2012

An Iterative Association Rule Mining Framework to K-Anonymize a Dataset

Michael Hayes
Western University, mhayes34@uwo.ca

Miriam A M Capretz
Western University, mcapretz@uwo.ca

Jefferey Reed
Western University, jreed26@uwo.ca

Cheryl Forchuk
Western University, cforchuk@uwo.ca

Citation of this paper:
Hayes, Michael; Capretz, Miriam A M; Reed, Jefferey; and Forchuk, Cheryl, "An Iterative Association Rule Mining Framework to K-Anonymize a Dataset" (2012). *Electrical and Computer Engineering Publications*. 36.
https://ir.lib.uwo.ca/electricalpub/36

An Iterative Association Rule Mining Technique to K-Anonymize a Dataset

Michael Hayes, Miriam A.M. Capretz
Department of Electrical and Computer Engineering
Western University
London, Ontario, Canada N6A 5B9
Email: {mhayes, mcapretz}@uwo.ca

Jeff Reed, Cheryl Forchuk
Arthur Labatt Family School of Nursing
Western University
London, Ontario, Canada N6A 5B9
Email: {jreed26, cforchuk}@uwo.ca

ABSTRACT

Preserving and maintaining client privacy and anonymity is of utmost importance in any domain, and especially so in healthcare, as loss of either of these can result in legal and ethical implications. Further, it is sometimes important to extract meaningful and useful information from existing data for research or management purposes. In this case it is necessary for the organization who manages the dataset to be certain that no attributes can identify individuals or groups of individuals. This paper proposes an extendable and generalized framework to anonymize a dataset using an iterative association rule mining approach. The proposed framework also makes use of optional domain rules and filter rules to help customize the filtering process. The outcome of the proposed framework is a preprocessed dataset which can be used in further research, with confidence that anonymity of individuals is conserved. Evaluation of this research will also be described in the form of a case study using a test dataset provided by the Lawson Health Research Institute in London, Ontario, Canada as a part of their Mental Health Engagement Network (MHEN) study.

I INTRODUCTION

Protecting the privacy and anonymity of individuals in a database is crucial for maintaining the integrity of the organization who manages the data. Privacy and anonymity-protection is arguably even more important in fields such as healthcare where breaches in privacy or anonymity result in legal or ethical implications. This problem becomes even more difficult when it becomes necessary to mine useful and meaningful relationships from the data. For example, revealing a new trend or relationship in the healthcare field can allow researchers to develop new techniques and advances which can help save lives, prevent illness or prepare for periods of a higher influx of clients. For instance, understanding that the flu in Canada is most common in late September to early November allows hospitals and clinics to better prepare themselves for the increase of in-clients [1].

In addition, with the growing popularity of integrating new technology into healthcare and other industries, it is becoming increasingly important for healthcare providers to be able to empower their clients and themselves with technology while still maintaining a high level of privacy and security. The healthcare industry is one in which much of the information that flows between different nodes in a healthcare environment (care providers, lab results, lab assistants, etc.) is considered personally identifiable information (PII). PII deals with any information which can identify a single client within a system [2]. Some examples of PII include Health Card Numbers, Social Insurance Numbers or Postal Codes if the sample size is small enough. PII can also be a combination of several attributes; in a larger data set the combination of Last Name and Birthday may uniquely identify a client so the pair {Last Name, Birthday} can be considered PII. In any healthcare setting, a breach of a client's PII can lead to severe or catastrophic results which has led to many legislative responses such as PHIPA [3] in Canada and HIPAA in the United States [4] which aim to protect client's PII. Knowing all this, it is very difficult, even for a domain expert, to identify all attributes and combinations of attributes which are personally identifiable or sensitive to an individual.

Currently, many organizations rely on expert knowledge and experience to determine what attributes need to be protected to maintain their client's privacy, however this is not always sufficient. Beyond this, it is very difficult for external researchers to attempt to extract meaningful relationships from a health-organizations data. Many times for a dataset to be distributed to research groups someone must manually go through the dataset the researchers wish to use and "black-out" the data which can be considered personally identifiable to a client. This process is long, arduous and delays the research process for many projects. Even for non-research oriented projects, should a dataset need to be distributed to a healthcare team, it must be filtered such that the healthcare team cannot identify individual clients. There have been some solutions proposed to this problem; for example, research by Friedman et al. [5] developed extensive work in providing k-anonymity to data mining techniques. K-Anonymity is a privacy model used to ensure that an external attacker can only identify groups of tuples of k-size [5]. Their work consisted of applying the k-anonymity model to data mining techniques such as clustering and association rule mining, most notably developing decision trees based on the outputs of these techniques. Other work by Seisungsittisunti and Natwichai [6] describe an incremental process to ensure privacy is maintained throughout association rule mining. These techniques make assumptions that the dataset will undergo association rule mining after the k-anonymity step.

Given the limitations of strict regulations, a need to perform research on sensitive data, and few real-world implementations to remedy the limitations, this paper will propose a framework that will allow any organization, healthcare providers included, to perform a pre-processing step which will find those attributes that can identify a subset of *k*-size in the dataset. This will be done using an iterative association rule-mining algorithm which makes use of dynamic class attributes to define k-subsets in the dataset. Knowing what attributes can narrow the dataset to such a size, the framework also provides a filtering step which will anonymize those attributes such that they can only define a broader subset of individuals. This framework will be extendable such that the organization or custodian of data can decide how small or large *k* should be, what rules to follow to perform the filtering algorithm and any exceptional attributes that should not be included in the filtering process. In addition, the framework can sit atop any existing infrastructure and needs only the raw dataset to produce a k-anonymized dataset. Ultimately, with this framework in the hands of healthcare custodians, the datasets which they govern will be k-anonymous and thus distributable to research and management groups.

The following sections of the paper are organized as follows: Section II will describe related works in the field of k-anonymizing datasets, and privacy in healthcare. Section III will introduce the technical concepts that will be used in the formalization of this framework, including data mining techniques and k-anonymization. Section IV will outline the approach taken by the proposed research. The framework will be applied in a case study in Section V; this will include background information on the current state of healthcare in Ontario, Canada as well as the project underway at the Lawson Health Research Institute in London, Ontario (Lawson). Finally, Section VI will describe concluding thoughts and ideas for future work in this area.

