycle inferences with the actual unique column combinations (mpmUCCs) in the dataset, which may identify individuals uniquely [40]. In Figure 3c, the evolution of the mean summed cycle inference and its deviation is depicted. The mean summed cycle inference is in this context simple all inference for one record summed up, and the mean over all sums. With an increasing number of describing attributes, one observes a decreasing mean summed cycle inference as well as a decreasing standard deviation. That meets our expectations, since with more options the inferences gravitate towards its lower boundary. The spike can be ascribe to the addition of a describing attribute with a high cardinality. Statistical speaking, the outliers to the mean summed cycle inference still exist, it becomes however more complicated to find. Exactly those outliers represents those data records (rows) with a higher risk of compromising anonymity. As long as no unique tuple (mpmUCCs) remain which can be abused to link back original data records, we can ensure anonymity [40, 41]. Figure 3b compares the *data informative value* of sanitized data sets including established GPDR compliant anonymisation mechanisms from previous work as well as the presented approach of Bayesian network in a distributed setting. The performance of applying Bayesian network in a decentralized setting performs significantly better than corresponding GDPR compliant anonymisation algorithms like compartmentation, generalisation and suppression.

VI. Summary & Conclusions

In this paper, we presented a novel approach of using Bayesian networks for identifying and circumventing private data exposure in high-dimensional and distributed data silos. Through the accompanied transition of relational data into an adjacency matrix, several advantages have been deviated. Besides privacy compliant consolidation of high-dimensional data in distributed silos, we elaborated on leveraging manifolds and multigrid mechanisms for exact complexity reduction in Bayesian networks. We showed, how data exposure can be identified and thwarted through the inferential probability within some adjacency matrix. Further, continuous (delta) learning in the context of Bayesian networks was discussed especially for accelerating computation time and its effect on decentralized settings.

In future work, we will deepen the assessment of the presented approach towards its time performance, data exposure and potential on preventing homogeneity and background knowledge attacks. Modeling different attack scenarios require more insights and shall be picked up in future work as well. Additional efforts may be given to evaluate potential side effects when using probabilistic data representation in the anonymisation context and more experiments on time complexity reduction through manifolds and multigrid mechanisms are desirable as well.

We have made an exemplary source code and documentation publicly available[1].

References

[1] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *CoRR*, vol. abs/cs/0610105, 2006. [Online]. Available: http://arxiv.org/abs/cs/0610105

[2] M. Barbaro and T. Zeller, "A face is exposed for aol searcher no. 4417749," Aug 2006. [Online]. Available: http://www.nytimes.com/2006/08/09/technology/09aol.html

[3] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.

[4] I. S. Rubinstein and W. Hartzog, "Anonymization and risk," *Wash. L. Rev.*, vol. 91, p. 703, 2016.

[5] J. Polonetsky, O. Tene, and K. Finch, "Shades of gray: Seeing the full spectrum of practical data de-intentification," *Santa Clara L. Rev.*, vol. 56, p. 593, 2016.

[6] Y.-A. De Montjoye, L. Radaelli, V. K. Singh *et al.*, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015.

[7] P. Vessenes and R. Seidensticker, "System and method for analyzing transactions in a distributed ledger," Mar. 29 2016, uS Patent 9,298,806. [Online]. Available: https://www.google.com/patents/US9298806

[8] B. Kayyali, D. Knott, and S. Van Kuiken, "The big-data revolution in us health care: Accelerating value and innovation," April 2013.

[9] G. Aue, S. Biesdorf, and N. Henke, "ehealth 2.0: How health systems can gain a leadership role in digital health," Dec 2015.

[10] R. Massey, "How the gdpr will impact life sciences and health care," Feb 2017.

[11] E. Schadt and S. Chilukuri, "The role of big data in medicine," Nov 2015.

[12] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2004, pp. 223–228.

[13] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[14] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 503–512.

[15] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 338–340.

[16] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '11. New York, NY, USA: ACM, 2011, pp. 193–204. [Online]. Available: http://doi.acm.org/10.1145/1989323.1989345

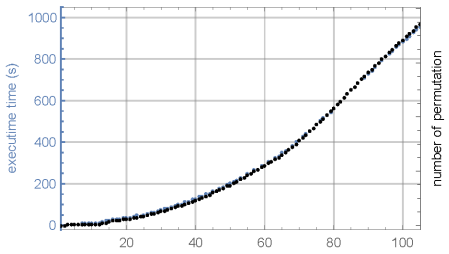
[17] D. Leoni, "Non-interactive differential privacy: a survey," in *Proceedings of the First International Workshop on Open Data*. ACM, 2012, pp. 40–52.

[18] F. Kohlmayer, F. Prasser, C. Eckert, and K. A. Kuhn, "A flexible approach to distributed data anonymization," *Journal of biomedical informatics*, vol. 50, pp. 62–76, 2014.

