Finding ways to aggregate patient records so as to preserve patient privacy while also providing useful information to medical researchers

Kevin Raina, Dr.Aaron Smith (Supervisor)
Department of Mathematics-Statistics

Introduction

Canadian medical researchers would like to be in possession of patients' medical records as the amount of useful information in the data allows researchers to analyze trends and notice irregularities. There is a lot of useful work that can be done with these medical records, but releasing such data to medical researchers would be considered a breach of the patients' privacy. For this reason, the researchers are unable to access the raw data; however, this does not limit them from accessing and using anonymized data: data where the individual can not be detected from their record. There are many ways to anonymize health data, but a common method is to aggregate (combine) patient records. In order to perform the aggregation in an optimal manner, it is necessary to understand the trade-off between privacy and usefulness, and to take both properties into consideration. In this project, we will implement computational approaches to determine the optimal aggregation of patient records so as to preserve privacy while still providing useful information to researchers.

Methodology

1.Aggregation Strategy:
The population of interest is all Canadians living in census sub-divisions, as defined by Statistics Canada in [10], with populations of 5000 or more. Types of sub-divisions include Municipalities, Cities, Towns, Villages etc. Furthermore, the size of the population of interest is 29,383,430. Intuitively, our data consists of 29,383,430 records. Table 1 illustrates how a particular record would appear in the database.

Name	Country	Province	Division	Sub-Division	Age	Gender	Medical Status for Influenza
Vik Chisthi	Canada	Ontario	Lambton	Lambton Shores	22	M	Recovered

Table 1: A patient's record in the database

The person's medical status for Influenza is a sensitive variable, their gender, location and age are indirect identifiers and their name is a direct identifier [2]. We will consider 16 aggregations of the 29,383,430 records based on all combinations of the nodes (vectors) from domain generalizations [1] as shown in Figure 1.

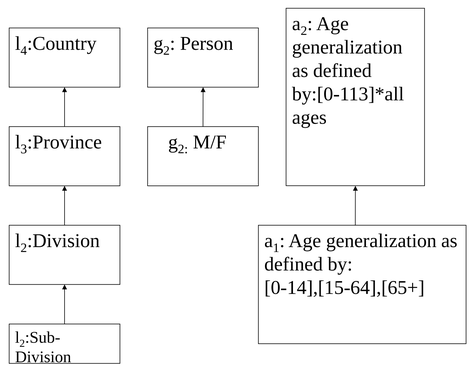

Figure 1: Domain generalizations of three quasi-identifiers: location, gender and age

To understand what a particular node in Figure 1 represents, consider L1. L1 partitions the patient's location based on country ,province, census division and census subdivision as defined by Statistics Canada [9,10]. Each point in Figure 2 depicts a part in the partition of our population of interest in British Columbia. Each colour in Figure 2 aggregates these parts/census subdivisions into census divisions.

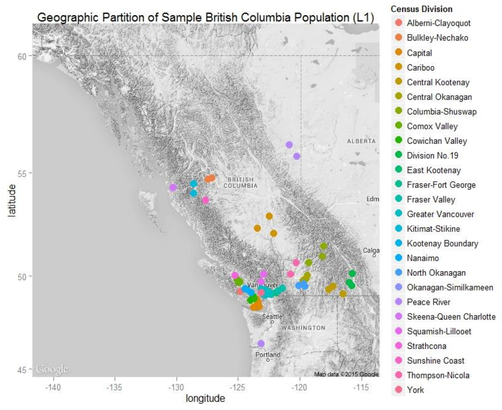

Figure 2: Geographic Partition of sample British Columbia Population (L1)

2. Objective Functions

2.1 Privacy

We require our partition to have the property P(re-identification) < T, where T is a set threshold. This means that very little information about any individual is recoverable, and the data is considered acceptable to release [11,12].

2.2. Information loss

We will also consider a measure of information loss as defined in [6]. Information is lost when partitioning the data in a certain way and this loss is typically larger for broadly defined partitions. We will attempt to minimize the amount of information loss if we are presented with more than one acceptable partition.