II RELATED WORK

Privacy- and anonymity-preserving algorithms and methods have been popular topics in recent years with research focused on how to maintain the privacy and anonymity of large datasets. Research proposed by Aggarawal and Yu [7], Agrawal and Srikant [8], and Seisungsittisunti and Natwichai [6] all focus on promoting privacy in the data mining field. The research proposed in these papers aim to tackle the problem of privacy or anonymity-preservation using different techniques. Each technique has drawbacks and benefits; for instance Seisungsittisunti and Natwichai [6] discusses privacy preservation using incremental steps to associatively classify a dataset. Closely related to the work proposed in this paper, there are similar benefits in accuracy. However, their research describes a specific algorithm in lieu of a generalized framework that can be applied to several domains. Also dissimilar, their research takes a post-processing route to determining the association rules of the data, where as the framework proposed in this paper incorporates Association Rule Mining (ARM) in the pre-processing step to ensure high levels of confidence. Their work is not evaluated against any dataset; instead, examples are given throughout the paper to make concepts clearer.

Aggarwal and Yu [7] and Agrawal and Srikant [8] both discuss approaches to providing privacy whilst data mining. Aggarwal et al. use an approach known as condensation, where the dataset is split into groups of k-size before the algorithm begins. Subsequently, each group is then anonymized separately based on statistical factors. The drawback in their research is that for the algorithm to produce sufficiently effective anonymity, the anonymous group size must be at least 15-20 individuals; whereas the general framework proposed in this work allows for any k-size to produce sufficiently anonymous results. Agrawal and Srikant [8] take a different approach, using Gaussian or Uniform perturbations to produce randomizations in the data. However, their approach does not allow for a general "privacy-level" to be defined, and thus the user cannot be guaranteed of any privacy. Other recent work has proven that the perturbation approach actually provides no privacy at all [9].

Research done by Brickell and Shmatikov [10] and Yang et al. [11] focus on providing anonymity in data mining but do not have a focus in privacy or k-anonymity as cornerstones to the research. Brickell and Schmatikov [10] have done work on creating efficient anonymity-preserving data collection. This work is closely related to cryptography and they do introduce many cryptographic processes; however, the over-arching theme of their research is quite similar to our proposed research in creating a protocol to anonymize data mining. Yang et al. [11] also claim to creating an anonymous-aware method of data-collection. Their research takes a less cryptographic approach and uses a more data mining technical oriented approach. The proposed research of Yang et al. is closely related to the proposed research of this paper as it provides an anonymous-aware protocol; however, Yang et al. do not use k-anonymity to provide a k-anonymous framework as this research does. Taking an approach which does not use k-anonymity constrains the research to pre-defined privacy levels, unless another proposition is made to extend the research. There is the benefit of creating a solution which is optimized for certain situations. This is less desirable in some situations, particularly in healthcare where many organizations do not share a unified database schema and more generalized solutions are necessary - such as those incorporating k-anonymity.

A central focus of this paper is on protecting client's privacy through anonymity, especially in the healthcare setting. Atzori et al. [12] provide an in-depth look to promoting privacy awareness in data mining, specifically for healthcare in Europe. Atzori et al. tend to focus on sanitizing the dataset by additive or subtractive calls to the dataset; and while similar to this research, it involves an extra element of addition queries along with subtractive queries as our proposed framework uses. Their research does not differentiate between public and private features, nor do they discuss an easy extension to their work to solve this. Further, the algorithm proposed in their research is of exponential complexity; however, they do propose a method to reduce the complexity by an order of magnitude. Other work in k-anonymity has been done by Friedman et al. [5], which relates most closely with this work. Their research stretches from modeling k-anonymity in existing tables, structures and datasets as well as providing data mining algorithms for future queries. However, their research differs in that it only suggests algorithms based on the clustering and decision-tree classification approach whereas this paper address association rule mining specifically. Given some of the limitations of clustering and classifications algorithms, namely producing "best-guesses'' for the classifier, the results may not always be accurate. Friedman et al. provide a theoretical top-down approach for achieving k-anonymity using clustering without discussing an evaluation. However, Friedman et al. do provide an evaluation of their induced decision tree and classification approach using 30,162 records for training and 15,060 records for testing. They found that the algorithm proposed produces good accuracy; however, there is some error, ranging from 17-20% using the classification approach.

III CONCEPT INTRODUCTION

This section shall introduce the concepts that will be used throughout the rest of this paper related to data mining and k-anonymity.

1 ASSOCIATION RULE MINING

Association rule mining (ARM) can be defined as the process of finding patterns, associations, correlations or causal structures in a set of items or objects [13]. Some further definitions related to ARM are provided below.

Dataset: A dataset **D** is a collection of data {d^1, d^2, d^n} representing tuples defined for a schema *S*. Each tuple d^i can be represented by a set of attributes $A \subset S$.

Class Attribute: A class attribute *C* is an attribute $\in A$ which is nominally defined for every tuple in the dataset.

Association Rule: An association rule is an implication of $E \rightarrow F$ where $E \in D$ and $F \in D$ and $E \cap F = \emptyset$.

Feature Vector: A feature vector is an attribute in the database schema; for example, Postal Code and Gender may both be features vectors in a database storing demographic information.

Precedent: The precedent side of an association rule includes all feature vectors found in *E.*

Consequent: The consequent side of an association rule includes all feature vectors found in *F.*

Metric: Support: The support of an association rule is given as the percentage of transactions in D that contain both *E* and *F (*or E ∪ F). Support($E \rightarrow F$) = supportCount(E ∪ F) / countOfAllTransactions().

Metric: Confidence: The confidence of an association rule is given as the percentage of transactions in D containing E that also contain F. Confidence($E \rightarrow F$) = supportCount(E ∪ F) / supportcount(E).

Candidate: A candidate *I* is one item or set of items $\in D \mid I$ meets minimum support requirements. That is, for a given support rule of 50%, a candidate-item is a single feature or set of features found in the precedent side of an association rule whose support is greater than 50%.

Candidate-k Set: A candidate-k set is a list of k-candidate keys which hold minimum support requirements.

The metrics, support and confidence, will be used in the proposed approach section of our research to define rules that identify k-subsets of the dataset, as well as in the evaluation stage of our research to determine the fitness of the proposed framework.

2 K-ANONYMITY

Anonymization techniques generally fall into four broad categories that often overlap, these are: *generalization*, *suppression*, *perturbation,* and *permutations*. Most work regarding k-anonymization falls within the generalization and suppression categories [14]; that is, coarsely placing attributes into large set ranges, and removing attributes all together, respectively. A large body of research has focused on k-anonymity and modifications to it. Aggarwal [15] focused on the concept of perturbation in their research where they added noise to generate a probabilistic model of k-anonymity. Machanavajjhala et al. [16] proposed ℓ-diversity whereby constraints are placed on the generalization rules. Our research will focus on the definition of classic k-anonymity, with future works aimed at comparing the performance between the proposed model and other anonymization models.

