Geographic Partitioning Techniques for the Anonymization of Health Care Data

by

William Lee Croft

Supervisors:
Prof. Jörg-R. Sack and Prof. Wei Shi

A Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Computer Science
in

Master of Computer Science
Carleton University
Ottawa, Ontario, Canada
April 1, 2015

Abstract

As the demand for the availability of detailed health care data sets continues to increase, organizations are faced with the conflicting interests of releasing this important information while protecting the confidentiality of the individuals to whom the data pertains. A major concern when releasing health care data is the geographic information which has a large influence on the re-identifiability of the data and yet is essential for many research applications. In this work, a novel system for health care data anonymization is presented. At the core of the system is the aggregation of an initial regionalization guided by the use of a Voronoi diagram. The process is broken up into major components for which different approaches are presented and tested. Testing is conducted via an implementation designed to run and analyze the results of the various combinations of approaches. In addition, a comparative test is conducted with another application, GeoLeader, which uses an alternative process for anonymization through geographic aggregation. It is shown that the Voronoi system is capable of producing comparable results with a much faster running time.

Acknowledgments

I would like to thank both of my supervising professors, Prof. Jörg Sack and Prof. Wei Shi, for their support throughout the course of this thesis. Their guidance and valuable input have been a much appreciated asset.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Domain and Motivation

In the field of health care research, relevant and detailed data sets are an essential asset. There is a high demand for health care data to be available for researchers to use. However, due to the fact that health care data is of a sensitive nature, it is necessary to protect the privacy of the respondents and patients whose information is in the data sets [2,4,13,16,17,41,55]. These data sets contain confidential information which could include details such as the medical conditions of patients or medications which they take. The protection of privacy is therefore very important in order to ensure that sensitive information is not disclosed, and by extension, to generate a level of confidence that this information will be kept safe, thus increasing respondent participation and disclosure waivers and allowing for this type of data to be made available to researchers [2, 13, 41].

Directly identifying information such as names are immediately removed from data sets, however there are ways through which the data can still be re-identified and used in potentially harmful ways [13]. An elaboration on threats associated with health care data sets and methods of protection against them is provided in the literature review. Although there has been much work in the field of anonymizing

health care data sets, concerns still remain as to whether sufficient levels of protection are achieved and how much important data is lost when anonymizing a data set [14, 15, 22, 64]. In particular, it is important to maintain as much geographic precision in the data as possible. Geographic information is becoming increasingly important for medical research as many studies involve observing the propagation of diseases across geographic areas [32]. High precision geographic information is required in order for location-critical research such as spatial epidemiology to be conducted on anonymized data sets, however patients living in small geographic areas tend to be more easily re-identifiable due to a high level of distinctness on their attributes [5, 50, 67]. Because of this privacy risk, it is difficult to disclose small geographic areas, limiting the ability of researchers to perform detailed geospatial analysis. The aggregation of small geographic regions into larger ones raises levels of anonymity; however existing approaches often result in a greater degree of information loss than necessary. For example, a technique known as cropping entails the creation of larger regions through the removal of characters from the end of postal codes, however these regions may then be larger than necessary [6, 30]. The goal of this research is therefore to propose how to anonymize health care data while maintaining a high degree of geographic precision. A modular system is proposed which is based around the anonymization of a data set through the use of geographic partitioning guided by a Voronoi diagram.

1.2 Overview of Contributions

In this work, a novel and configurable system to achieve k-anonymity [64,65] on a data set is presented. The system, Voronoi-Based Aggregation System (VBAS), achieves anonymity in a data set purely through generalization (coarsening of the level of precision) of the geographic attributes and the application of suppression of outlying records while still maintaining a high degree of geographic precision in the resultant

data set. Existing geographic-based methods of anonymization such as suppression of small regions can lead to heavily censored data sets [11, 25], whereas cropping can lead to an excessive loss of geographic precision [6, 30]. Similarly, aggregation through the use of geographic generalization hierarchies may lead to excessive loss in geographic precision and requires an a-priori construction of the hierarchy [15, 18]. Additionally, if geographic generalization is not carried out carefully, there is a risk in the loss of geographic contiguity and compactness, creating regions which are not ideal for analysis done by researchers on the resultant data [53]. Any loss in geographic precision has negative effects on the ability to effectively analyze a data set, thus the preservation of as much geographic precision as possible is important [45].

In this thesis, we focus our efforts on the generalization of geographic attributes combined with suppression in order to achieve anonymity while reducing the loss in geographic precision. We do not touch on issues such as data security, levels of trust, or risks associated with longitudinal data. VBAS addresses the problem at hand through the aggregation of small regions of fine granularity into larger regions while satisfying specific criteria employed to measure the quality of the aggregation. This criteria includes a distance measure which indicates the compactness of the new regions as well as information loss metrics used to measure the levels of suppression, geographic precision, record discernibility and non-uniform entropy.

The configurability of the system refers to the ability to select the desired quasi-identifiers [16] on which to achieve anonymity as well as the ability to choose from different approaches for each component of the system. VBAS is designed in a modular fashion to allow for easy substitution and comparison of different component approaches. This is intended not as a benefit for end-users who are more likely to want a single option to use but rather for domain experts in order to provide a framework which gives the ability to easily compare the merits of the various approaches

and their combinations. This configurability allows for additional component approaches to be easily incorporated for further testing while providing the ability to analyze their effectiveness.

The system consists of components used to:

- Approximate an appropriate number of aggregated regions
- Select locations at which to place Voronoi sites
- Construct the Voronoi diagram and perform aggregation
- Analyze the aggregation

The core of VBAS is the use of the Voronoi diagram to guide the aggregation of regions. The application of the Voronoi diagram offers a novel approach to aggregation in this context. By representing each initial region as a point in the plane determined by its centroid, the Voronoi diagram provides a simple and efficient means to group together regions which are appropriate for aggregation by grouping the regions based on the Voronoi cell in which their centroid is located. This results in an aggregation consisting of contiguous and compact regions. The Voronoi diagram, as well as the quality of the aggregation, are dependent upon appropriate output from the other components of the system.

The following approaches supplied for the other components are primarily drawn from existing work and adapted to function in this context.

1. The main approaches used to approximate the number of Voronoi sites are adapted from dynamic GAPS cutoff models [11].

2. The use of suppression (although not technically considered one of the components, as later discussed) is adapted from the typical use of suppression in k-anonymity algorithms [65].

3. Among the approaches for the selection of site locations is another contribution

of this work. A clustering algorithm which seeks to optimize levels of anonymity, ADC (Anonymity-Driven Clustering), is also presented. ADC is essentially a heavily modified k-means [8] algorithm. While following the basic framework of k-means to perform iterative optimization, ADC employs an alternative objective function for the anonymity level of an aggregation dictated by the current clusters. This change necessitates modifications for both the optimization step as well as the convergence criteria in order to maintain a monotonic change of the objective function throughout the iterative optimization.

4. Various commonly used information loss metrics [14, 15, 22, 64] are adapted and applied in the aggregation analysis component.

As an approximation is used to determine the desirable number of aggregated regions, a small range is placed around the approximation and the process is applied for different numbers of regions at certain intervals in the range. To provide the ability to focus on the more pertinent results, a solution vector is built for the chosen analysis measures and only the non-Pareto-dominated results are kept [39]. Additionally, the variables on which to base the equivalence classes used for anonymity checks can be easily selected to allow for customization of the resultant data set. This ability to configure and compare different combinations of approaches facilitates testing and provides a means for domain experts to determine which approaches are most effective in various regards.

1.3 Application Overview

An application has been developed to provide a testable implementation of VBAS. At the base of the application is the representation of the geographic regions. A region must be stored as a data structure with information about its geographic aspects,

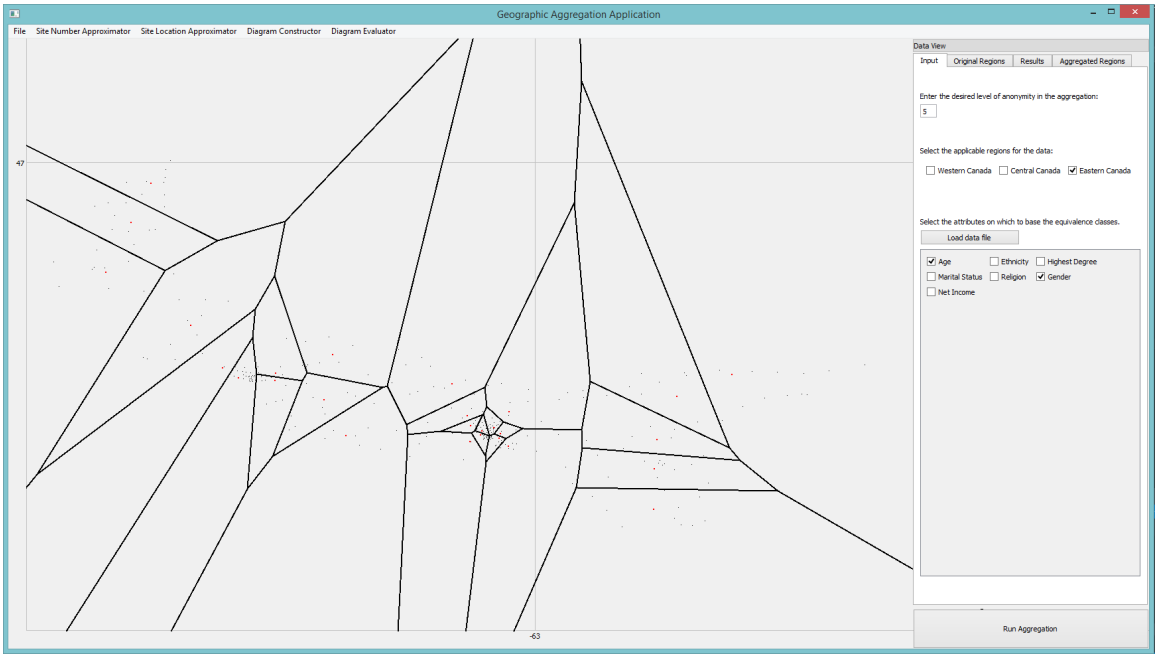
data records, equivalence classes, and anonymity. The regions are designed in such a way that it is possible to efficiently merge and query them. The efficiency of these operations is necessary for the approaches used in each component.

The four components of the system are incorporated into the application using the strategy design pattern [1] to allow for different strategies to be easily selected for each component. Via a graphical interface, the user can load in a data set and select the desired approaches for each component and the quasi-identifiers on which to base the equivalence classes. The system can then be run and the resultant aggregation is displayed for visual inspection. All non-Pareto-dominated solutions are stored and can be selected to view the aggregation and the results of the analysis.

The application minimally requires two files as input and a selection of quasi-identifiers on which to perform anonymization. The first file contains the data set which is to be anonymized. This file must contain header information about the attributes of the data set in addition to all of the records. The other required file is a region file which contains information about the initial regionalization. The region file must supply an ID for each region along with a coordinate representation of the region. The user may configure additional settings such as the desired level of anonymity and the chosen approaches or they may run the aggregation on default settings. The application output is a set of aggregated regions which each have a set of records associated with them.

[1] The strategy pattern is a design pattern described in the book *Design Patterns: Elements of Reusable Object-Oriented Software* [21]. This pattern involves the use of a client which executes an algorithm (the strategy) and returns the result. Any strategy can be supplied to the client as long as it has the same input parameters and output data type.