3. Optimization Problem/Algorithm

We want to find the partition with the minimal information loss, that satisfies P(re-identification) < T. A partition that satisfies this condition will be defined as the optimal partition. In order to do this, we will run an optimization algorithm mentioned in [7]. Briefly, our algorithm takes as input a lattice of partitions/nodes as constructed in [7] as well as a given T, and outputs the optimal node. Figure 3 illustrates how a lattice would appear for our aggregation strategy.

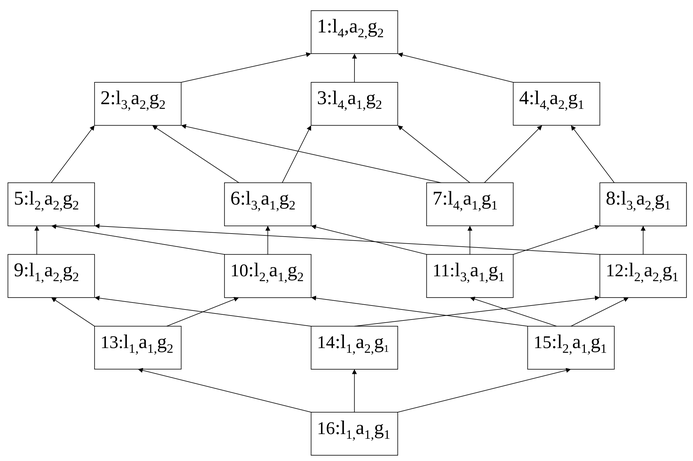

Figure 3: A Lattice as defined in [1], that corresponds to our aggregation strategy

Results

We found the partition that satisfies the optimization problem in section 3, for 0 0007 < T < 1. This was enough to show the distribution of the optimal node based on T for all possible values of T (0 to 1). Chart 1 depicts the optimal node based on the value of 100T. Hence, if we were given a threshold, we would be able to determine the optimal partition. An information loss measure was computed over the range of T in Chart 2.

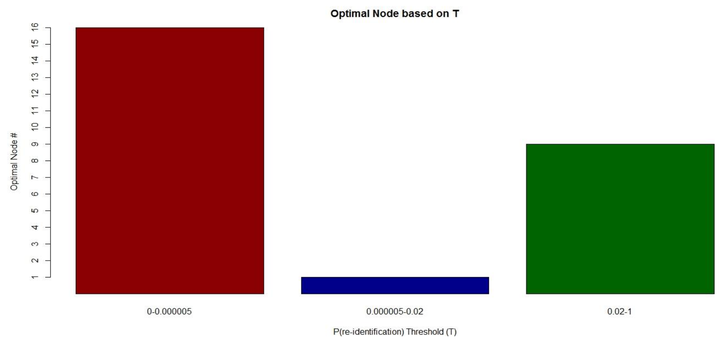

Chart 1: Optimal node based on the value of 100T

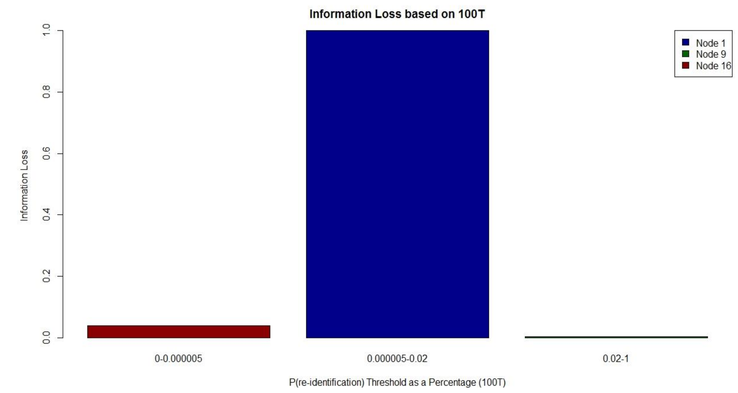

Chart 2: Information loss based on the value of 100T

Conclusion

The interpretation of the results would be as follows: Let us assume, that T = 0.05. Then, as mentioned in [5], we are setting the threshold to account for a very risky environment. From chart 1, we can tell that the optimal node is node # 9. Hence, given the threshold, the optimal way to aggregate patient records, is to aggregate them solely by census sub-division. We see from chart 2 that node # 9 is as good as it gets for information loss compared to the other optimal nodes. It should also be noted that the trade-off between privacy and usefulness still holds for reasonable values of T.