K-Anonymity is a model which describes whether a dataset can be considered anonymous with respect to some value k. In general this means that for a dataset to be k-anonymous an attacker must not be able to reveal the identity of at most k-individuals, or that every individual shares each data point with another k-set of individuals. For example, for a dataset to be 10-anonymous, each attribute of each tuple must be the same as the same attribute in, at minimum, 10 other tuples. K-Anonymity is useful in situations where the user wishes to know the anonymity of a dataset with respect to a set of entries in the dataset, rather than individual entries. The rest of this paper will refer to an *anonymized* dataset as one that complies to the k-anonymous model described here. This model relies on two main assumptions [5]:

1. The database owner can identify columns, or attributes, in the dataset as quasi-identifiers and private-identifiers. Quasi-identifiers are those attributes which may have external links to public datasets which an attacker may use to gain knowledge on the dataset. Combinations of attributes such as name and address may be considered quasi-attributes as they can be found, also in combination, in a city's public directory.

2. The algorithm is fully transparent to the attacker. That is, the attacker knows which attributes are considered public and also knows how the algorithm is structured. This is similar to all well-used cryptographic algorithms.

The difficulty in structuring a k-anonymous algorithm is that the owner of the dataset must be able to identify all those attributes which have external links to public datasets [5]. In lieu of this limitation, this algorithm will assume all attributes are public unless specifically stated otherwise by the database owner. Specifically for the purposes and evaluation of the framework in this paper, no external databases will be referenced and thus only the subject database of the framework need be used.

IV PROPOSED APPROACH

1 CONCEPTUAL FRAMEWORK

The conceptual framework will be described in this section to provide a solution that iteratively, and automatically, k-anonymizes a dataset for use in research or distribution. The proposed framework is represented in Figure 1 using a sequence diagram. As shown in the figure, the conceptual framework parses a new dataset as follows:

1. Discover the database structures in the subject (research) database, and any reference public databases. This is represented as a function called aggregateDBs which takes the subject database, and any reference databases as inputs. This function returns an aggregation of the datasets given to the function.

2. Perform association rule mining, using a singular class attribute as the demographic primary key. This is done through the performARM function, taking the aggregated dataset as a parameter and returning a list of association rules, ruleSet.

3. Check if the entire dataset is anonymized, using optional user-defined domain rules to restrict filtering of attributes. This is peformed by the function resultInference which either takes a ruleSet or a ruleSet and a domain rule set, *dr*, as parameters.

   a) If the dataset does not meet all the rules for acceptable anonymization, proceed to Step 4
   b) If the dataset meets all the domain rules for acceptable anonymization, proceed to Step 5

4. Perform the filtering algorithm to trim the attributes that do not meet anonymization criteria. This step can optionally include filter rules provided by the domain expert. Similar to the result inferencing step, this step is represented by the filterSet function which takes as parameters the ruleSet or the ruleSet and the filter rules, *fr*. Once the filtered dataset is created, return to Step 2, using the new dataset.

5. The result is a k-anonymized dataset that follows the domain rules input to the framework.

From this outline, one can see that the required inputs to the framework are:

1. A subject dataset

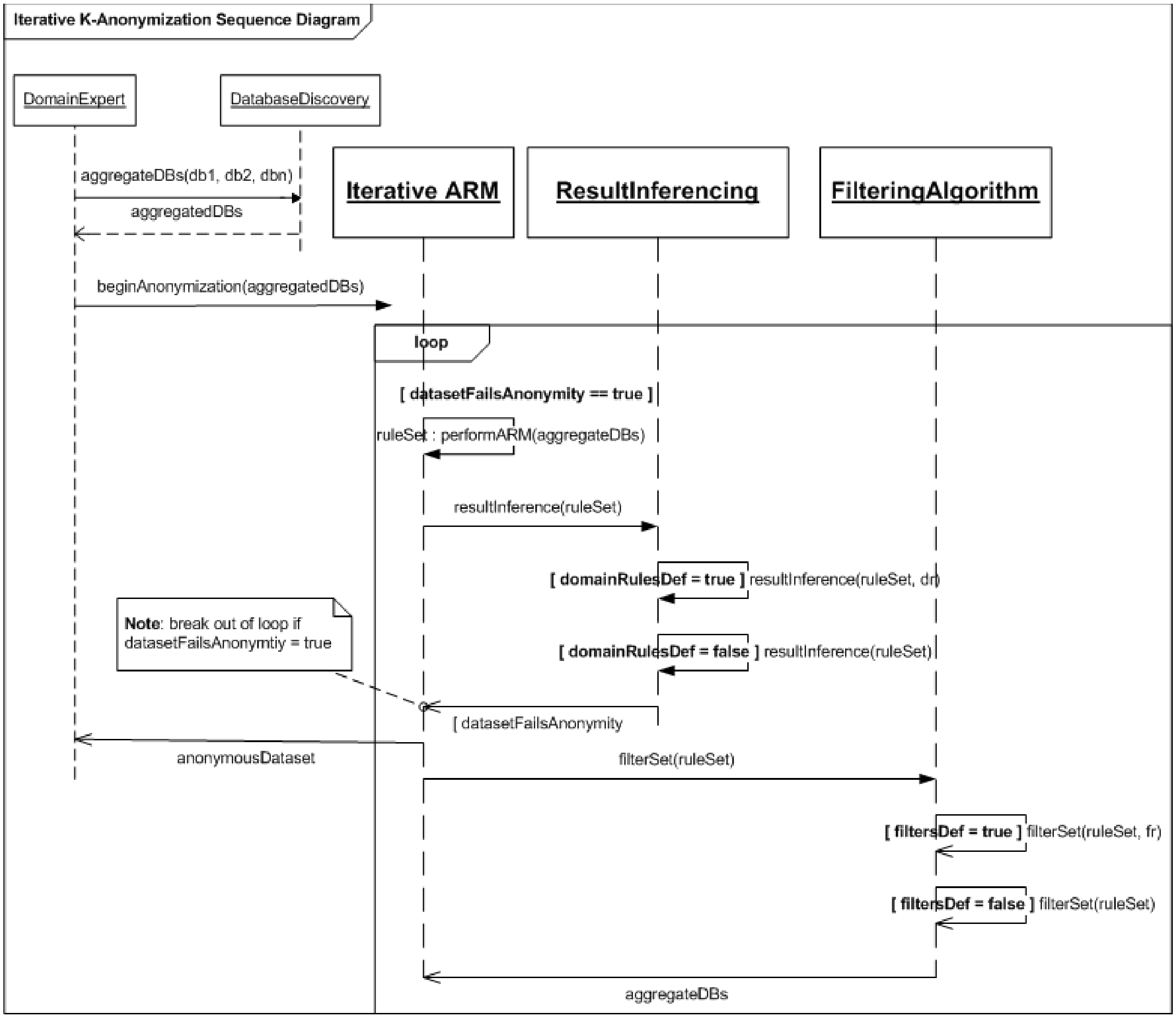
Iterative K-Anonymization Sequence Diagram
DomainExpert
DatabaseDiscovery
aggregateDBs(db1, db2, dbn)
aggregatedDBs
Iterative ARM
ResultInferencing
FilteringAlgorithm
beginAnonymization(aggregatedDBs)
loop
[datasetFailsAnonymity == true]
ruleSet : performARM(aggregateDBs)
resultInference(ruleSet)
[domainRulesDef = true] resultInference(ruleSet, dr)
[domainRulesDef = false] resultInference(ruleSet)
Note: break out of loop if datasetFailsAnonymtiy = true
[datasetFailsAnonymity
anonymousDataset
filterSet(ruleSet)
[filtersDef = true] filterSet(ruleSet, fr)
[filtersDef = false] filterSet(ruleSet)
aggregateDBs