Figure 1: Application Screenshot



A screenshot of the application with an aggregation displayed. The input tab is open on the right hand side. Approach selection is done through the top menus.

Chapter 2

Literature Review

This chapter provides a review of the related literature for the work done in this thesis. First, a general review on anonymity with respect to the release of health care data sets is covered, followed by a review of general methods and specific approaches used to achieve anonymity. Following this, is a review on work related to the evaluation of such approaches with respect to the utility of the resultant data sets. Finally, to direct the subject matter towards the main focus of this work, a review is provided on anonymity approaches which concentrate on the geographic aspects of the data set, followed by a review of some pertinent computational geometry algorithms.

2.1 Anonymity with Health Care Data Sets

While the release of health care data sets for research purposes is a desirable occurrence, and indeed in many cases there is a high demand for this [2,4,13,16,17,41,55], it entails a high risk with regards to disclosure of confidential information relating to patients and respondents. Data custodians have the responsibility of protecting the privacy of all individuals to whom the data pertains. A breach in confidentiality may incur legal repercussions as well as negatively impact the reputation of the data custodian [2,41]. Should there be a lack of faith in the security of the data, individuals

will be less likely to allow organizations to use their information [13].

Disclosure Risks For any health care data set, whether it is being released for research purposes or being released to the public domain, there are risks involved with its release. The field of study which focuses on anonymity in health care data sets aims to develop approaches to deal with these risks. In order to delve into the approaches used to protect against these risks, it is important to first have an understanding of what types of risks exist when working with health care data sets. It should be noted that the domain of this problem pertains purely to the enforcement of privacy in the data that is released and has nothing to do with the security of the original data set. In other words, the assumption is that the data is given from the data custodian(s) to other parties to make use of. The goal is not to ensure that unwanted parties cannot obtain the data but instead to ensure that it is not possible to retrieve any confidential information from the released data which puts confidentiality at risk.

The general risk which must be guarded against is the re-identification of the data. This means that the data must be de-identified such that it can be distributed with a sufficient degree of confidence that the information in the data set cannot be used to discern any information about specific individuals [4, 13, 16, 17, 41, 55]. There are 3 main types of attributes in a data set which are relevant for this context. The first type is directly identifying variables such as names or identification numbers. These attributes must be entirely stripped from a data set during de-identification. Next, there are quasi-identifiers which are typically demographic type attributes such as age, geographic location, salary, etc. The last type is confidential attributes which contain the relevant medical information about the records [16]. De-identification is achieved by modifying the data set, typically the quasi-identifier values, in such a way that one cannot re-associate any of the resultant records with directly identifying attributes

or discern information about individuals in any other way. In other words, there should be no way to tie together confidential information with specific individuals after de-identification [4, 16, 17, 55].

De-identification of Data De-identification should be done in such a way that the resultant data set has as high of a level of utility as possible while keeping the risk of re-identification as low as possible. There are two main types of re-identification which must be protected against. The first is identity disclosure. This is when an attacker (a party attempting re-identification) is able to associate a de-identified record with the individual to whom it pertains. This compromises any confidential information stored in that particular record regarding the individual. The other type is attribute disclosure which occurs when an attacker is able to confirm that an individual exists in a particular data set or a subset of it such that this knowledge reveals some information about that individual. For example, if an attacker is able to confirm that an individual is in a data set which only contains information about people with diabetes then the attacker has confirmed that the individual has diabetes, even without having associated a particular record with them [4, 13, 16, 55].

Since there are no standardized methods for measuring the overall risk of re-identification in a data set, different strategies for analyzing the risks associated with the data set must be used instead. One very important factor is the distinctness of records in the data set. In this context, distinctness refers to how many records are indistinguishable from each other in terms of their quasi-identifier values. If many records have identical quasi-identifier values then they have a low level of distinctness whereas if there are very few records which have the same values then they have a high level of distinctness. If a record is unique, an attacker could cross-reference the record against other external data sets in order to re-identify it. Such external data sets can be either private data sets which an attacker happens to have access to or

they may simply be public data sets. Often, data sets with common demographics are made available to the public by governments and can be used in this way [4,13,16,55].

Since distinctness is based on the quasi-identifier values of records, the more demographic information which is available in the data set, the higher the risk will be as the inclusion of further details raises the levels of distinctness of the records. Similarly, finer granularity of the quasi-identifier response categories also presents a higher risk [13]. For example, if there are two similarly sized data sets, A and B, where the data set A has only respondents ages and B has both respondents ages as well as their gender, the distinctness of the records in B will be higher. Furthermore, if the data set B categorizes respondents' age by intervals of 5 years but another data set, C, also with gender and age, categorizes its respondents' age by intervals of one year, then the distinctness of the records in the data set C will be even higher.

If a record in a data set is completely unique then it is highly susceptible to re-identification. For example, if a health care data set has only one record of a male who is 78 years old and lives in a specific postal code region then an attacker may be able to look that information up in another data set and find the name of that individual. Even when records are not unique, a high level of distinctness is still undesirable as an attacker could make a guess during re-identification with a high probability of being correct if only a small number of records are the same.

2.2 Methods to Achieve Anonymity

Reducing Distinctness As most measures of disclosure risk on data sets focus on the level of distinctness in the data, the methods used to protect against this risk generally try to reduce the level of distinctness. The methods which are primarily used to do this are generalization, suppression and distortion. Generalization is typically applied globally to the data set by modifying the granularity of the response categories

Figure 2: Attribute Generalization

Age	Gender
23	Male
24	Male
37	Male
39	Male
25	Female
26	Female
33	Female
36	Female
44	Female

Generalized

Age	Gender
20-29	Male
20-29	Male
30-39	Male
30-39	Male
20-29	Female
20-29	Female
30-39	Female
30-39	Female
40-49	Female

The age attribute generalized to intervals of 10 years.

for quasi-identifier attributes [3, 15, 34, 64, 65]. For example, if one of the attributes in a data set is the age of the patient and is currently discretized by intervals of one year, this could be generalized by widening the interval to 10 years as shown in Figure 2. Such a generalization would cause more records to have the same age values, thus decreasing the level of distinctness. Suppression is typically done at a local level by completely removing records from the data set which are unique or which are causing problems in terms of obtaining the desired level of protection [3, 15, 34, 64, 65]. For example, if a data set contains only a single record which is completely unique, that record could be suppressed to ensure that the data set no longer has any unique records. Figure 3 shows a data set in which a unique record must be suppressed. Distortion involves modifying the values of the quasi-identifier attributes within some acceptable range of their original value in order to render the records more difficult to return to their original state [17, 68].

Although these methods are generally used in the fashion just described, there are

Figure 3: Record Suppression

Age	Gender
20-29	Male
20-29	Male
30-39	Male
30-39	Male
20-29	Female
20-29	Female
30-39	Female
30-39	Female
40-49	Female

The record highlighted in red is unique and must be suppressed.

also variations of how they may be applied. Suppression, which is typically used at the local level as described, can instead be used at the global level to suppress one or more attributes of the data set rather than its records. Doing so would have a much greater impact on the resultant data. Alternatively, suppression can be made even more local by suppressing a single attribute of a record [17]. Similarly, generalization can be applied at a local level in order to generalize only the values of specific records. Doing this, however, may have negative impacts on the data utility as the records will no longer have the same level of accuracy [68].

While these methods are effective at reducing the levels of distinctness in a data set, they are generally not sufficient when only a single one of them is used. As such, they are mostly used in combination [3, 15, 34, 58, 64]. Additionally, there is need for some type of framework to guide the application of the methods such that their use reduces the levels of distinctness as much as possible while detracting from the utility of the data as little as possible.

Measures of Protection The measure of protection sought to be achieved by an approach is an important factor and is one way through which these different approaches can be categorized. One of the most commonly used measures is k-anonymity [3, 15, 34, 64, 65]. K-anonymity is a measure for the level of protection against identity disclosure through guesses made at which record belongs to an individual. This is achieved by controlling the minimum size of the equivalence classes in the data set. Any records which have identical values for all of their quasi-identifier attributes are considered to be in the same equivalence class. An equivalence class of size 1 is thus a unique record. For a data set to be considered k-anonymous, each equivalence class in the data set must have a cardinality greater than or equal to k. The value of k is a user-specified variable and the higher it is set, the stronger the protection is. An example of a 2-anonymous data set is shown in Figure 4. Each

Figure 4: Equivalence Classes and k-Anonymity

Age	Gender
20-29	Male
20-29	Male
20-29	Male
30-39	Male
30-39	Male
20-29	Female
20-29	Female
30-39	Female
30-39	Female
30-39	Female

Each group of records highlighted in a different shade is an equivalence class. The anonymity of the data set is bounded by the smallest equivalence class, making it 2-anonymous.

shade represents a different equivalence class.

Similar to k-anonymity, there are other measures of protection which can be applied. These include l-diversity [33] which reduces the risk of full or partial attribute disclosure by enforcing a required level of diversity on the confidential attributes found in each equivalence class and t-closeness [58] which is essentially an extension of l-diversity which enforces that the distribution of confidential attributes in each equivalence class must be sufficiently similar to the distribution of confidential

attributes across the entire data set. An alternative class of privacy protection measures is syntactic privacy protection. This includes measures such as ϵ-differential privacy [58] which requires that a sufficient amount of distortion is applied to the data in order to render the resultant data more difficult to re-identify.

Privacy Protection Frameworks Frameworks which aim to achieve k-anonymity typically do so by following predefined generalization hierarchies in some fashion [3, 15, 34, 40, 64, 66]. A generalization hierarchy is a tree of generalization steps for an attribute of the data set where each node at any level of the tree represents the response categories for the attribute at that level of generalization. Traveling up the tree has the effect of generalizing the attribute to these predefined stages. Some approaches which use generalization hierarchies also have a maximum allowable level of suppression specified and attempt to find the best trade-off of generalization levels which can produce k-anonymity without exceeding the allowable suppression [15, 40, 64].

Other approaches aim to offer different guarantees of protection by generalizing k-anonymity [63] or extending it to more specific restrictions [33, 48, 58]. Additionally, some approaches look beyond a single data set in order to offer methods of protection against disclosure which could take place over a prolonged period of time if multiple versions of the data set are released or if an individual appears in data sets released from different sources. This is referred to as protection against trail re-identification [35, 36].

An additional consideration is privacy protection through feature selection [27, 28]. This technique, seen applied for privacy protection in data mining, is used to select which quasi-identifiers to include in the resultant data set such that a good trade-off is struck between reducing disclosure risk and maintaining accurate results.

2.3 Data Utility Metrics

While there are a number of different strategies which can be applied for the anonymization of a data set, it is important to pay attention to the side-effects induced by the anonymization; in other words, how much information has been lost in the process of anonymizing the data. As with the measurement of disclosure risk, there are no standardized measures to be applied in this area, however there have been various proposed measures.

Precision Metric One such measurement is a precision metric, first introduced in [64] and also seen applied in the system of [15]. This metric measures the precision lost in quasi-identifier attributes during generalization. In order to use the precision metric, a generalization hierarchy is needed for each attribute. The hierarchy is used to determine how much precision has been on lost an attribute by looking at how far the final form of generalization has traveled up the hierarchy. The farther up it has traveled, the greater the percentage of precision loss. The overall precision loss of a resultant data set is calculated as the average precision loss across all attributes.