Suggestions for future work

We only carried out this project for an aggregation strategy containing 16 partitions. There are millions of ways to aggregate data, and expand the number of options so as to make the optimization more effective. In addition, our utility metric was revolved around properties of the data itself, rather than being problem-related . Hence, implementing the algorithm for several utility functions corresponding to different scenarios would be ideal and may change the distribution of the optimal node based on the threshold risk. Furthermore, considering other optimization algorithms to improve computation time would also be useful.

Contact

Kevin Raina
Department of Mathematics-Statistics/ University of Ottawa
Email: krain033@uottawa.ca

Acknowledgements

I would like to thank Dr.Smith for supervising the project and U.R.O.P. for giving me this opportunity.

References

1.El Emam, K. (2013). De-Identification Methods. In *Guide to the De-Identification of Personal Health Information* (pp. 234-235). Boca Raton: Taylor & Francis Group,LLC.
2.El Emam, K. (2013). Scope, Terminology and Definitions. In *Guide to the De-Identification of Personal Health Information* (pp. 127-128). Boca Raton: Taylor & Francis Group,LLC.
3.El Emam, K. (2013). Measuring the Probability of Re-identification. In *Guide to the De-Identification of Personal Health Information* (p. 187). Boca Raton: Taylor & Francis Group,LLC.
4.El Emam,K.,Kamal Dankar,F.(2009) In *Protecting Privacy Using k-Anonymity*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/
5. El Emam, K. (2013). Choosing Metric Thresholds. In *Guide to the De-Identification of Personal Health Information* (p. 228). Boca Raton: Taylor & Francis Group,LLC.
6. Xu,J., Wang,W.,Pei,J.,Wang,X.,Shi,B.,Wai-Chee Fu,A. (2006). In *Utility-Based Anonymization Using Local Recording.* http://www.cs.cuhk.hk/~adafu/Pub/localrecoding-kdd06.pdf
7.El Emam,K.,Kamal Dankar,F.,Romeo,I.,Jonker,E.,Amyot,D.,Cogo,E.,...Bottomley,J.(2009). In *A Globally Optimal k-Anonymity Method for the De-Identification of Health Data.* Retrieved from http://jamia.oxfordjournals.org/content/jaminfo/16/5/670.full.pdf
8. Statistics Canada. 2012. "Age (131) and Sex (3) for the Population of Canada, Provinces,Territories,Census Divisions and Census Subdivisions, 2011 Census." "Canada,Provinces,Territories,Census Divisions and Census Subdivisions." "2011 Census of Canada: Topic-based tabulations." *Census*. Statistics Canada Catalogue no. 98-311-XCB2011023. Ottawa,Ontario.Last updated January 17th, 2013. http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/tbt-tt/Rp-eng.cfm?LANG=E&APATH=3&DETAIL=1&DIM=0&FL=A&FREE=0&GC=0&GID=0&GK=0&GRP=1&PID=102010&PRID=0&PTYPE=101955&S=0&SHOWALL=0&SUB=0&Temporal=2011&THEME=88&VID=0&VNAMEE=&VNAMEF= (Accessed January 13,2014).
9.Statistics Canada.(2012). *Census Division (CD)*. Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo008-eng.cfm
10. Statistics Canada.(2012). *Census subdivision (CSD). Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo012-eng.cfm*
11. El Emam, K. (2013). Measuring the Probability of Re-identification. In *Guide to the De-Identification of Personal Health Information* (p. 177). Boca Raton: Taylor & Francis Group,LLC.
12. El Emam, K.(2013). Measuring the Probability of Re-Identification. In *Guide to the De-Identification of Personal Health Information* (p.180). Boca Raton: Taylor & Francis Group,LLC.