Figure 1. Iterative K-Anonymization Framework

2. [**Optional**] Domain Rules outlining what features to ignore
3. [**Optional**] Other public datasets
4. [**Optional**] Filter rules, determining how to trim the data if it is not anonymous.

The following sub-sections will describe in more detail the association rule mining in Section IV-2, the result inferencing and filtering algorithm steps in Section IV-3 and Section IV-4 respectively.

2 ITERATIVE ARM WITH DYNAMIC CLASS ATTRIBUTE

As the proposed framework aims to provide a way to use association rule mining to anonymize a dataset for future use, it is important to understand how the association rule mining actually takes place. In classic association rule mining implementation, those relationships that have a high degree of confidence and support are used to indicate relationships that occur often and in large groups. This is adequate when trying to find features that occur frequently as well as frequently occurring relationships between them, but this approach cannot be applied to

Listing 1 Modified Association Rule Mining

```
Function AssociationRuleMiner
Inputs: Dataset data,AnonymityFactor k
Output: NonAnonFeatures[]

data = createClassAttribute(data)
results = performARM(data)
For each AssociationRule a in results
  If support(a) == (k/sizeOf(data)) Then
    If confidence(a) == 100 Then
      addFeaturesToNonAnonFeatures(a)
    End if
  End if
End For

return NonAnonFeatures[]
```

the problem of finding the number of tuples referenced using combinations of feature vectors. In other words, this framework needs to find those features which can identify a small group of tuples, rather than an existing feature vector.

Given the disparity between classic association rule mining and the goal of this research, an altered association rule mining technique will be proposed. During the association rule mining stage, an additional column will be added to the dataset whose domain is simply *1*. This will be defined as the new class attribute for the dataset. Essentially, this is used as a classical counter to determine how many individual tuples are identified in each association rule. Recall that the support of an association rule is the proportion of the transactions in the dataset for which the rule holds true. If the domain expert wishes the dataset to be k-anonymous, then the association rule mining step should indicate all rules which have a support threshold of *less than* k / sizeOf(transactions). Also recall that the confidence of an association rule is the "correctness" of the rule. That is, for the association rule $X \rightarrow Y$, what proportion of the dataset where X appears, does Y appear. Given that Y is constant throughout every tuple in the dataset, the algorithm should then only find those rules whose confidence is 100%, indicating that X always identifies k-tuples. The algorithm described here is formalized in pseudo-code in Listing 1.

Relating the algorithm to a common example in the healthcare field, should the algorithm produce the association rule *PostalCode='N5Y2N2', Gender='Male'* → *1* with a support of k / sizeOf(transactions) and with confidence of 100% then it is known that there exists k tuples in the dataset where Postal Code is N5Y2N2 *and* Gender is Male. These features are then marked as features which k-identify a subset of the data and thus must be filtered. Another important note is that the marking process will also include information regarding what candidate-k set the rule came from. For instance, those association rules that only have one feature on the precedent side of the rule will be considered as higher weighted. These higher weighted attributes will be used first in the filtering process to minimize the number of features that will lose precision to gain anonymity.

3 RESULT INFERENCING

The output of the automated association rule mining step is a list of the feature vectors which have high associations with our class attribute, along with a reference to their candidate-k set. This output will be used as the input to the result inferencing step, along with optional domain rules as a secondary input, to determine which features to exclude. Ultimately, this step finds those unique features that are found to define a subset of size k in the dataset. The algorithm will first go through all the candidate-1 features that are passed to the function. Referring to the healthcare example, if the following association rules are found to be non-anonymous from the association rule mining set, then the features *Postal Code* and *Gender* will be checked first.

1. *PostalCode='N5Y2N2', Gender='Male'* → *1*

2. *Gender='Male', Region='LONDON', Age='25'* → *1*

The features are checked against the list of domain rules provided by the domain expert, note that providing domain rules is an optional step. The domain rules are a list of exceptions the domain provider may wish to include in the algorithm. Options for the domain rules are:

Listing 2 Result Inferencing Algorithm

```
Function resultInferencer
Inputs: AssociationRules[] ar,DomainRules[] dr
Output: NonAnonAssociationRules[]

For each AssociationRule ar in
                    AssociationRules[]
  For each DomainRule dr in DomainRules[]
    If meetsRule(ar, dr) Then
     continue;
    Else
     addToNonAnonAssociationRules(ar);
    End If
  End For
End For

return NonAnonAssociationRules[]
```

Feature	Rule Type	Rule Option
PostalCode	Candidate	Candidate-1
PostalCode	Candidate	Candidate-3
Gender	Threshold Greater	7
ALL	Candidate	Candidate-4

Table 1: Domain Rule Sample

1. Ignore features if they fall within a candidate-"x" key
2. Ignore features if the number of appearances is less than a threshold
3. Ignore features if the number of appearances is greater than a threshold

These options are provided to the domain expert should they wish to have more control over the features that are found. It is not recommended to add constraints as this will mean the k-anonymous database may have lost more precision due to the algorithm requiring extra iterations to find non-excepted rules. However, this is included to ensure the framework is flexible and extendable for any domain. The function **meetsRule**(ar, dr), shown in Listing 2, takes in the two inputs: an association rule (ar) and the domain rules (dr). An example set of domain rules is shown in Table 1. The function first determines if the association rule can be ignored due to domain rule constraints. This function returns a boolean which references whether or not the association rule is anonymized as per the domain rule given. An example of a domain rule set is given in Table 3. This domain rule set shows four rules; the first indicating that all rules including Postal Code as a Candidate-1 feature should be ignored; the second indicating that all rules including Postal Code as a Candidate-3 feature should be ignored; the third rule indicates that rules including Gender must have a threshold below 7 occurrences; and the fourth rule indicating that all features that are found in a Candidate-4 set should be ignored. To further explain the third rule, if Gender is found in 8 different association rules then it occurs in more than 7 associations and thus will be ignored from the filtering step. The domain expert can use these additional constraints to retain precision for certain features within the dataset, depending on the requirements of the individual requesting the anonymous dataset.