Discernibility Metric Another information loss metric is discernibility, introduced in [10] and seen applied in the systems of [14, 15]. The discernibility metric assigns a penalty for each record in an anonymized data set proportional to the number of other records which are indistinguishable from it. Thus, equivalence classes which are overburdened will cause a large penalty to be applied. Similarly for each record which is suppressed, a penalty equivalent to the size of the data set is assigned as those records have become completely indistinguishable from every other record.

Non-Uniform Entropy Metric A third information loss metric is one of non-uniform entropy, introduced in [22] and seen in [15]. This metric works under the

assumption that information loss is higher on a data set with a uniform distribution of the generalized attribute than on a data set with a non-uniform distribution on that attribute. For example, consider a very simple data set which only has a gender attribute and this attribute is generalized from two options of 'male' or 'female' into a single option of 'person'. For a version of this data set with an even distribution of entries which are male and female, when picking an entry at random from the generalized data set and trying to determine the gender of the individual, there would be a 50% chance of the individual being male and a 50% chance of the individual being female. There is no bias towards either of the two options so the level of information loss is very high since there is no way to make a good guess at what the gender was originally with only this information. Now consider a version of the data set with an uneven distribution. Suppose that 90% of the entries in the original data were male and 10% were female. Picking an entry at random from the generalized data set and attempting to guess the original gender is now easier. Knowing that the distribution was uneven, there would be a 90% chance of being correct when guessing male and a 10% chance of being correct when guessing female. This bias gives a strong indication that the chosen record is very likely to be male.

The information loss measurement for non-uniform entropy is made using the probabilities of correctly guessing the original value of a record's attribute given its generalized value. The probability measurement is taken for each attribute of each record across the data set and is used to compute an information loss value. Higher values indicate a higher degree of information loss.

2.4 Geographic Partitioning to Achieve Anonymity

When anonymizing a data set, one strategy which can be used is to focus on the population sizes of the regions into which the data entries are grouped. Since a sufficient level of protection can be attained by ensuring that the level of distinctness in the data set is sufficiently low, grouping records into large regions can be used as a means to raise the level of protection. Should data records be grouped into relatively small regions, as long as the geographic attributes of the data set are released, the level of anonymity is likely to be too low due to an overly fine granularity on the geographic precision. The finely grained precision causes there to be a large number of equivalence classes with low cardinalities. Using larger regions in which to group the records is essentially a form of generalization solely on the geographic attributes of the data set. Widening the geographic regions has the effect of decreasing the number of distinct equivalence classes in the data set, thus producing remaining equivalence classes with higher cardinalities. Due to the fact that the other quasi-identifiers of the data set will still play a large role in determining the number of equivalence classes, this approach is dependent on appropriate granularities of the other attributes as well.

Cutoff Sizes Work in this area has shown that for a sufficiently large population, data will have an acceptable level of anonymity if that entire population is given the same geographic attribute [23–25]. In other words, for data sets with typical quasi-identifiers, a lack of geographic precision is sufficient to maintain anonymity; there will be enough identical records in the data set to protect privacy. Due to this fact, there are anonymization approaches which focus on reaching a safe level of privacy purely through the modification of the geographic attributes of the data set

in order to preserve all other data in its original form. There is still, however, a desire to maintain as much precision as possible in the geographic identifiers for research purposes, thus creating a trade-off.

The simplest of these approaches are the ones which make use of a single standard cutoff size. The cutoff size acts as an indicator for what minimum population size is needed in order for a region to be considered safe. Any regions in the data set which have a population under this cutoff size are suppressed. In the United States, The Bureau of Census employs a 100, 000 population size cutoff [25]. Similarly, Statistics Canada uses a 70,000 population size cutoff for their Canadian Community Health Survey [59] and the British Census uses a 120,000 population size cutoff [37]. A drawback of a standard cutoff size is its inflexibility due to the fact that the cutoff size will never change. If a cutoff size has been tailored to a specific set of attributes then it is well-suited for a particular survey but cannot be easily applied to other surveys or data sets with different variables. Carelessly applying a cutoff to another data set risks under or over-suppression. For example, despite the 100,000 cutoff used by the United States Bureau of Census, a 250,000 cutoff is used for their Survey of Income and Program Participation due to a higher perceived threat to privacy based on the contents of the data set [24].

Even for the intended data set, there are risks of under and over-suppression depending on how the cutoff size is enforced. Some approaches simply suppress all regions which do not meet the cutoff size [11]. This means that even if a region with a population size under the cutoff would have been sufficiently anonymous, it is suppressed regardless since the approach simply does not analyze local anonymity. Other approaches use cropping to raise population sizes by removing one or more digits from a postal codes identifiers to decrease geographic precision [6]. With cropping, there is a chance of a loss of more geographic precision than necessary. Regardless of how a cutoff size is enforced, there is still no guarantee that the data set is sufficiently

anonymous unless each final region has its level of anonymity checked.

Modifying Geographic Attributes An alternative approach to the problem is to mask the geographic data in such a way that its disclosure does not reveal any confidential information. One way that this can be achieved is by modifying the geographic attributes of each record in the data set in order to reduce their accuracy. This is typically done by adding some type of noise to the attribute values [1, 70]. One approach uses linear programming to reduce the probability of re-identification by shifting the geographic identifier values of the records [30]. The algorithm uses an objective function which seeks to minimize the shifts in distance and checks a neighborhood around each record to iterate over potential geographic shifts. While such approaches allow for a fine granularity to be maintained on the geographic identifiers, the level of accuracy is reduced which may present difficulties during analysis of the data [5, 31, 51].

Another approach is the modification of the geographic attributes in order to reduce their precision. The reduction of precision is generally achieved by widening the geographic areas referred to by the attributes in the data set [1, 11, 12, 70]. Assigning widened regions, however, presents two main problems: determining how to define the borders of the new regions, and determining how large the new region must be in order to provide sufficient guarantees of anonymity. The first problem can be addressed through the use of generalization hierarchies or cropping however these techniques may lead to a greater loss in precision than necessary. For the problem of determining how large the new regions should be, the use of a standard cutoff size would present similar problems to those previously discussed, thus an alternative is desirable. For this, dynamic cutoff sizes can be used [11, 12]. A dynamic cutoff size is computed for a data set by taking into account the attributes of the data set and calculating an approximation of the minimum population size required in a region for

the cardinality of the equivalence classes on the supplied attributes to be sufficiently high.

One approach employs dynamic cutoff sizes while addressing the problem of determining borders through the use of an adjacency matrix [9]. Using expanding radii, adjacent regions are merged together in order to reach the required cutoff size. Although this approach allows for greater control over the region sizes, the maintenance of the adjacency information is not practical for scenarios involving a large number of regions.

2.5 Related Computational Geometry Problems

As the main focus of the work in this thesis is based around the modification of the geographic attributes of a data set in order to achieve anonymity, a brief review is provided on some well-known problems in computational geometry which can be applied to this end.

Voronoi Diagram The Voronoi diagram [19] is an effective tool for geographic partitioning when one wishes to compute the areas nearest to a particular set of points. More specifically, given a set of points S in the two-dimensional Euclidean plane, the Voronoi diagram divides the plane into regions such that each region has an associated site from the set S and consists of all points in the plane nearer to its site than any other site in S. An important property of the Voronoi diagram is that the Voronoi regions are all convex.

Clustering Clustering algorithms also constitute a useful set of tools. In particular, k-means [8] is a simple clustering algorithm which produces clusters that have a strong tie with the regions created in a Voronoi diagram. Through an iterative optimization process, k-means continuously re-adjusts a set of k cluster centers until convergence

is reached on an objective function. The objective function aims to minimize the sum of the squared distance between each data point and the nearest cluster center. Each point is assigned to a cluster which pertains to the center it is nearest to. As such, when convergence is reached, the points in a cluster also happen to be exactly the points which would fall inside a Voronoi region for each cluster center if the centers were used as sites to construct a Voronoi diagram. Although the k-means algorithm does not compute clear-cut borders as the Voronoi diagram does, the clusters themselves can still be seen as regions of sorts.

Point Location As this work involves diagram representations of geographic regions, point location is also an important concept which is employed. In this context, this involves supplying a point as a query and determining the region in which the point is located as the result of the query. A number of different point location algorithms exist [54]. Of these, the random incremental trapezoidal decomposition algorithm [56] is used in this work as it is a part of a library, CGAL [7], used in the implementation.

Chapter 3

VBAS Details

In this chapter, an explanation is provided on the components of the system as well as how they are combined to achieve a desirable aggregation. Following this are the theoretical details for each of the approaches employed for these components.

3.1 Description of VBAS and its Components

The system developed in this work to achieve anonymity consists of five main stages as well as the application of suppression at appropriate points during the process. The five stages, in order, are:

- Loading the initial data into regional representations
- Approximating an appropriate number of aggregated regions
- Selecting locations at which to place Voronoi sites
- Constructing the Voronoi diagram and performing aggregation
- Rating the aggregation

Two pieces of input are required for the system, an initial regionalization and the data set to be anonymized. Implementation-specific details of the input and output will be addressed in Chapter 4. Table 1 provides a cross-reference of where the details for each part of the system can be found.

Table 1: System Details Cross-Reference

Topic	**Theoretical Details**	**Implementation Details**
Input and Output	n/a	Chapter 4, Section 2
Loading Data	n/a	Chapter 4, Section 3
Site Number Approximation	Chapter 3, Section 2.1	Chapter 4, Section 4
Site Location Selection	Chapter 3, Section 2.2	Chapter 4, Section 5
Aggregation of Regions	Chapter 3, Section 2.3	Chapter 4, Section 6
Application of Suppression	Chapter 3, Section 2.4	Chapter 4, Section 7
Aggregation Rating	Chapter 3, Section 2.5	Chapter 4, Section 8

It should be noted that the term anonymity is used from here onwards to refer to k-anonymity. The terms required and sufficient anonymity therefore refer to reaching a level k of anonymity. In other words, if k is set to 5, then the required level of anonymity means that each equivalence class in the data set must have a cardinality of at least 5. Similarly, the current level of anonymity refers to the smallest cardinality across all equivalence classes in the data set and the current anonymity of a region refers to the smallest cardinality of the equivalence classes in that region.

The core idea of VBAS is that a Voronoi diagram can be used as a tool to determine a geographic partitioning which achieves anonymity while satisfying other desirable objectives such as the creation of contiguous and compact regions. To use the Voronoi diagram in this way, the initial regions of the input data set are transformed into a point set representation. These initial regions must correspond to some predefined geographic partitioning of contiguous, high precision areas. This can be handled by using a geographic partitioning such as census tracts or smaller units if they are available. Since such a high level of geographic precision cannot be safely released, the regions must be aggregated together until a sufficiently safe population size is reached in all aggregated regions. To this end, a number of regions (typically

much smaller than the number of points) is approximated and a number of Voronoi sites equal to this approximated value are placed. In the resultant Voronoi diagram, all points (initial regions) which fall within the same Voronoi region are aggregated together into a single new region. Since the Voronoi regions themselves are contiguous, the aggregation of contiguous subregions inside the Voronoi regions will produce contiguous aggregated regions. Additionally, the Voronoi diagram consists of convex regions which aids in the creation of compact aggregated regions.

While the Voronoi diagram is the driving force in determining this new geographic partitioning, it is dependent upon the sites provided as input, thus the selection of the number of sites and their placement is very important. The two components which handle those jobs use the schema of the input data as well as the initial regions and the records loaded into them in order to make approximations which are well suited to the input data.

