Background and Specification Progress Report

Matthew Beardwell

December 2021

Abstract

This project focuses on implementing, assessing, and adapting the algorithms described in [5] to other problems. I will discuss the paper's background in privacy preservation and outline the requirements, design, and specification of the programming deliverables.

Contents

1 Introduction

Sequential data mining can be found in areas such as business and biology. Companies profit, geneticists can analyse DNA and protein sequences, and end users get a better service. However, the release of sequential data comes with concerns for the individual's privacy. Using hospital patient RFID location tracking from Wang et al. [28] as an example, "some event patterns may be sensitive in that they reveal an individual's private information. For example, an observation that a doctor leaves a patient's room and then immediately enters a psychiatrist's office might serve as an indication that this patient is experiencing psychiatric problems" which poses the issue that legitimate sequential data analysis "may disclose private information about individual patients as a side-effect". Some works have looked at privacy preservation in DNA sequences [22], 'user context' streams based on mobile sensor data [18], 'transit data' [8], and user trajectories [27]. In this report, the focus is on the work of Bonomi et al. [5] who deal with anonymising non-time-stamped event sequences but in Section 2, I discuss the surrounding area of privacy preservation techniques with various data types and privacy goals. Section 3 looks into what existing works have tried and section 4 discusses the specification, requirements, and design of the reports deliverable - a program that implements the approaches in Bonomi et al. [5].

2 Background

Data mining often results in privacy violations either of individuals or of sensitive information that was often unknowingly contained in the data. The research area of privacy preserving data mining can be broken down into two main areas: *data anonymisation* and *data sanitisation* (also known as *knowledge hiding*).

Data anonymisation and data sanitisation are both methods to hide some information in data before it is analysed but they have different privacy goals. Literature define the criteria that decide whether the data is correctly anonymised or sanitised depending on their privacy goals and propose an algorithm that solves the problem of cleaning the data such that the utility of the data for mining is optimised and that the privacy criteria are met.

2.1 Data Anonymisation

Data anonymisation privacy goals aim "at preventing a data recipient from inferring information about individuals whose information is contained in the input dataset ... This includes inferences about the identity of an individual (identity disclosure), about whether or not an individual's information is contained in the output dataset (membership disclosure), as well as inferences that generally depend on an individual's information (inferential disclosure)." [2]. Re-identification of data that has users' names and identifying information 're-moved' is still possible by matching 'quasi-identifiers' - precise and often unique attributes contained in the data, such as age and sex, that when combined allow the data attributes to re-attributed to an individual because they are also contained in another dataset that has not had the names and identifying information suppressed.

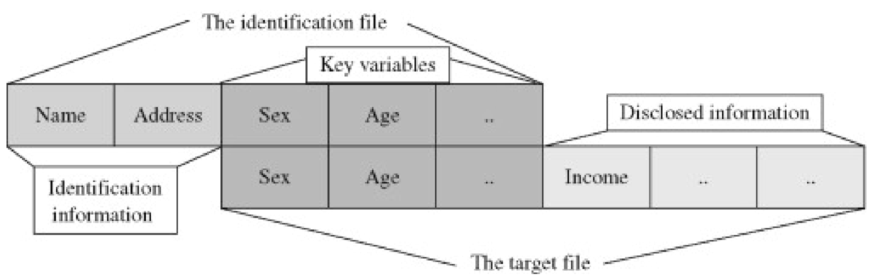

Figure 1: Diagram taken from [12] that illustrates how quasi-identifiers can be used in re-identification and attribute disclosure attacks. The QI's are shown as 'Key Variables'.

Figure 1 combines two records from different datasets, one with names attached and one without. The identities of individuals in the de-identified data are now revealed by comparing the datasets and further sensitive information about them can be attributed to them.

2.2 Data Sanitisation

As described in [3], “Data sanitization ... aims at concealing patterns modeling confidential knowledge by limiting their frequency, so that they are not easily mined from the data”. An example of confidential knowledge is business information that might give a competitive advantage which, when released, may damage the business releasing it. The release of business information for data mining such as selling product sales data might be advantageous but can only be profitable if no proprietary knowledge that gives a competitive advantage is given. Suppression is not sufficient as mining patterns in data may allow for compromising inferences that may not have been known to be contained in the data prior to its release. [1] outlines a privacy concern in sequential pattern mining - “In mobility data, some typical mobile behaviors (i.e., frequent patterns) might be considered sensitive for political or security reasons. For instance, mobility patterns corresponding to national security activities”. More sophisticated theoretical techniques will have to be developed to conceal this information. Methods that conceal confidential information are discussed in Section 3.