4 FILTERING ALGORITHM

Should the output of the result inferencing algorithm produce features that identify groups of individuals of k-size, the filtering algorithm must be used to trim the features. By trimming the features, the dataset becomes more anonymous since a trimmed feature will most likely reference a larger group of individuals. The trimming process can occur in one of three ways:

1. Based on the length of the smallest value within the feature divided by 3.
2. Based on pre-determined cluster groups defined in the optional Filter Rules reference file.
3. Set the attributes to "null"

The first case will run for those feature vectors who are nominal attributes and do not have a filter rule defined in the optional reference file. The second case will run for any non-nominal attributes or nominal attributes that have a filter rule defined in the optional reference file. Finally, for all features not defined in the external reference file or features that are non-nominal, the filter algorithm will reduce all the values in the dataset to "null". The trimmed, or null value will then replace the old value in the dataset and be used in future iterations of the iterative ARM step. One important note is that the use of three in the first option for filtering is done through choice. This can be chosen by the domain expert and will help in determining the number of iterations required to anonymize the dataset.

Relating this step to the running healthcare example and using the nominal attribute Postal Code, the algorithm will run for the first trimming option given no external reference rules. If the smallest value of the nominal feature is six, the filtering algorithm will trim all postal code values by removing the last two digits from the initial value. For a postal code **N5Y2N3**, the resulting value will be **N5Y2**. This trimmed value will then replace the old value in the dataset and thus used in future iterations of the Automated Association Rule Mining step. The filtering algorithm is formalized in pseudo-code in Listing 3.

Listing 3 Filtering Algorithm

```
function filterAlgorithm
Inputs: NonAnonAssociationRules[],
        OPTIONAL: FilterRules[], Dataset
Output: FilteredDataset

For each Feature f in
                    NonAnonAssociationRules[]
  If usedFeatures.includes(f) Then
    continue;
  Else
    If FilterRules[] has values Then
      For each FilterRule fr in FilterRules[]
        If meetsRule(f, fr) Then
          FilteredDataset =
            trimWithRules(dataset, f, fr);
        Else
          continue;
        End If
      End For
    Else
      shortestValue = findShortest(dataset,f);
      FilteredDataset = trimWithSize(dataset, f,
          shortestValue/3);
      usedFeatures.add(f));
    End If
  End If
End For

return FilteredDataset
```

V CASE STUDY: MHEN STUDY

1 LAWSON

Lawson is an organization based out of London, Ontario, Canada and is the research arm of both the London Health Science Center (LHSC) and St. Joseph's Healthcare, London (St. Joseph's). Also in collaboration with the University of Western Ontario, Lawson aims to understand the basis of wellness and the dysfunctions of the body that result in disease. From this information, Lawson develops and researches ways to deliver healthcare innovations to both London and the rest of Canada. Currently, a major project underway at Lawson involves working with TELUS and TELUS's Canadianized version of Microsoft HealthVault, entitled TELUS health space^{TM}. This project is entitled the Mental Health Engagement Network (MHEN). This project is funded by Canada Health-Infoway and as a result requires rigorous privacy and security checks at milestones in the project, such as ensuring they follow the healthcare standards and operating procedures in Ontario. Finally, as this project proceeds to later iterations, the researchers at Lawson and affiliated research centers would like to perform data mining to better understand those clients enrolled in the study. This leads to a major motivation of this work: creating an extendable solution that will allow any researcher to have access to the dataset, while giving Lawson peace of mind that the dataset does not reveal information on individual clients.

To better understand the coming sections regarding Lawson, the MHEN project and the overall motivation for this work, some organization-specific definitions will be provided. These terms describe those used within the MHEN study in discussions between Lawson and TELUS to understand the context and scope of the MHEN study. The terms will also be used throughout this case study to maintain the relationship between this work and the work done at Lawson. The following list describes these terms:

Patient: A patient is a person currently under care at one of the mental health facilities in London.

Client: A client is a patient who is participating in the mental healthcare project with Lawson and interacting with commercial software.

Community Partner: A community partner is an organization in London that is involved in the mental healthcare projects with Lawson. Patients and doctors at these organizations will be clients and care providers in the research studies.

Care Provider: A certified doctor at one of the participating mental healthcare facilities in London.

2 HEALTHCARE STANDARDS IN ONTARIO

As this paper focuses on a project currently underway in London, Ontario, discussions on the current standards of healthcare in Ontario will be presented. For many healthcare projects in Canada, funding is provided by Canada Health Infoway. To maintain funding, several conditions must be continuously met and delivered throughout the project. In particular these include standards set forth by the Ontario Privacy Commission (IPC) [17] and the Office of the Privacy Commissioner of Canada (PCC) [18]. Beyond this, Canada Health-Infoway (CHI) themselves have a set of privacy standards that all projects must comply with [19]. Together, the PCC, IPC and CHI define what privacy actually means in the context of healthcare, as well as a set of several guidelines each project must meet. Particularly, Canada Health Infoway has a set of ten guidelines for any project that is related to the electronic health record in Canada [20]. These guidelines outline exactly what a healthcare project must prove to be considered a valid Canada Health Infoway project. This is important to understand for this work proposed in this paper as the evaluation of the proposed algorithm will be used within a project that is required to follow Canada Health Infoway's set of ten guidelines.