Although the main focus of the system is to achieve anonymity through the generalization of the geographic attributes, this is well complemented by an appropriate use of suppression. As such, there are various points at which suppression can be applied. In particular, there are opportunities for suppression when the data has been initially loaded, before any aggregation has occurred, and then again after the aggregated regions have been determined. The suppression used after aggregation is in fact necessary, whereas the suppression used beforehand is simply beneficial for reducing information loss.

The final stage is the rating of the aggregation which has been produced. This is important both to provide an indication of the quality of the final result as well as to have the ability to compare the various algorithms which are tested in this work.

3.2 Theory and Algorithm Details

3.2.1 Site Number Approximation

Naïve Anonymity-Based

The simplest approach taken to approximate the required number of sites uses the average expected anonymity across the initial regions to calculate how many regions must be aggregated together such that they would become sufficiently anonymous. The expected anonymity of a region is an approximation of its level of anonymity based on the total number of possible equivalence classes and the number of records in the region. The approximation is made by dividing the population of the region by the total number of possible equivalence classes as shown in Equation 1. Assuming an even distribution of equivalence classes in the region, this would give the number of members in each equivalence class. Although the distribution will not truly be even, this can still serve as an approximation; suppression is applied later to take care of any outliers. The average of this value is taken across all of the initial regions.

Let:

a be the expected anonymity

p be the region population

e be the total number of different equivalence classes

$$a = \frac{p}{e} \tag{1}$$

The aggregation of two regions creates a new region in which the cardinality of each equivalence class is the sum of the cardinalities of that equivalence class in the two original regions. For example, if one region has 3 records in the equivalence class of *Female, Age 46, Married* and another region has 2 records in the same equivalence

class, the aggregation of these regions will have 5 records in that equivalence class. Since the anonymity of a region is determined by the lowest non-zero cardinality of its equivalence classes, the aggregation of two regions can produce a region with a higher level of anonymity anonymity if the regions share members in all of their equivalence classes of the lowest cardinality. This is due to the fact that the aggregated region will have the sums of the original cardinalities, meaning that all of its equivalence classes must have a higher cardinality than those of the original regions.

If all regions had members in exactly the same equivalence classes then the aggregation of two regions is guaranteed to produce a new region with anonymity that is at least the sum of the two levels of anonymity of the initial regions. This is because their levels of anonymity represent their lowest equivalence class cardinality. If the lowest equivalence class was the same in both of the regions, then the new cardinality is the sum of the two anonymity levels. If their lowest equivalence classes differed then the anonymity level of the aggregated region is guaranteed to be higher than the sum of the two original anonymities. These facts can be applied to make a conservative estimate of the number of regions which must be merged together in order to reach the required level of anonymity. This estimate can be made as the required level of anonymity over the average expected anonymity of the initial regions, as shown in the Equation 2.

Let:

r be the expected number of initial regions required in an aggregated region

k be the required level of anonymity

e be the total number of different equivalence classes

$$r = \frac{k}{e} \tag{2}$$

The total number of initial regions would then be divided by the required number of regions which has just been computed to approximate how many aggregated regions there must be; in other words, the number of sites needed. Differing distributions of equivalence classes in the regions which are merged together may lead to an over-approximation in the desirable number of sites, resulting a greater degree of suppression at the end of the aggregation process. To account for this, the result can be multiplied by a factor which can be any real number from 0 to 1 to reduce the approximation. This final calculation is shown in the Equation 3.

Let:

s be the approximated number of sites

R be the set of initial regions

d be the distribution offset factor such that $0 < d \leq 1$

r be the expected number of initial regions required in an aggregated region

$$s = \left\lceil \frac{|R|(d)}{r} \right\rceil \tag{3}$$

Dynamic GAPS

The two dynamic Geographic Area Population Size (GAPS) models of [11] can each be used to create another site number approximation approach. These models are intended to be used to estimate an appropriate GAPS cutoff value for a given data set. The motivation behind the models stems from the previously discussed downfalls of using a standard cutoff value; namely, the issues which arise if the cutoff value is too high or too low for the data set it is applied to. The authors conduct a study on a sample of a Canadian census data set in order to produce a model which can be applied to any data set to determine an appropriate cutoff size.

The study is based on the known relationship between uniqueness of records in a region and the population size of the region. Specifically, as the population size increases, the level of uniqueness decreases. Based on this observation, it is clear that the level of uniqueness in a region can be decreased by increasing the population size of the region through the widening of the region. Another known piece of information about this relationship is also key to the study: as the population size reaches a certain point, the level of uniqueness begins to plateau.

Using these two pieces of information, the authors construct a regional model with multiple levels of nested regions. Each step up through the levels of nesting yields a small increase in the regional population. They also construct various quasi-identifier models by using different combinations of the quasi-identifiers from their data set in order to create variations on what is provided as input. They then observe the levels of uniqueness present in the various models at the different levels of nested regions and plot the data in a graph of uniqueness against regional population size. The most appropriate cutoff size is taken to be a point on the graph where the uniqueness approaches the asymptotic value of zero.

Based on the results, they provide two models which users can apply to a data set to calculate a dynamic cutoff. These models are based on the values of either entropy or max combinations of the input data set. Max combinations refers to the total number of equivalence classes in the data set which is calculated as the product of the number of response categories across each quasi-identifier as shown in Equation 1. The entropy calculation is shown in Equation 2.

Let:

Q be the set of quasi-identifiers in the data set

$|q_i|$ be the number of response categories in a quasi-identifier q_i

$$MaxCombs = \prod_{q_i \in Q} |q_i| \tag{1}$$

Let:

L be the size of the largest equivalence class

t_k be the number of equivalence classes of size k

N be the total number of records in the data set

$$Entropy = -\sum_{k=1}^{L} t_k \left(\frac{k}{N}\right) \left(\log \frac{k}{N}\right) \tag{2}$$

The two models are from these calculations are:

$$Cutoff = e^{B_0} \left(MaxCombs^{B_1}\right) \tag{3}$$

$$Cutoff = e^{B_0} \left(Entropy^{B_1}\right) \tag{4}$$

In the models, B_0 and B_1 are model parameter estimates which are determined in their study through Tobit regression.

The application of their work in VBAS is based on models that were specifically provided for three regions which make up the country of Canada. These models are shown in the Table 2.

Table 2: Regional GAPS Cutoff Models

Region	**Entropy Model**	**MaxCombs Model**
Western Canada	$1588(Entropy^{0.42})$	$1588(MaxCombs^{0.42})$
Central Canada	$1436(Entropy^{0.43})$	$1436(MaxCombs^{0.43})$
Eastern Canada	$1978(Entropy^{0.304})$	$1978(MaxCombs^{0.304})$

The entries in the table show the equations used for each of the GAPS models for the 3 regions of Canada that were studied.

3.2.2 Site Location Selection

Naïve Anonymity-Based

The simplest site location approach employed is a selection of a subset of the initial region points as sites based on the anonymity of these initial regions. When selecting s sites, the point representations of the s regions with the lowest anonymity levels are chosen. If a tie must be broken between two regions which have the same level of anonymity, the region which has more instances of equivalences classes at the tied cardinality is considered to be less anonymous. The rationale for this tie breaker is that the region with more equivalence classes at this level has a greater population which is at this level of risk and thus the region itself is more at risk. Further tie breaks are not considered beyond this and regions are arbitrarily chosen if necessary.

Naïve Density-Based

This approach works under the assumption that the initial regions are relatively similar to each other in terms of population size. This assumption is based on the use of initial regions which conform to certain standards such as census tracts or dissemination areas. As such, it is reasonable to assume that there is some approximate number of regions which could be expected to create a sufficiently anonymous region when aggregated together. This assumption follows along the same line of reasoning

that there is little benefit in creating regions with populations much greater than a cutoff size. Based on this reasoning, the approach aims to place sites such that the resultant regions will each contain roughly the same number of initial regions.

Since the aggregation of regions is guided by the computation of a Voronoi diagram on top of the point set of initial regions, a good way to achieve this would be a uniform distribution of the sites if it were the case that the initial region points were also evenly distributed. However, since the points represent regions in which people live, it is a given that the distribution will not be even. There will be dense areas of points which are more heavily populated and sparse areas of points with lower populations. Due to this uneven distribution of points, an even distribution of sites would produce aggregated regions with great differences in the number of points they contain. To account for this, the sites should be distributed based on the density of the initial region points. Having a greater number of sites in the high density areas will result in Voronoi regions of smaller geographic size but with more comparable numbers of points to the Voronoi regions in the sparsely populated areas.

The proposed naïve approach to this is to compute the smallest bounding rectangle which encases all of the initial region points and to create a grid on this rectangle. Each cell of the grid will be evaluated and assigned a number of sites to be placed within it based on its density. A greater number of grid cells allows for more precise assignments of sites in areas of high density, however if the precision is too fine, it will hamper the assignment of sites in less dense areas as large areas with numerous low density cells may be overlooked. The number of grid cells is thus restricted to being roughly equivalent to the number of sites to place in order to avoid the creation of cells which are greatly underpopulated.

Ideally, the number of grid cells would be equal to the number of sites to place. This amounts to a packing problem which s unit squares must be used to completely fill the area of an arbitrary rectangle, in this case the bounding rectangle. The grid

cells can be considered unit squares without loss in generality since the bounding rectangle can be arbitrarily scaled to satisfy this. As such packing problems have been well studied, it is known that it is not always possible to completely fill the area of the rectangle in such a problem; there may be wasted space [20, 38]. Since wasted space in this context represents geographic areas which have not been accounted for, any solution with wasted space cannot be accepted.

To accommodate this, other restrictions on the problem can be modified. The number of cells needs not be exactly matched to the number of sites. As previously discussed, it is simply necessary that the number of cells be close to the number of sites in order to produce cells of a reasonable resolution. Additionally, the cells do not necessarily need to be square-shaped. Instead, the dimensions of the cells are restricted such that one dimension may not be greater than twice the length of the other dimension. A proof on this restriction is provided along with the relevant equations below. The restriction on the ratio of the cell dimensions is enforced due to the fact that the grid cells are meant to approximate an area covered by one or more Voronoi regions. Elongated grid cells are undesirable as the Voronoi regions are more likely to cross the boundaries of multiple cells.

In the following equations, the numbers of rows and columns used in the calculations are initially kept as real numbers rather than enforcing that they must be integers as will eventually be necessary. This is done to allow a simple calculation of numbers of rows and columns whose product is the number of sites to be placed. Additionally, as will be proven, when the numbers of rows and columns are real numbers, the width to height ratio of the cells can be easily restricted to 1:1.

Without loss in generality, assume that the width of the bounding rectangle is greater than the height. Equation 1 shows that the product of the number of rows and the number of columns should be equal to the number of sites.

Let:

s be the number of sites

r' be the number of rows (as a real number)

c' be the number of columns (as a real number)

$$s = r'\,(c') \tag{1}$$

Equations 2 and 3 define the number of columns with respect to the width to height ratio of the bounding rectangle and the number of rows.

Let:

ϕ be the width to height ratio of the bounding rectangle

w be the width of the bounding rectangle

h be the height of the bounding rectangle

$$\phi = \frac{w}{h} \tag{2}$$

$$c' = r'\,(\phi) \tag{3}$$

Equations 4 and 5 define the dimensions of the cells.