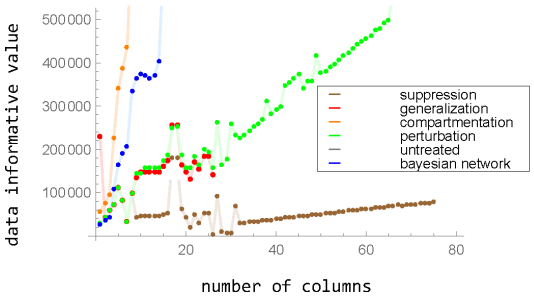
[19] N. Mohammed, B. Fung, P. C. Hung, and C.-K. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 4, p. 18, 2010.

[20] X. Zhang, L. T. Yang, C. Liu, and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 2, pp. 363–373, 2014.

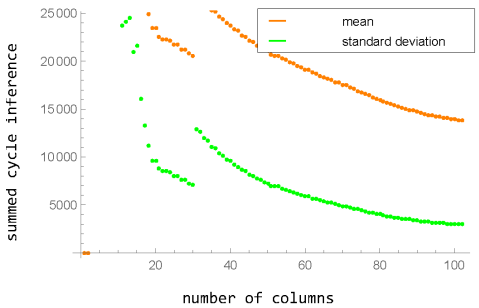
[1] https://github.com/jaSunny/MA-enriched-Health-Data



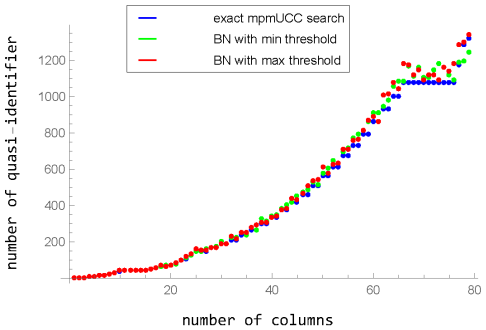
(a) Time Complexity and Permutation growth over number of columns



(b) Data informative value evolution of various GDPR compliant algorithms



(c) Mean summed cycle inferences and its standard deviation



(d) Precision of data exposure identification through Bayesian inferences

Fig. 3: Metric visualisation over the number of columns

[21] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, p. 25, 2017.

[22] S. E. Fienberg and J. Jin, "Privacy-preserving data sharing in high dimensional regression and classification settings," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, p. 10, 2012.

[23] J. Vaidya and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 206–215.

[24] J. Vaidya, M. Kantarcıoğlu, and C. Clifton, "Privacy-preserving naive bayes classification," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 17, no. 4, pp. 879–898, 2008.

[25] D. Meng, K. Sivakumar, and H. Kargupta, "Privacy-sensitive bayesian network parameter learning," in *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004, pp. 487–490.

[26] R. Wright and Z. Yang, "Privacy-preserving bayesian network structure computation on distributed heterogeneous data," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 713–718.

[27] B. Zhang, V. Dave, N. Mohammed, and M. A. Hasan, "Feature selection for classification under anonymity constraint," *arXiv preprint arXiv:1512.07158*, 2015.

[28] B. Efron, "Bayes' theorem in the 21st century," *Science*, vol. 340, no. 6137, pp. 1177–1178, 2013.

[29] D. M. Chickering, D. Geiger, D. Heckerman *et al.*, "Learning bayesian networks is np-hard," Technical Report MSR-TR-94-17, Microsoft Research, Tech. Rep., 1994.

[30] P. Dagum and M. Luby, "Approximating probabilistic inference in bayesian belief networks is np-hard," *Artificial intelligence*, vol. 60, no. 1, pp. 141–153, 1993.

[31] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[32] K. Stüben, "An introduction to algebraic multigrid," *Multigrid*, pp. 413–532, 2001.

[33] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A multigrid tutorial*. SIAM, 2000.

[34] S. R. Fulton, P. E. Ciesielski, and W. H. Schubert, "Multigrid methods for elliptic problems: A review," *Monthly Weather Review*, vol. 114, no. 5, pp. 943–959, 1986.

[35] P. Vaněk, J. Mandel, and M. Brezina, "Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems," *Computing*, vol. 56, no. 3, pp. 179–196, 1996.

[36] J. Carr, *Applications of centre manifold theory*. Springer Science & Business Media, 2012, vol. 35.

[37] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 796–809, 2008.

[38] J. Wang, Z. Zhang, and H. Zha, "Adaptive manifold learning," in *Advances in neural information processing systems*, 2005, pp. 1473–1480.

[39] L. N. Olson and J. B. Schroder, "PyAMG: Algebraic multigrid solvers in Python v4.0," 2018, release 4.0. [Online]. Available: https://github.com/pyamg/pyamg

[40] S. v. S. M. U. Nikolai J. Podlesny, Anne V.D.M. Kayem, "Minimising information loss on anonymized high dimensional data with greedy in-memory processing," in *International Conference on Database and Expert Systems Applications*. Springer, 2018.

[41] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving anonymization of set-valued data," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 115–125, 2008.