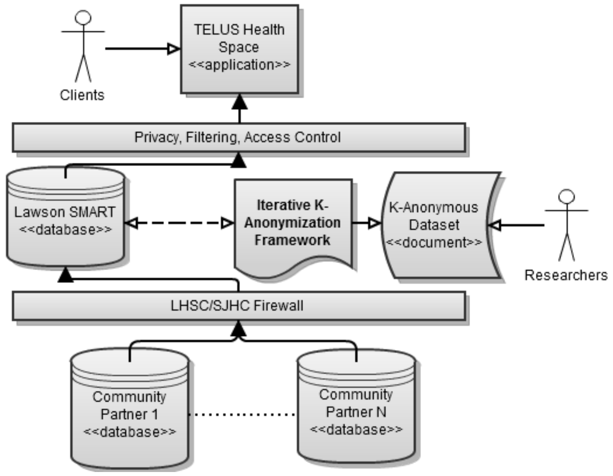
Clients
TELUS Health Space <<application>>
Privacy, Filtering, Access Control
Lawson SMART <<database>>
Iterative K-Anonymization Framework
K-Anonymous Dataset <<document>>
Researchers
LHSC/SJHC Firewall
Community Partner 1 <<database>>
Community Partner N <<database>>

Figure 2. Iterative K-Anonymization Framework Incorporated into MHEN

One of the main Acts in Canada regarding privacy and security in healthcare is the Personal Information Protection and Electronic Documents Act (PIPEDA) [3]. PIPEDA is a requirement for private-sector organizations to collect, use or disclose personal information using written consent and only for purposes that are stated and reasonable. PIPEDA defines a set of attributes considered under the Act as personal information; including, name, IP Address, age, height, weight and several others. The aim of PIPEDA is to ensure that people are comfortable in knowing that their personal information is protected and consent must be expressly given, unless otherwise specified to use or disclose their information. In addition to PIPEDA, another Act entitled the Personal Health Information Protection Act (PHIPA) [21] governs the explicit protection of data in the health sector in Canada. The equivalent Act in the United States is the Health Insurance Portability and Accountability Act of 1996 [4]. Specifically, PHIPA outlines the rules for collecting, using and disclosing personal health information and providing individuals with access to their personal health information. As defined by these purposes, organizations such as Lawson are required to follow PHIPA since they will be using personal health information in the data flow of their project. These regulations must be kept in mind when discussing the solution provided in this paper as any health organization that wishes to perform data mining, or something similar, will need to abide by these laws. Due to this, the provided solution aims at providing the end-user with a dataset, of which they can be confident that it protects the confidentiality, privacy and anonymity of its clients. The resultant k-anonymous dataset would then meet the requirements set out in PIPEDA and PHIPA.

3 THE MHEN STUDY

Mental health has become a large focus of interest in clinical and research environments because of its ubiquitous nature in the world around us. The Canadian Mental Health Association estimates that twenty percent of Canadians will personally experience a mental illness in their lifetime and most others will experience mental illness indirectly through a family member, friend, or colleague. Mental illness affects people of all ages, regardless of education, income level, or culture [22]. Mental health disorders are characterized by alterations in thinking, mood or behaviour associated with significant distress and impaired functioning [23]. Not only do these illnesses affect the individual, they also affect the healthcare system and society at large. In 2006, The Institute of Health Economics found that the economic burden of mental disorders in Canada amounted to approximately $52 billion from both direct and indirect costs and losses [24].

Growing bodies of evidence and literature suggest that prevention, early-interventions and ongoing support are the most cost-effective and sustainable ways of addressing the burden that mental illness places on the healthcare system. An innovative and sustainable method of delivering healthcare of this kind may be to utilize smart technologies to augment and support the treatment of mental health clients, while relieving the use of primary healthcare services. The Mental Health Engagement Network (MHEN) is a new study that will develop and evaluate the use of technologies in mental health to demonstrate how to more effectively and efficiently distribute healthcare services. An overview of the MHEN infrastructure with the proposed iterative ARM framework is represented in Figure 2. The MHEN project will utilize the TELUS health space^{TM} consumer health platform along with a customized personal health record application and interactive tools that support a new way for clients to receive health services, ongoing monitoring and regular communication with their healthcare team. The data for the personal health records is stored in a centralized Lawson SMART database. This is an aggregation of health records at various community partners throughout London, shown in Figure 2 as databases 1 through N. Ultimately, the MHEN study, on a wider scale, has the capability to reduce or prevent acute episodes of mental illness and reduce the severe pressures on an already over burdened healthcare system [25]. The overall hypothesis is that smart health information technology will improve the quality of life among the study participants and will reduce mental health-related costs on the healthcare system. To verify this hypothesis, a standardized evaluation framework is used to facilitate systematic research reviews on the project outcomes for economic, policy, ethical and effective analysis outcomes for the project [25].

Given the number of research possibilities created through the MHEN study, there should exist a way for all researchers to have access to the wealth of data. Provided with the proposed framework in this paper, Lawson can be confident that the k-anonymous dataset distributed to the research teams will not breach privacy or confidentiality regulations detailed in PHIPA, PIPEDA and internal organization regulations. Specifically for the MHEN project, a domain expert at Lawson will be in charge of supervising the framework. That is, they will be responsible for supplying the peripheral files to the framework, as well as pass it the database source file. The end result for the domain expert, and Lawson, is a dataset which represents an anonymized version of the original dataset. Using this framework, Lawson can be confident and comfortable in allowing external or internal researchers access to the MHEN study dataset, as well as any other dataset they manage, without breaking client confidentiality and privacy regulations.

4 SAMPLE DATASET

The dataset used in the evaluation of the framework mirrors the table real-world dataset being used in the MHEN study; however, the data is not real client data but rather test-data that Lawson

ID	Age	Gender	Region	Income	Married	FV	Cov	Current	LT	SIN
19828	22	Male	LONDON	22038	N	N	Y	N	N	Y
20211	43	Male	RURAL	879	Y	Y	Y	Y	N	N
03838	11	Male	LONDON	56752	N	N	Y	Y	N	N
19822	32	Female	RURAL	72339	N	Y	Y	Y	N	N
14722	31	Female	LONDON	4038	Y	Y	Y	N	Y	Y
49381	44	Male	LONDON	4038	Y	Y	Y	N	Y	Y
31222	72	Female	OTHER	43048	N	N	Y	Y	N	Y

Table 2: Sample Dataset

Feature Vector	Possible Values
ID	ID10001-10601, inclusive
Age	0-34, 35-54, 55+
Gender	Male, Female
Region	LONDON, RURAL, DISTANCE, OTHER
Income	0-22000, 22001-45000, 45001+
Married	YES, NO
First Visit	YES, NO
Health Coverage	YES, NO
Current Client	YES, NO
Long Term Client	YES, NO
SIN Available	YES, NO

Table 3: Feature Vector Domain

used to test the functional and non-functional requirements of the database. Shown in Table 2 is a small sample of the larger dataset that will be used to initially test the accuracy and usefulness of the proposed solution. Some feature vectors were abbreviated due to space constraints; each feature maps directly to those listed in Table 3. As seen the dataset is quite simple and limited to one table. In the dataset are mostly binary attributes such as married, first visit, health coverage and SIN available, and there are nominal attributes such as region which take on the values "London, Rural, Distance and Other". This snippet is also representative of the dataset which will be used and discussed in the evaluation section below.