Let:

w' be the width of a cell

h' be the height of a cell

$$w' = \frac{w}{c'} \tag{4}$$

$$h' = \frac{h}{r'} \tag{5}$$

Equations 6 through 10 prove that the ratio of the cell dimensions is 1:1.

Let:

ϕ' be the width to height ratio of a cell

$$\phi' = \frac{w'}{h'} \tag{6}$$

$$\phi' = \frac{w(r')}{h(c')} \tag{7}$$

$$\phi' = \frac{\phi(r')}{c'} \tag{8}$$

$$\phi' = \frac{\phi(r')}{\phi(r')} \tag{9}$$

$$\phi' = 1 \tag{10}$$

Finally, Equations 11 through 14 provide a derivation to compute the number of rows using the number of sites and the width to height ratio of the bounding box.

$$s = r'\,(c') \tag{11}$$

$$s = r'\,(r')\,(\phi) \tag{12}$$

$$\frac{s}{\phi} = r'^2 \tag{13}$$

$$r' = \sqrt{\frac{s}{\phi}} \tag{14}$$

Using Equation 14, the true numbers of rows and columns which will be used can be calculated as shown in Equations 15 and 16.

Let:

r be the number of rows (as an integer)

c be the number of columns (as an integer)

$$r = \begin{cases} 1 & \text{if } \left\lfloor \sqrt{\frac{s}{\phi}} \right\rfloor < 1 \\ \left\lfloor \sqrt{\frac{s}{\phi}} \right\rfloor & \text{if } \left\lfloor \sqrt{\frac{s}{\phi}} \right\rfloor \geq 1 \end{cases} \tag{15}$$

$$c = \begin{cases} 1 & \text{if } \left\lfloor \frac{s}{r} \right\rfloor < 1 \\ \left\lfloor \frac{s}{r} \right\rfloor & \text{if } \left\lfloor \frac{s}{r} \right\rfloor \geq 1 \end{cases} \tag{16}$$

In Equations 17 through 23 it is now shown that the width to height ratio of a cell created using the integer values for the rows and columns will never have one dimension more than twice the length of its other dimension as long as there are 2 or more rows in the grid. Still working under the assumption that the width of the bounding rectangle is greater than the height, the number of rows in the grid must be less than or equal to the number of columns. The largest impact on the ratio of the cell dimensions would occur when r' is infinitesimally smaller than some integer value such that the application of the floor function produces an r value which is nearly 1 less than r'. Additionally, in the scenario producing the largest impact, the value of c' is already an integer, meaning the that value of c is equal to c'. This simulates the largest possible change in the width to height ratio of the cells. For simplicity, the change in the value of r' is assumed to be 1 in the equations below.

$$\phi' = \frac{\frac{w}{c}}{\frac{h}{r}} \tag{17}$$

$$\phi' = \frac{w(r)}{h(c)} \tag{18}$$

$$\phi' = \frac{w(r'-1)}{h(c')} \tag{19}$$

$$\phi' = \frac{\phi(r'-1)}{c'} \tag{20}$$

$$\phi' = \frac{\phi(r'-1)}{r'(\phi)} \tag{21}$$

$$\phi' = \frac{r'-1}{r'} \tag{22}$$

$$\phi' = 1 - \frac{1}{r'} \tag{23}$$

From Equation 23, it is clear that when r' is greater than or equal to 2, the cell ratio will be greater than or equal to 0.5 and will be approach 1 as r' grows. In practice, it is expected that the number of rows will rarely be 2 or less.

The final step is to determine how many sites should be alloted to each cell based on the density of initial region points in the cells. The expected number of points in a grid cell under even distribution of the regional points is calculated as the total number of points over the number of cells as shown in Equation 24.

Let:

p be the number of initial region points

p' be the expected number of initial region points in each cell

$$p' = \frac{p}{r(c)} \tag{24}$$

Each cell then has its local density calculated as the number of points in the cell over the expected number of points as shown in Equation 25.

Let:

p_i be the number of initial region points in a cell c_i

d_i be the density of cell c_i

$$d_i = \frac{p_i}{p'} \tag{25}$$

Each site then has a number of sites assigned to it equal to the floor of its density. This is shown in Equation 26.

Let:

s_i be the number of sites assigned to a cell c_i

$$s_i = \lfloor d_i \rfloor \tag{26}$$

The remaining density of each cell is then adjusted to its local density minus the number of sites that ware assigned to the cell as shown in Equation 27. After iterating over each cell in this way, any sites remaining to be placed are assigned to cells based on the remaining densities of the cells. The remaining sites are placed one per cell, iterating over the cells in descending order of their remaining density until no more sites remain to be placed. Using this approach, each cell is given a number of sites proportional to its local density. An example of the division of cells and assignment of sites is shown in Figure 5.

Let:

d'_i be the adjusted density of cell c_i

$$d'_i = d_i - s_i \tag{27}$$

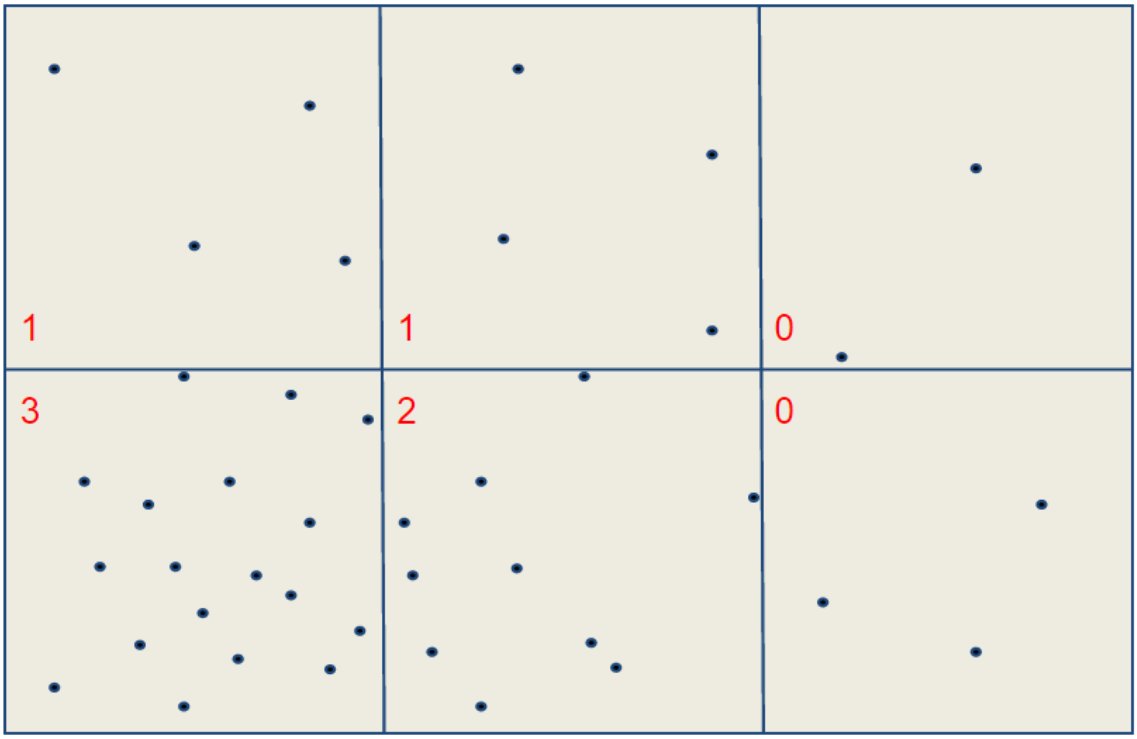
Balanced Density

The balanced density approach also creates a subdivision of cells as the naïve density approach did, however some of the restrictions now differ:

- The number of cells must be equal to the number of sites to be placed
- Each cell should have roughly an equal population
- The cells are no longer required to all have the same dimensions

By ensuring a roughly even distribution of the population across the cells, the sites can then have their distribution matched more closely to that of the global population by placing exactly one site in each cell. To achieve this, initial approximations are made for a number of rows to use and a number of cells to place in each row, however these approximations may be modified on the fly. The reason that it may be necessary

Figure 5: Naïve Density Example



The black dots represent the initial region points. With 7 sites to place and a width to height ratio of 1.6, the number of rows is calculated as 2 and the number of columns is calculated as 3. Based on these numbers, 6 cells of equal dimensions are created. The red numbers indicate how many sites are alloted to each cell.

to modify one of the approximations is that the exact population of each initial region point is not known until it is processed and added to a row or cell. If the population of the final point added to a row or cell should differ greatly from the expectation then it become necessary to make adjustments in order to maintain a distribution of population across the cells which is roughly even.

Initially, a number of rows roughly equal to the number of cells per row is approximated by taking the floor of the square root of the number of sites to place. The floor must be taken to ensure that the numbers of rows and columns are integer values, however the product of the rows and cells per row will then be short of the number of sites to place. To handle this, it is checked if the number of rows can be incremented by 1 such that the product of rows and cells per row is still less than or equal to the number of sites, and if the check succeeds then the value is adjusted to this new number. The reason this may be possible is because although the floor function was only applied once, the result was used for both the number of rows and the number of cells per row, thus there is potential for one of those two values to be incremented by 1. The calculations for the numbers of rows and cells per row are shown in Equations 1 and 2. The product of the two values may still be less than the number of sites, however this is inconsequential as the true number of cells placed in each row will be adjusted on a row by row basis.

Let:

s be the number of sites

r be the number of rows

c be the number of cells per row

$$c = \lfloor \sqrt{s} \rfloor \tag{1}$$

$$r = \begin{cases} c+1 & \text{if } c(c+1) \leq s \\ c & \text{if } c(c+1) > s \end{cases} \tag{2}$$

With this done, a division between each row must now be created. Each row should have roughly an equal population across the initial regions whose points are contained within the row. The ideal population per row is calculated as the total population divided by the number of rows as shown in Equation 3. The value is rounded to the nearest integer.

Let:

p be the total population

p' be the ideal population per row

$round(x)$ be a function which rounds x to the nearest integer value

$$p' = round(\frac{p}{r}) \tag{3}$$

The divisions between each of the rows are created by first sorting all of the region points by their y coordinates and then walking through the sorted coordinates until a number of points have been passed such that the sum of the populations of those points is greater than or equal to the ideal row population. It is not possible to guarantee that the population of the row matches the ideal population since the final point which is passed may make the row's population larger than the ideal value. Due to this, the population of the row before and after the final point was added are compared. If the population after the addition is closest to the ideal population then the point is kept in the row, otherwise, it is not kept in the row and will become part of the next row. Equation 4 shows how this decision is made.

Let:

$r_{p'}$ be the population of a row just before it passes the ideal population

$r_{p''}$ be the population of a row just after it passes the ideal population

$I(r)$ be an indicator function for a row r where its value is 1 when the final point should be included in the row and 0 when it should not

$$I(r) = \begin{cases} 1 & \text{if } r_{p''} - p' \leq p' - r_{p'} \\ 0 & \text{if } r_{p''} - p' > p' - r_{p'} \end{cases} \tag{4}$$

Once the final point of a row is decided, a division is created by assigning all of the points of the current row to a set and then continuing on to determine the points of the next row. Due to the fact that the actual population of a row may be less than or greater than the ideal population, it is possible that all points are assigned to rows before the approximated number of rows has been created or that the approximated number of rows is reached but too many points are left for the final row. In the first scenario, any rows which were not created are simply omitted without consequence. As will be seen, the number of cells per row is adjusted based on the population of the row. This means that if fewer rows were created than expected, those rows will have a greater number of cells. Despite the adjustments, the same number of cells is still created and all points are still accounted for. In the second scenario, all remaining points are assigned to the final row. As explained, each row has its number of cells recalculated based on its population. Therefore, even if the final row has a greater population than expected, it will simply account for this by having a greater number of cells as well.