3 Literature Review

Data anonymisation works can be categorised by the operations they perform on the data to achieve their privacy goals. Eyupoglu et al. [13] categorises these operations into five main types - “In order to preserve privacy, there are five types of anonymization operations, namely generalization, suppression, anatomization, permutation and perturbation.”.

Generalisation approaches [22, 20, 26, 23] work by replacing some values with more non-specific ones and typically rely on heuristics. In the case of non-numerical data, a value such as 'Male' could be generalised to 'Gender'. In numerical data, a value such as 26 could be put into a bin of size 10, replacing it with the range [20,29]. Suppression (withholding values, e.g. replacement of some data with asterisks) is less common - “The drawback of suppression-based approaches is the data loss inflicted by deleting symbols in the released data” [5]. Susan and Christopher [25] outline anatomisation in their approach which “dissociates the correlation observed between the quasi identifier attributes and sensitive attributes (SA) and yields two separate tables with non-overlapping attributes” thus preventing re-identification attacks. Permutation [2], “disassociates a relation between a quasi-identifier and sensitive attribute by dividing a number of data records into groups and mixing their sensitive values in every group” [13]. Perturbation [9, 6] “tampers the data by the addition of noise, aggregation of values, swapping of values, or generation of artificial data or by the encryption of the data” [25]. Differential privacy, a widely used mechanism for achieving data anonymisation, relies on perturbation by adding noise.

Heuristics have largely been used to solve data anonymisation problems [6, 7, 14, 20, 22, 23, 26, 2]. Duchi et al. [11] solve with stochastic gradient descent to satisfy differential privacy. Fawaz and du Pin Calmon [10] solve using line search.

Some knowledge hiding problems rely on suppression [21, 1, 15, 18, 24, 28, 16], perturbation [17, 4], and permutation [19]. Greedy heuristics are quite common in data sanitisation [1, 15, 18, 24]. Integer programming [16, 4], dynamic programming [21], and binary integer programming [17] also appear. Wang et al. [28] develop a hybrid approach combining linear programming and some optimisation heuristics. Gwadera et al. [19] is another example of a hybrid approach that mixes some heuristics with conditional random search.

4 Specification, Design, and Requirements

4.1 Problem Overview

Bonomi et al. [5] implements heuristics to solve the 'Minimum Utility Loss Generalisation' problem where the goal is to minimise the utility lost in the data while still satisfying privacy goals based on mutual information. The data consists of a set of sequences with subsequence patterns that model confidential knowledge. The sensitive subsequence patterns are hidden through generalisation - a technique whereby symbols are replaced by ancestor symbols in a taxonomy tree.

Two algorithms are proposed in Bonomi et al. [5]. The top-down approach replaces all symbols with their most distant ancestor in the taxonomy tree and refines each, increasing the utility metric, until the privacy criteria are not met. The bottom-up approach starts off with the optimal utility as no generalisation is applied and replaces symbols by their ancestors, decreasing the utility metric, until the privacy goal is finally met. The conlusion is that "the bottom-up approach achieves similar utility results as the top-down approach while incurring considerably lower running time" [5].

In this work, the goal is to implement the algorithms in the work, evaluate their effectiveness, and adapt the approaches to other problems.

4.2 Design

The algorithms will be programmed in C++. User interaction will take place through the console and data will be input and output through the presence and writing of .txt files. There will also be a possibility to input and display data through the console. Below, I outline the specific requirements for the program.

4.3 Requirements

Few requirements for the implementation of the algorithms discussed in Bonomi et al. [5] are necessary. Here I discuss functional and user requirements of the implementation.

4.3.1 Functional Requirements

1. The top-down and bottom-up algorithms should be implemented in C++

   > *There are no requirements for how this program file should be compiled, executed, or distributed.*

2. The program will read from a .txt file contained in the same directory as the program file a set of arrays of strings. Each array of strings represents a sequence for which we aim to hide sensitive patterns. The file will also need to contain information for a taxonomy tree T, and a privacy level ϵ.

3. The format required will be told to the user through the use of a console and if the read file is not in the correct format, an appropriate error will be given.

4. The program will write to a .txt file two sets. One where the sequences have been sanitised by the top-down approach and the second by the bottom-down approach. This will also be accompanied by the utility metrics achieved by each approach on the data.

5. The program will provide an option to input the data, that would otherwise be given in the text file, through a console.

4.3.2 User Requirements

1. User interaction in the console must be straight forward to read and understand. Appropriate errors will be given if necessary.

2. If a user requests to input the data through the console, it must be straight forward to do so given that the console must explain to the user the formatting requirements.

3. If a user inputs the data through the console, a prompt will ask the user if they would like the sanitised sequences to be saved to a file. If not, it is written to the console. Utility metrics will be given in the console regardless as well as written to the file.