5 APPROACH EVALUATION

The proposed framework has been implemented in Java using the open source Weka library [26] which includes pre-defined data mining algorithms. The data which has been used to evaluate the correctness and logic of the algorithm is based on test data provided by Lawson that mimics the structure of their production database. Table 3 shows the feature vectors used to describe the data; the name of the feature vector is listed in the left column and the domain of each attribute is listed on the right.

The dataset used consisted of 600 unique tuples, each with values for all of the feature vectors. The algorithm was run to achieve 3-anonymity for the dataset. For brevity's sake, the dataset shown is one that has undergone a pre-processing step to translate the numerical attributes into nominal attributes, these attributes were age and income, along with removing the id feature. The other input to the framework was the list of domain rules for each attribute to meet. For the first run of the implementation the domain rule file consisted of just one rule: *Postal Code -- Candidate -- Candidate1*, this follows the format represented in Table 1. This should be expanded in future work to ensure multiple rules coexist and that rules are implemented using first-come first-serve order. The last input to the framework is the list of filtering rules to be used when an attribute is discovered to be non-anonymous. This file consisted of a comma-separated list of attributes and their corresponding filter rule. Each line in the document is considered a rule, once a rule has been executed it is removed from the list of rules. This means that a single attribute may have multiple rules. For example, in the case of the feature *Region*, first execution of the filtering algorithm will trim the last character of the data stored in every tuple's *Region* attribute. On second iteration, it will trim the next character; and finally one more last character. If at this point the filtering algorithm needs to filter the attribute again, it will replace the values with "null". The filter rule file is illustrated in Listing 4. Finally, again for simplicity's sake, the algorithm does not reference other public databases to aid in finding k-anonymous patterns with public databases. For the attributes which have a domain of only two values, the trimming rule is simply to set their value to "null".

With the optional external documents defined and referenced, the framework was then executed over the Lawson dataset. The framework executed over the 600 tuples in the dataset and the results indicated that the framework iterated three times over the dataset. The framework found the following rules in the first iteration of the dataset:

1. *Age='35-54', First Visit='YES', Long Term='YES'* → *1*
2. *Age='35-54', Married='YES', Has Health='YES'* → *1*
3. *Age='35-54', First Visit='YES', Has Health='YES'* → *1*
4. *Age='35-54', Income='0-22000'* → *1*
5. *First Visit='YES', Long Term='YES', Income='0-22000'* → *1*
6. *First Visit='YES', Long Term='YES', Has Health='YES'* → *1*

Given these results, the framework then used the smallest candidate-key set, which for this iteration is a candidate-2 set, and filtered according to the given domain and filter rules. The features included in this iteration were *Age,* and *Income* as these were both found in candidate-2 sets. Since there were no domain constraints for these rules, and given the filter rules, both attributes were set to null for all tuples in the dataset. Upon second iteration of the algorithm the following association rules were found:

1. *First Visit='YES', Long Term='YES'* → *1*
2. *First Visit='YES', Long Term='YES', Has Health='YES'* → *1*

Again, given these results the framework then used the smallest candidate-key set, which is still a candidate-2 set.

Listing 4 Filter Rule - External File

```
id, -3
id, -2
id, -1
id, null
age, null
gender, null
region, -1
region, -1
region, -1
region, null
married, null
children, null
first time, null
current, null
long term, null
has health, null
has sin, null
```

The features included in this filtering process are now *First Visit* and *Long Term* as these were found in the only candidate-2 set. Since there were no domain constraints for these rules, and given that they are both binary attributes, the values of each attribute was changed to null for every tuple in the dataset. The framework then iterated over the dataset once more to find new associations rules but came up with none that matched the requirements of 100% confidence and less than *k/sizeOf(dataset)* support. Ultimately, the framework found four features that identified groups of individuals of size 3 within the database. When these features were found, the framework filtered them based on the rules discussed in Section IV-4. In the end, the framework produced a dataset, shown in Table 4, that is 3-anonymous and thus able to be distributed to anyone wishing to use the data. This was done in a timely fashion, requiring just three iterations over the dataset to produce the k-anonymous dataset.

The first performance measure of the algorithm can be illustrated using the asymptotic bounds of the solution. The worst-case algorithm proposed has a running time of O(*a*d)* where O refers to Big-O notation, *a* refers to the number of association rules found, and *d* refers to the number of domain rules. As the number of domain rules is optional, for the case of no defined domain-rules the worst-case running time is O(*a*n*) as the algorithm needs to loop through all

Age	Gender	Region	Income	Married	FV	Cov	Current	LT	SIN
null	Male	LONDON	null	N	null	Y	N	null	Y
null	Male	RURAL	null	Y	null	Y	Y	null	N
null	Male	LONDON	null	N	null	Y	Y	null	N
null	Female	RURAL	null	N	null	Y	Y	null	N
null	Female	LONDON	null	Y	null	Y	N	null	Y
null	Male	LONDON	null	Y	null	Y	N	null	Y
null	Female	OTHER	null	N	null	Y	Y	null	Y

Table 4: K-Anonymous Sample Dataset

instances in the dataset to determine the filtering rule. If we constrain the number of association rules found, as the algorithm is only interested in those less than k/sizeOf(dataset), we can say the running time is simply O(n).

The proposed solution's performance can further be evaluated with respect to the resulting dataset and the amount of information still remaining. Based on the small study, it was found that 4 of the 10 attributes needed to be addressed by the algorithm. Unfortunately, these four attributes were binary in nature and thus were coarsely filtered to "null" in all cases. The resulting dataset thus has 60% of the initial data. These results indicate that while the algorithm was successful in removing non-obvious PII, only 60% of the initial data remains for future mining purposes. The majority of this data loss is due to the inherent domain of the attributes; however, there is certainly room for further evaluation and assurance that attributes with nominal domain retain some level of information.