Each row must be addressed individually due to the fact that they may have different numbers of cells. The number of cells to be created in a row is calculated based on the percentage of the total population which falls within the row. The total number of sites is multiplied by the decimal representation of this percentage to determine how many cells must be created in the row in order to allot an appropriate

number of sites. This is shown in Equations 5 and 6.

Let:

r_p be the population of a row

r_α be the decimal percentage of the total population in a row

r_c be the number of cells assigned to a row

$$r_\alpha = round(\frac{p}{r_p}) \tag{5}$$

$$r_c = round(s(r_c)) \tag{6}$$

The division between row cells is handled in exactly the same fashion as the division between the rows except that now the points of the row must be sorted by their x coordinates. By walking through the sorted points of the row, a division is created between each cell using the same calculation as shown in Equation 4 to determine the most appropriate population for each cell. Each time the points of a cell are determined, they are stored in a set.

As with the rows, since the populations of each cell are not guaranteed to match the ideal number, there may be a greater population than expected in the final cell or there may be multiple sites left to be assigned by the final cell. In the first scenario, the cell is simply allowed to have a greater population than the others. In the second scenario, one site is assigned to the final cell. For the remaining sites, the cells of the row are sorted by their populations. Then, in order to descending population until all remaining sites have been alloted, a cell is taken and split in half to create a new cell in which a site can be assigned.

By creating the cells in this way, a number of cells equal to the number of sites are created. While the populations of each cell will not be exactly the same, they will be similar to each other. An example of the division of cells is shown in Figure 6.

One site is placed per cell at the median of the points in that cell. Equations 7 and 8 show the computation of the median for a cell.

Let:

P be the set of points in a cell

p_i.x be the x coordinate of a point p_i

p_i.y be the y coordinate of a point p_i

$m.x$ be the x coordinate of the median

$m.$ be the y coordinate of the median

$$m.x = \frac{\sum_{p_i \in P} p_i.x}{|P|} \tag{7}$$

$$m.y = \frac{\sum_{p_i \in P} p_i.y}{|P|} \tag{8}$$

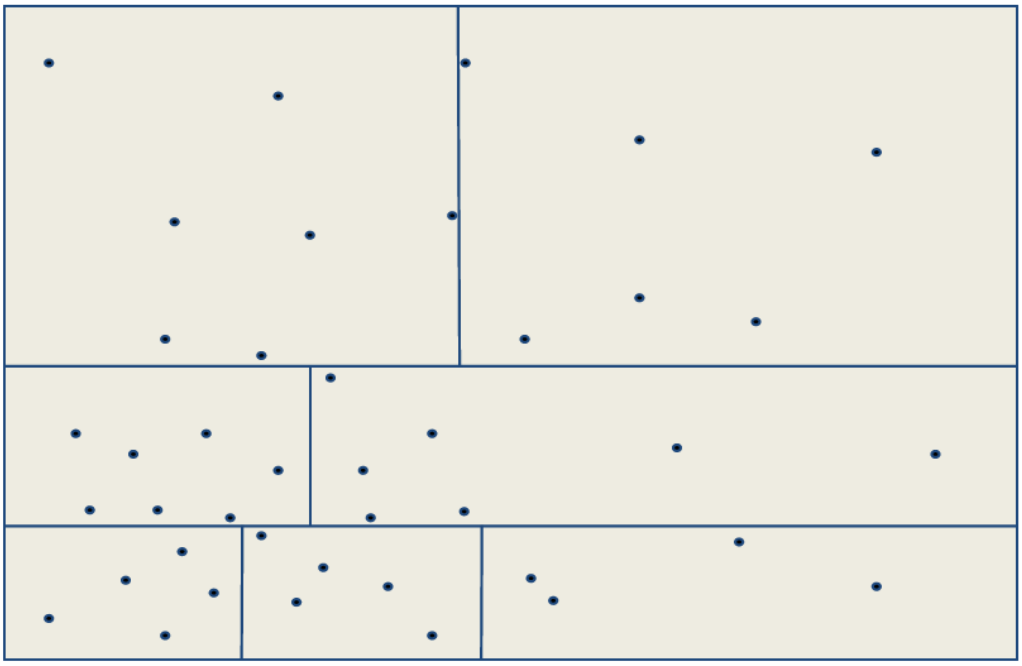
Anonymity-Driven Clustering

The Anonymity-Driven Clustering (ADC) approach is based on the k-means algorithm [8]. A set of points is taken as input and an iterative optimization process is run until convergence is reached. In order to adapt the algorithm to use anonymity as the clustering criterion, the following modifications were made:

1. A different objective function is employed which aims to reduce global anonymity.
2. The relocation of cluster centers during the optimization step has been redesigned to ensure that the move is beneficial for the new objective function.
3. The convergence criteria has been modified to accommodate these changes.

The point set of initial region points is used as the input point set for the algorithm, and the number of clusters to compute is set to the desired number of sites. The

Figure 6: Balanced Density Example



With 7 sites to place, the number of rows is calculated as 3. Note that for simplicity, each point in this example is assumed to have the same population.

cluster centers represent the sites of the Voronoi diagram and their final positions at convergence are the locations which are taken as the output of this approach. The assignment of points into clusters is the same as with k-means; each point is assigned to the cluster of the nearest center. While the initial cluster centers can be placed randomly, in order to improve the quality of the results, different seeding methods are preferable. Since all of the site location approaches share the same objective, any other site location approach can be employed as a seeding method for this clustering algorithm.

Anonymity-Based Objective Function In order to adapt the k-means algorithm, it is necessary to substitute the objective function for one which considers the desirable aspects of the aggregated regions, namely their levels of anonymity. The most important factor is the global anonymity of the current aggregation, however

it is also important to consider the number of aggregated regions which sit at the current global anonymity. While the global anonymity acts as an indicator of the overall quality of the aggregation, the number of aggregated regions sitting at the current global anonymity acts as an indicator of how far away the current state of aggregation is from raising its global anonymity. In other words, the higher the global anonymity is, the better, and the lower the number of aggregated regions sitting at the global anonymity is, the better.

The objective function that is employed is shown below in Equation 1. Higher values indicate a better aggregation. The global anonymity is multiplied by the number of aggregated regions so that an objective function value may never be dominated by one with a lower global anonymity. This is important due to the fact that just as the global anonymity is about to increase, only one aggregated region will remain at the current level of anonymity, and after the increase has occurred, all aggregated regions will sit at the new level of anonymity. A derivation to show that the objective function maintains its monotonicity in this scenario is shown in the Equations 2-4.

Let:

α be the current global anonymity

R be the set of aggregated regions

R_α be the set of aggregated regions with an anonymity of α

$$\alpha\left(|R|\right) - |R_\alpha| \tag{1}$$

$$\alpha\left(|R|\right) - 1 < \left(\alpha + 1\right)\left(|R|\right) - |R| \tag{2}$$

$$\alpha\left(|R|\right) - 1 < \alpha\left(|R|\right) + |R| - |R| \tag{3}$$

$$\alpha\left(|R|\right) - 1 < \alpha\left(|R|\right) \tag{4}$$

The aggregated regions used during the computation of the objective function value correspond to the current clusters. In other words, each cluster represents an aggregated region consisting of the initial regions points in the cluster. This representation of aggregated regions is possible due to the relationship between the k-means algorithm and the Voronoi diagram. That is, the points belonging to each cluster center in k-means are exactly the points which would fall within the Voronoi region for the corresponding site. This is due to the fact that both k-means and the Voronoi diagram compute membership based on the distance to the nearest cluster center or site respectively. The substituted objective function therefore accurately reflects the actual anonymity levels for the aggregated regions which would be formed by the selection of site locations at any time during the optimization process.

Optimization Step With a new objective function in place, an optimization operation must be determined which can only affect the objective function value in a monotonic fashion. In the k-means algorithm, each cluster center would be adjusted to the median of the points in its current cluster. Due to the fact that the median and the objective function both rely upon the same measurement (distance) this operation can ensure a monotonic decrease in the objective function value. With the modified objective function in the present algorithm, moving cluster centers to the median of the cluster no longer provides such a guarantee. To determine an alternative operation, the nature of the new objective function must be taken into consideration. From Equation 1, it is clear that reducing the number of aggregated regions at the current global anonymity will increase the objective function value. This reduction, however, must be achieved by raising the anonymity of those aggregated regions since a reduction in their anonymity would cause the global anonymity to drop, resulting in a lower objective function value.

This leads to the question of how to increase the anonymity of an aggregated

region by moving its cluster center. As discussed in the literature review, an increase in population in general leads to higher levels of anonymity, however, at this point, the number of sites is already determined and there is now a concern to disturb the surrounding regions as little as possible in order to not incur significant decreases in the anonymity of other regions as a side-effect. For this reason, it is desirable to look more closely at how to influence the anonymity of the region. By the nature of k-anonymity, the anonymity of the region is bottlenecked by one or more equivalence classes that have a cardinality matching the current global level of anonymity. An increase in the cardinality of these equivalence classes would therefore allow for an increase of regional anonymity. Of course, all such bottlenecking equivalence classes must be addressed since even a single remaining bottleneck will still limit the regional anonymity. Additionally, care must be taken to not decrease the cardinalities of other equivalence classes such that new bottlenecking equivalence classes are created.

To increase the cardinality of an equivalence class in a particular region, its neighborhood can be checked for members of the same equivalence class. The neighborhood, in this context, is the union of the Voronoi cell with the polygon defined by the sites of the adjacent Voronoi cells. The Voronoi cell is used as part of the neighborhood in order to ensure that all current points in the cell will exert an influence over where the site will shift to. The polygon of the neighboring sites is chosen in order to allow the site shift to be influenced by other nearby points outside of the cell. The polygon of the sites is used rather than the union of the each entire adjacent cell to restrict which region points can influence the movement of the site as it has the potential to be shifted to any location within the neighborhood. If the site were to move outside of this polygon, its original region would be significantly altered as the neighboring regions would claim much of its previous regional area. Such a movement is undesirable as it may significantly change the equivalence class cardinalities in the affected regions, resulting in a much more complex analysis of the

effects of the move. A special case occurs when the Voronoi cell is unbounded since this will affect the nature of the polygon formed by the sites of its adjacent cells. In such cases, the site of the Voronoi cell being improved upon is also used as a vertex in the polygon. This avoids instances of polygons being formed as a single line if there are only two adjacent cells and also widens the neighborhood allowing for a greater degree of mobility.

Should the neighborhood have members of the equivalence class in question, the region can attempt to increase its equivalence class cardinality by taking some of those members. The proposed approach to do so is to create a collection of weighted points which constitutes all regional points in the neighborhood which have members in the equivalence class, weighted by the number of members they have in that class. The new site location is computed as the weighted median of the points. This weighted median is computed in the same way as the regular median of a set of points, however each point is used in the computation a number of times equal to its weight. By adjusting the center in this way, it is hoped that it will be drawn towards an area which is more heavily populated in the equivalence class and will take some additional members from the fringes of its neighbors. The calculations of the coordinates for the weighted median are shown in Equations 5 and 6.