References

[1] Osman Abul, Francesco Bonchi, and Fosca Giannotti. Hiding sequential and spatiotemporal patterns. *IEEE Transactions on Knowledge and Data Engineering*, 22(12):1709–1723, 2010.

[2] Giulia Bernardini, Huiping Chen, Alessio Conte, Roberto Grossi, Grigorios Loukides, Nadia Pisanti, Solon P Pissis, Giovanna Rosone, and Michelle Sweering. Combinatorial algorithms for string sanitization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(1):1–34, 2020.

[3] Giulia Bernardini, Huiping Chen, Alessio Conte, Roberto Grossi, Grigorios Loukidis, Nadia Pisanti, Solon Pissis, and Giovanna Rosone. String sanitization: A combinatorial approach. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) 2019*, 2019.

[4] Giulia Bernardini, Alessio Conte, Garance Gourdel, Roberto Grossi, Grigorios Loukides, Nadia Pisanti, Solon P Pissis, Giulia Punzi, Leen Stougie, and Michelle Sweering. Hide and mine in strings: Hardness and algorithms. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 924–929. IEEE, 2020.

[5] Luca Bonomi, Liyue Fan, and Hongxia Jin. An information-theoretic approach to individual sequential data sanitization. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 337–346, 2016.

[6] Luca Bonomi and Li Xiong. A two-phase algorithm for mining sequential patterns with differential privacy. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 269–278, 2013.

[7] Jianneng Cao, Panagiotis Karras, Chedy Raïssi, and Kian-Lee Tan. ρ-uncertainty: inference-proof transaction anonymization. *Proceedings of the VLDB Endowment (PVLDB)*, 3(1):1033–1044, 2010.

[8] Rui Chen, Benjamin CM Fung, Bipin C Desai, and Nériah M Sossou. Differentially private transit data publication: a case study on the montreal transportation system. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–221, 2012.

[9] Rui Chen, Benjamin CM Fung, Bipin C Desai, and Nériah M Sossou. Differentially private transit data publication: a case study on the montreal transportation system. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–221, 2012.

[10] Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1401–1408. IEEE, 2012.

[11] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[12] Mark Elliott. Statistical disclosure control. *Encyclopedia of social measurement*, 2005.

[13] Can Eyupoglu, Muhammed Ali Aydin, Abdul Halim Zaim, and Ahmet Sertbas. An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, 20(5):373, 2018.

[14] Liyue Fan and Hongxia Jin. A practical framework for privacy-preserving data analytics. In *Proceedings of the 24th International Conference on World Wide Web*, pages 311–321, 2015.

[15] Aris Gkoulalas-Divanis and Grigorios Loukides. Revisiting sequential pattern hiding to enhance utility. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1316–1324, 2011.

[16] Aris Gkoulalas-Divanis and Vassilios S Verykios. An integer programming approach for frequent itemset hiding. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 748–757, 2006.

[17] Aris Gkoulalas-Divanis and Vassilios S Verykios. Exact knowledge hiding through database extension. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):699–713, 2008.

[18] Michaela Götz, Suman Nath, and Johannes Gehrke. Maskit: Privately releasing user context streams for personalized mobile applications. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 289–300, 2012.

[19] Robert Gwadera, Aris Gkoulalas-Divanis, and Grigorios Loukides. Permutation-based sequential pattern hiding. In *2013 IEEE 13th International Conference on Data Mining*, pages 241–250. Ieee, 2013.

[20] Yeye He and Jeffrey F Naughton. Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*, 2(1):934–945, 2009.

[21] Grigorios Loukides and Robert Gwadera. Optimal event sequence sanitization. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 775–783. SIAM, 2015.

[22] Bradley A. Malin. Protecting dna sequence anonymity with generalization lattices, 2004.

[23] Reza Sherkat, Jing Li, and Nikos Mamoulis. Efficient time-stamped event sequence anonymization. *ACM Transactions on the Web (TWEB)*, 8(1):1–53, 2013.

[24] Xingzhi Sun and Philip S Yu. A border-based approach for hiding sensitive frequent itemsets. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[25] V Shyamala Susan and T Christopher. Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes. *SpringerPlus*, 5(1):1–21, 2016.

[26] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Local and global recoding methods for anonymizing set-valued data. *The VLDB Journal*, 20(1):83–106, 2011.

[27] Manolis Terrovitis, Giorgos Poulis, Nikos Mamoulis, and Spiros Skiadopoulos. Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1466–1479, 2017.

[28] Di Wang, Yeye He, Elke Rundensteiner, and Jeffrey F Naughton. Utility-maximizing event stream suppression. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 589–600, 2013.