VI CONCLUSIONS AND FUTURE WORK

The evaluation of this framework showed promising results. Given a dataset of 600 tuples, the framework took under two seconds to anonymize the dataset using an iterative, dynamic class attribute, framework. The process took just three iterations to find associations which produced results that identified groups of 3-tuples. The framework also successfully referenced the customizable, optional external reference files to determine filter and domain rules. In the end, the framework produced a workable dataset that can be used in research or management without concern for client anonymity. Compared to some other approaches, this framework produced a dataset with 100% confidence metric for the association rules, identifying those attributes to filter. There is some inherent loss of precision due to coarsely-filtering some attributes, and finely-filtering others.

This paper has proposed a framework for creating a k-anonymous dataset using association rule mining. This process begins with discovering the structure of the dataset, through to performing classic association rule mining on the dataset and then iteratively performing k-anonymous checks and association rule mining to create a final set of data which is k-anonymous based on the primary key of the dataset. This framework is realized through two main algorithms: the result inferencing algorithm and the filtering algorithm. The result inferencing algorithm is used to compare and reference other datasets to determine the k-anonymity value of each data point which show high correlation to the primary key. Then the filtering algorithm takes those values which do not meet k-anonymity thresholds and hold high correlations to the primary key of the dataset and filters them accordingly. This framework is flexible in that it can run without any external data to "help" the algorithms run; that is, a set of domain rules, other reference public datasets or filtering rules. It can also run using all these rules instead of default-set values. By using this framework an organization can be confident in knowing their data mining procedures are anonymity-enhanced, as well as knowing that their client privacy and security are maintained throughout the entire process.

In terms of future work, given that the implementation and evaluation of the proposed architecture excluded some aspects in the initial proof of concept, it would be pertinent to include these in future iterations of the project. For example, the healthcare domain has public reference datasets that could be included to ensure that multi-dataset attacks will not be possible. Related to this, the work provided here indicated that PII is removed with respect to the subject database, it would also be important to ensure there are no correlations remaining to public datasets. Also, the paper glossed over part of the attribute-binning process which could be optimized, or even included in the filtering step, in future iterations of this work. Also, as with all proof of concept architectures, it will be important to test and evaluate this framework over larger datasets. As Lawson moves forward with the MHEN project and more data is collected this will be possible. Finally, while it is extremely important that patient privacy is preserved, it is also important for the anonymized data to remain useful. Therefore, an in-depth analysis at the performance of the proposed solution in terms of the amount of data lost should be considered.

References

[1] Public Health Agency of Canada. (2012, Aug.) Expore flu Trends in Canada. [Online]. http://www.google.org/flutrends/ca/#CA

[2] U.S. General Services Administrations. (2012, May) Personally Identifiable Information. [Online]. http://www.gsa.gov/protal/content/104256

[3] Office of the Privacy Commision of Canada. (2009, July) A Guide for Individuals: Your Guide to PIPEDA. [Online]. http://www.priv.gc.ca/information/02_05_d_08_e.asp

[4] U.S. Department of Health and Human Services. (2012, Aug.) Understanding Health Information Privacy. [Online]. http://www.hhs.gov/ocr/privacy/hipaa/administrative/index.html

[5] R. Wolff, and A. Schuster A. Friedman, "Providing Kanonymity in Data Mining," *VLDB Journal*, pp. 789-804, July 2008.

[6] and J. Natwichai B. Seisungsittisunti, "Incremental Privacy Preservation for Associative Classification," in *Proceedings of the ACM First International Workshop on Privacy and Anonymity for Very Large Databases*, New York, 2009, pp. 37-44.

[7] and P.S. Yu C.C. Aggarwal, "A Condensation Approach to Privacy Preserving Data Mining," in *EDBT*, 2004, pp. 183-199.

[8] and R. Srikant R. Aggarwal, "Privacy-Preserving Data Mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, New York, 2000, pp. 439-450.

[9] W. Du, and B. Chen Z. Huan, "Deriving Private Information From Randomized Data," in *Proceedings of the 2005 ACMSIGMOD International Conference on Management of Data*, 2005.

[10] and V. Shmatikov J. Brickell, "Efficient Anonymity-Preserving Data Collection," in *KDD*, 2006, pp. 76-85.

[11] S. Zhong, and R.N. Wright Z. Yang, "Anonymity-Preserving Data Collection," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, New York, 2005, pp. 334-343.

[12] F. Bonchi, F. Giannotti, and D. Pedreschi M. Atzori, "Anonymity Preserving Pattern Discovery," *VLDB Journal*, pp. 703-727, July 2008.

[13] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.*: Morgan Kaufmann Publishers, 2006.

[14] L. Sweeney, "K-Anonymity: A model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, October 2002.

[15] C.C. Aggarwal, "On Unifying Privacy and Uncertain Data Models," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, Washington, 2008, pp. 386-395.

[16] D. Kifer, J. Gehrke, and M. Venkitasubramaniam A. Machanavajjhala, "L-Diversity: Privacy Beyond K-Anonymity," *Transactions on Knowledge Discovery of Data*, March 2007.

[17] IPCO. (2012, Aug.) Information and Privacy Commisioner of Ontario. [Online]. http://www.ip.on.ca/english/Home-Page/

[18] Government of Canada. (2012, Aug.) Office of the Privacy Commisioner of Canada. [Online]. http://www.priv.gc.ca/index_e.asp

[19] Canada Health-Infoway. (2012, July) Privacy Reports for Canadian Health Projects. [Online]. http://www.infowayinforoute.ca/index.php/resources/reports/privacy

[20] Canada Health-Infoway. (2006, July) EHR - Privacy and Security Overview. [Online]. https://www.knowledge.infowayinforoute.ca/EHRSRA/doc/EHR-Privacy-Security-Outline.pdf

[21] ServiceOntario. (2010, Aug.) Personal Health Information Protection Act, 2004. [Online]. http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_04p03_e.htm

[22] Canadian Mental Health Association. (2012, July) Understanding Mental Illness. [Online]. http://www.cmha.ca/bins/content_page.asp?cid=3

[23] Health Canada. (2002, Aug.) A Report on Mental Illness in Canada. [Online]. http://www.hc-sc.gc.ca/pphb-dgsppsp/publicat/miicmmac/index.html

[24] P. Jacobs, and C. Dewa K. Lim, "How Much Should We Spend on Mental Health?," Institute of Mental Health, 2008.

[25] A. Rudnick, J. Hoch, M. Godin, L. Donelle, D. Neal, and D. Corring C. Forchuk, "Mental Health Engagement Network: Connecting Clients with their Health Team," in *SMART2012: First International Conference on Smart Systems, Devices and Technologies*, 2012.

[26] Machine Learning Group. (2012, Apr.) Weka 3 - Data Mining with Open Source Machine Learning Software in Java. [Online]. http://www.cs.waikato.ac.nz/ml/weka