Let:

N be the set of weighted points in the neighborhood

w_i be the weight of a point n_i

$n_i.x$ be the x coordinate of a point n_i

$n_i.y$ be the y coordinate of a point n_i

$m.x$ be the x coordinate of the weighted median

$m.y$ be the y coordinate of the weighted median

$$m.x = \frac{\sum_{n_i \in N} n_i.x(w_i)}{\sum_{n_i \in N} w_i} \tag{5}$$

$$m.y = \frac{\sum_{n_i \in N} n_i.y(w_i)}{\sum_{n_i \in N} w_i} \tag{6}$$

Acceptance Criterion Before committing the change, the new objective function value is computed. The change is only committed if the value has increased. In this way, each optimization step that is actually committed is guaranteed to monotonically affect the objective function. The criteria of the objective function are purposely kept somewhat loose. Stricter requirements were found in experimentation to limit the ability to make necessary moves, thus producing worse results. Equations 7 and 8 show the computation of the objective function value v before the change and Equations 9 and 10 show the computation of the objective function value v' after the change. If $v' > v$, the change is committed.

Let:

- α be the global anonymity before the change
- R be the set of aggregated regions before the change
- E_i be the set of equivalence classes of an aggregated region R_i
- v be the objective function value before the change
- R_α be the set of aggregated regions with an anonymity of α before the change

$$\alpha = \min_{R_i \in R}(\min_{E_{ij} \in E_i}(|E_{ij}|)) \tag{7}$$

$$v = \alpha\,(|R|) - |R_\alpha| \tag{8}$$

Let:

α' be the global anonymity after the change

R' be the set of aggregated regions after the change

E_i be the set of equivalence classes of an aggregated region R_i

v' be the objective function value after the change

R'_α be the set of aggregated regions with an anonymity of α after the change

$$\alpha' = \min_{R_i \in R'} (\min_{E_{ij} \in E_i} (|E_{ij}|)) \tag{9}$$

$$v' = \alpha' (|R'|) - |R'_{\alpha'}| \tag{10}$$

Using the described process during the optimization step, iterative optimization runs by selecting the region of the lowest level of anonymity. If multiple regions sit at this level then one is chosen arbitrarily. Optimization is then attempted on each equivalence class of the region at its lowest cardinality. If the local anonymity of a region improves during the optimization step, the iteration through the equivalence classes of that region is ceased as it is no longer necessarily among the regions of lowest anonymity. When this happens, a new region of the lowest level of anonymity is selected (this may be the same region in some cases).

Convergence The final consideration is the stopping criteria. Optimization can cease under two conditions. The first is that the data set has reached a global level of anonymity which is sufficient. In this case, no further optimizations need to be made. If this does not happen, optimization will stop when convergence is reached. Convergence is defined in this context as the time when none of the bottlenecking regions can be improved any further. Since optimization starts back at the beginning of the list of regions every time an improvement is made, convergence is reached if optimization ever reaches the end of the list of regions.

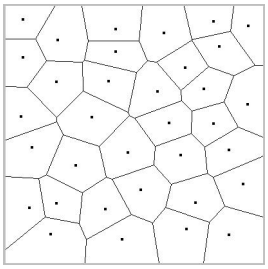
This convergence, as with the convergence of the regular k-means algorithm, represents a local minimum. It should be noted that there is potential for further improvements by making other moves such as optimizations of non-bottlenecking equivalence classes. Making such moves may improve local anonymity in other areas or create additional moves which can then lead to improvement in the bottlenecked areas. It is, however, difficult to analyze how such moves will affect the regions overall. As such, these improvements are simply not considered since their inclusion may lead to unnecessarily longer running times and present greater difficulty in defining convergence criteria.

3.2.3 Construction of Geographic Aggregation

Basic Voronoi Diagram

The construction of the Voronoi diagram is a well-known computational geometry problem. This is easily addressed by Fortune's sweep line algorithm [19]. An example of a basic Voronoi diagram is shown in Figure 7. To construct the aggregation, the sites selected from the previous component are first provided as input for the Voronoi diagram. Once the diagram is constructed, a point location structure must be set up to allow for queries to be made on the resultant diagram. This is handled using the point location algorithm of [56]. Using this point location structure, queries are made for each point representation of the initial regions. These regions are organized into disjoint sets based on the Voronoi region into which they fall, as indicated by the point location results. All initial regions which are in the same group (Voronoi region) are merged together into a single aggregated region. When this has been done for each Voronoi region, the resultant aggregated regions represent the final aggregation.

Figure 7: Voronoi Diagram



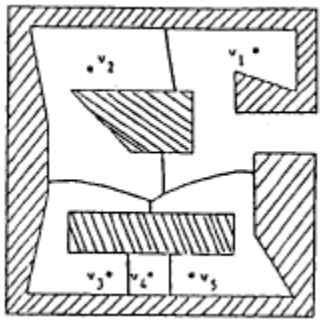
The sites of the Voronoi diagram are the black dots. It can be seen that for each site, there is a corresponding convex region. [47]

Voronoi Diagram with Obstacles

While a Voronoi diagram is useful for determining a geographic partitioning to achieve anonymity on a health care data set, it does not take into account certain geographic properties which may be of importance for research and analysis conducted using the anonymized data set. In particular, the modeling of geographic obstacles can provide useful information for researchers [42, 46]. Such obstacles may have significant effects which would be useful to observe such as an influence over the speed of the propagation of a disease [31, 52, 69]. To account for this, it is proposed to model these geographic obstacles as polygonal obstacles in the plane using a generalization of the Voronoi diagram, as presented in [26], which accounts for such obstacles in its construction.

The standard version of this algorithm computes shortest path queries in the two-dimensional Euclidean plane in the presence of polygonal obstacles. Using a quad-tree styled subdivision of free space, a continuous Dijkstra method is employed to propagate an approximate wavefront throughout the subdivision. During this wavefront propagation, all obstacle vertices which are covered act as generators and emit wavelets from their location. Interactions between the wavefronts are recorded as generator marks in a small neighborhood of cells around where collisions have occurred. After the propagation phase, the approximate wavefronts and generator marks recorded in each cell are used to compute exact wavefront collisions. This

Figure 8: Voronoi Diagram with Polygonal Obstacles



This Voronoi diagram consists of 5 sites, labeled V_1 through V_5 and polygonal obstacles represented by the shaded areas. It can be seen that some of the bisectors in this generalization are curved segments. [57]

collision information is used to construct a shortest path map in the free space of the plane. Queries can then be made using the shortest path map to find the shortest path from any point in the plane back to the source point.

A generalization of this algorithm included in the paper explains how to modify the algorithm for multiple source points. This simply amounts to initially starting a wavefront from each of the source points at the same time and using a marking system to indicate which source points first covered which obstacle vertices. If the multiple source points are taken as the sites for a Voronoi diagram then the resultant shortest path map is a geodesic Voronoi diagram of free space which is the generalization needed for modeling the effects of the geographic obstacles. A simple example of the generalized Voronoi diagram is shown in Figure 8.

3.2.4 Application of Suppression

Local Suppression

Suppression is typically used in combination with generalization to achieve anonymity on a data set while maintaining a lower degree of generalization. This is primarily due to outlying records which remain in equivalence classes of low cardinality even

after an appropriate amount of generalization has been applied [15, 64]. For this reason, suppression is beneficial when used alongside the process of aggregation as well. Once the Voronoi diagram has been computed and the aggregation has been determined, if any equivalence classes remain under the required level of anonymity in the aggregated regions, there would be a need for further aggregation. To avoid this, suppression can be applied locally within each aggregated region. The use of suppression in this way simply involves suppressing every record in every equivalence class of cardinality lower than the required level of anonymity.

Global Suppression

In addition to applying suppression after aggregation, a better result can also be obtained by applying suppression at the very beginning of the process. The approximation of the number of sites to use and the selection of their locations are based upon details from the input data set such as its size and initial equivalence classes. Therefore, any records which are certain to require suppression regardless of the amount of aggregation applied should be immediately removed from the data set. The removal of such records produces a refined input data set which allows for more accurate approximations to be made.

The records which can be immediately removed from the data set are those which will never be part of a sufficiently anonymous equivalence class regardless of how much aggregation is applied. To identify the records for which this is true, the input data set can be temporarily considered to be a single region, in other words all geographic precision has been lost. This means that equivalence classes are based entirely upon the other non-geographic attributes. It is certain that for any these equivalence classes which are not currently anonymous, any amount of geographic aggregation will not aid their case. Such records can therefore be immediately removed from the data set.

3.2.5 Aggregation Rating

General Objectives

When analyzing an aggregation that has been produced, it is necessary to have a means to rate it. There are multiple objectives which are desirable to satisfy in the resultant aggregation. Among these, attaining the required level of anonymity can be seen as a hard constraint whereas the other objectives can be seen as soft constraints. Although it is necessary to achieve the required level of anonymity, it is in fact undesirable to exceed it by any significant amount as this implies that more information has been lost than necessary. As such, one of the objectives is to minimize the average deviation from the required level of anonymity across the aggregated regions. This objective is implicitly handled through the careful approximation of the number of sites to use and the locations at which to place them however it can serve as a good measure when comparing different aggregations.

Another soft constraint is the desire to create compact regions. The larger the regions are, the less geographic precision they provide. This is an important factor since researchers such as spatial epidemiologists require fine geographic precision in the data they work with [43, 45, 53]. The use of the Voronoi diagram aids in this regard by producing regions which are convex. Other factors such as the number of sites used and their placement also have a large impact on compactness of regions. This objective can be given a measure as the sum of the distances between each initial region point and the site of the Voronoi region in which it falls. This measure will aid in the comparison of how well-suited the final regions are with respect to their shape and size.

A final general objective is the minimization of suppression. Every record which is suppressed during the process constitutes information which is lost, thus it is desirable to suppress as few records as possible. Since the amount of records suppressed during

the final stage of local suppression is dependent upon the quality of the aggregation, the final number of suppressed records may also act as a partial indicator of the quality of the solution. In other words, if the approaches leading up to this point have created a good aggregation then less suppression will be needed to achieve anonymity.

Information Loss Metrics

In order to get a more accurate measure of how much information is lost in a data set during anonymization, some previously proposed information loss metrics are used.

Precision Loss The first of these is the precision metric [15, 64] which uses a generalization hierarchy to measure how much information has been lost during the generalization of an attribute. In the process of aggregation with the Voronoi diagram, only one attribute is being generalized, however there is no hierarchy associated with its generalization. Due to this, an approximation of a hierarchy is adopted. It is assumed that the initial regions are roughly of the same size and that a step of generalization involves merging two regions sitting at the same level of generalization together into a single region. Such a hierarchy would take the form of a binary tree where all of its leaves represent the initial regions. The height of the hierarchy would therefore be $\log_2 |R|$ where R is the set of initial regions. Although in practice the regions are not merged two at a time, the level of generalization an aggregated region sits on this approximated hierarchy can be determined by $\log_2 |A_i|$ where A_i is the set of initial regions which make up an aggregated region i.

The precision loss of an attribute is typically measured as the number of steps taken up the generalization hierarchy over the total number of possible steps. The average precision loss is then taken across all of the different attributes of the data set to find the total precision loss [15, 64]. For the process of aggregation, there is only one attribute being generalized however unlike the expectation in the precision

calculation, it is generalized to a different degree for each aggregated region. As such, the calculation is modified to give total precision loss as the average precision loss across all aggregated regions. To calculate the total remaining precision, this average is subtracted from 1. The formula for the modified version is as follows:

Let:

A be the set of aggregated regions

A_i be an aggregated region from the set A

R be the set of original regions

$$1 - \frac{\left(\sum_{A_i \in A} \frac{\log_2 |A_i|}{\log_2 |R|}\right)}{|A|}$$

Thus, using this equation, a measure is given for the average loss in geographic precision that as occurred during aggregation.

Discernibility The information loss metric for discernibility [14, 15] is also employed. As this metric simply looks at the equivalence classes of records in the resultant data set there was no need for any modifications to be made. The formula assigns a penalty to all overburdened equivalence classes (ones which have deviated beyond the required level of anonymity). Aggregations in which the equivalence classes are poorly distributed across the aggregated regions will thus have a higher penalty. The higher the calculated value is, the greater the degree of information loss.

Let:

E be the set of equivalence classes

E_i be an equivalence class from the set E

k be the desired level of anonymity

$$\sum_{(|E_i| \geq k) \in E} |E_i|^2$$

Non-Uniform Entropy Finally the information loss metric for non-uniform entropy [15, 34] is applied as well. In order to make use of this metric, it is necessary to be able to measure the probability of correctly guessing the original value of an attribute given its generalized value. This calculation is shown in the following formula:

Let:

- a_r be the original value of the attribute
- b_r be the generalized value of the attribute
- n be the number of entries in the data set
- $I\,()$ be the indicator function
- R_i be original attribute value of the ith entry
- R'_i be the generalized attribute value of the ith entry

$$Pr\,(a_r|b_r) = \frac{\sum_{i=1}^{n} I\,(R_i = a_r)}{\sum_{i=1}^{n} I\,(R'_i = b_r)}$$

This probability measurement can then be used to measure the information loss across the whole data set. For the case of aggregation, there is only one attribute which must be checked per record in the data set - the geographic identifier. The following formula shows how this can be calculated (higher values indicate more information loss):

Let:

R_i be original geographic identifier of the ith entry

R'_i be the generalized geographic identifier of the ith entry

n be the number of entries in the data set

$$-\sum_{i=1}^{n} \log_2 Pr\left(R_i|R'_i\right)$$

Comparing Aggregations

Various measurements have been mentioned for rating the aggregation, as summarized in Table 3, however the task of selecting a single best aggregation out of a set of contenders based on these measurements cannot be simply done due to the fact that these measurements are not directly comparable with each other. In order to account for this, the concept of Pareto dominance [39] can be used to select aggregations which are non-dominated. With Pareto dominance, each solution is represented as a vector of objective values. A solution is said to be dominated if there exists another solution which has a value that is better or equal for each position of the vector and at least one value that is better. When comparing multiple solutions using Pareto dominance, only the non-dominated solutions are kept. This means that the Pareto optimal set may consist of multiple solutions which then require additional criteria for comparison.

For the purpose of this study, the Pareto optimal set is used to refine the original set of solutions to those which are non-dominated. Further human analysis on the remaining solutions is necessary in order to determine the merits of the different aggregations. For example, one aggregation may have more compact and regular shaped regions than the other solutions but have a higher amount of suppressed

Table 3: Aggregation Quality Measures

Measure	**Short Description**
Deviation of Average Anonymity	How far the average anonymity of the aggregated regions is beyond the required level of anonymity.
Average Distance	The average distance between the initial region points and their nearest site in the aggregation.
Suppression	The total number of suppressed records.
Precision Loss	The average percentage of loss in geographic precision across the aggregated regions.
Discernibility	The loss in discernibility between each record.
Non-Uniform Entropy	Entropy measurement for records with a non-uniform distribution of geographic attribute values.

For each measure, lower values indicate a higher quality aggregation.

records. It may only be possible to know what is preferable in such an instance based on the intended use of the final data set. Therefore, no attempt is made in this work to further reduce the set of solutions beyond those which are non-dominated.

Chapter 4

Implementation

Implementation-specific details are provided in this chapter for each of the approaches employed in VBAS. These details include time complexity analyses for each approach.

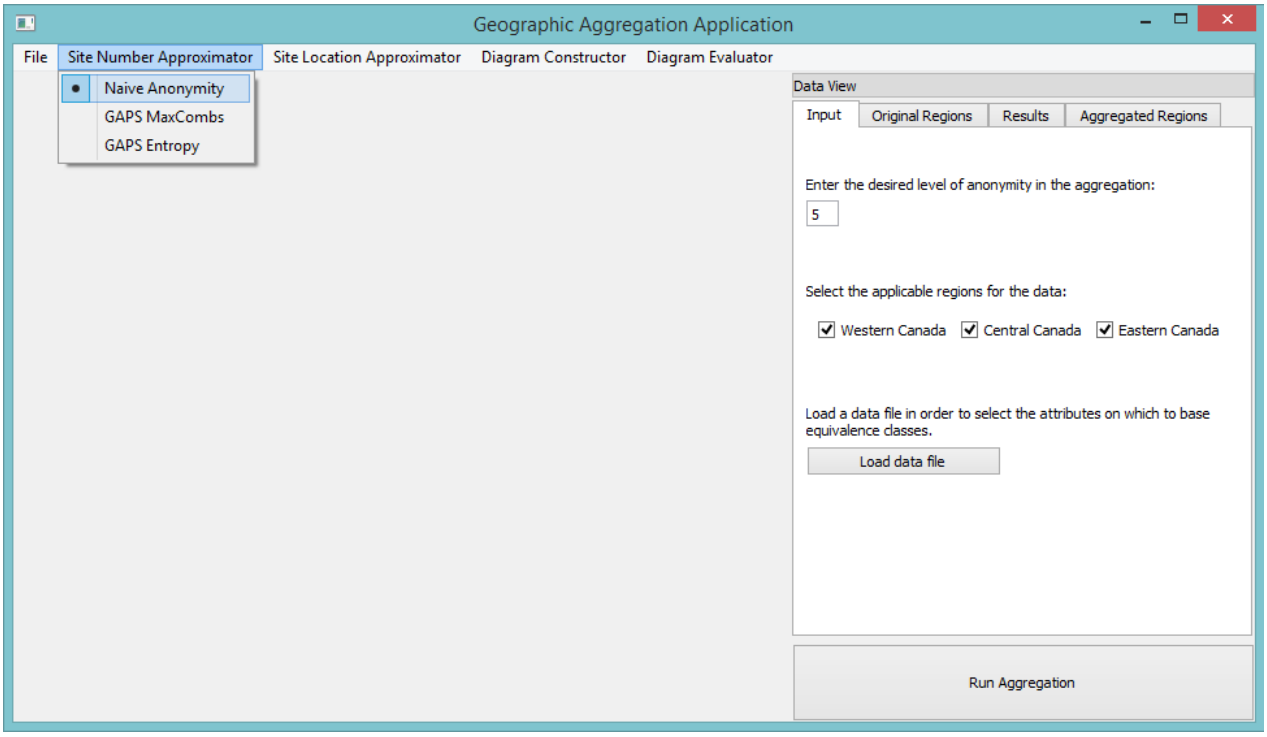
4.1 Implementation Overview

We have developed an application in order to test the proposed anonymization process and provide a tool which can be supplemented with additional approaches in order to conduct further testing and comparisons. The application includes a graphical interface to allow for easy selection of different approach combinations and a visual inspection of the results. Approach combinations as well as anonymity and quasi-identifier parameters can all be selected via the interface and applied to produce an aggregation which can then be analyzed.

The implementation has been written in C++, drawing on the Computational Geometry Algorithms Library [7] (CGAL) for the diagram representation, and Qt [49] for the user interface. The user interface is split into three main areas: the control panel, the approach selection menus, and the aggregation display area. A screenshot of the interface can be seen in Figure 1. The control panel, located on the right hand, is designed using tabs to select the different control areas. The main tab is the input

tab on from which files can be loaded and user preferences can be specified. The other tabs are used to view information about the data and the results. The four approach selection menus at the top of the window can be used to select the desired approach to apply for each of the system components. The aggregation display area, which takes up the majority of the window, displays the point representations of the initial regions as well as the Voronoi diagram used to determine the aggregation.

Figure 9: Application Screenshot



A screenshot of the application at launch. The input tab is currently selected on the control panel. Approach selection is done through the top menus. The aggregation display area is currently blank as no aggregation has yet been run.

The strategy design pattern [21] has been employed for the design of the components to allow for the approach used on each component to be easily substitutable. In this way, it is a simple matter for a user to select the desired approaches for each of the components at run time and then execute the anonymization process. The implementation details of each approach as well as other pertinent areas are discussed in the following sections.

Throughout this chapter, Table 4 should be referenced for all variables used in the time complexities.

Table 4: Time Complexity Variables

Variable	Definition
n	The number of records in the data set
d	The dimensionality of the equivalence classes (the number of quasi-identifiers selected)
r	The number of initial regions
e	The number of different equivalence classes
s	The number of sites (or aggregated regions)
i	The number of iterations of optimization (during clustering)

4.2 Input and Output

In order for the implementation to be run, two CSV (comma separated values) files must be loaded: the data set and a region file. The data set contains all of the records as well as information about the attributes of the data set. The expected format is that the first line contains the names of the attributes used in the data set. The second line contains the number of response categories for each of the attributes. The remaining lines each correspond to one record of the data set. The region file is also expected to contain the names of its attributes on the first line. This is simply needed to determine how to parse the data. Each line beneath the first corresponds to one of the initial regions. Each region in this file must have an ID and a coordinate value. Typically, longitude and latitude would be used for the coordinates but this does not need to be the case. It is important to note that the IDs of the regions must map to the geographic identifiers used for the records of the data set. If this is not initially the case then an additional mapping step is necessary.

The final representation of the data after aggregation is a set of aggregated regions which each contain a set of records and a set of initial regions. The set of initial regions represent all the regions which were merged together to create the aggregated region. The set of records are all records which are contained within the aggregated region. The anonymized data set is produced by adjusting the geographic attribute of each record to an ID representing the aggregated region within which it is grouped. Additionally, any records suppressed during the process would not be included in the resultant data set. Since the aggregated regions are dynamically created during the process, the ID of such a region alone means nothing. A user may chose to represent an aggregated region either as the set of initial regions which it covers or as the polygonal shape defined by its Voronoi region.

4.3 Representation of Regions

In order to effectively perform operations involving geographic regions, it is necessary to have an appropriate data structure to represent the regions. As the regions are dealt with in the Voronoi diagram as points in the plane, each region is given a point representation located at its centroid. Trivially, the region must also have a unique identifier and a container that holds all of the data records which pertain to the region, that is, all of the records whose geographic attributes place them within the region.

Equivalence Class Storage The most important aspect of the regional representation is how the equivalence classes in the region are handled. The maintenance of the equivalence classes is necessary in order to measure the levels of anonymity in each region. Recall that an equivalence class consists of records with the same quasi-identifier values. Since the data records are already separated by region, the inclusion

of the geographic quasi-identifiers is not necessary for these equivalence classes. The storage structure for this can essentially be thought of as a d-dimensional matrix or array where d is the number quasi-identifiers used to make up the equivalence classes. Each dimension has a size corresponding to the number of response categories for its quasi-identifier. As such, an equivalence class can be located in such a structure by ordering the quasi-identifiers and their response categories so that a sequence of d response indices derived from a data record can be used as a sequence of d indices to look up an equivalence class in a d-dimensional array.

In practice, it is desirable to design the implementation in such a way that both the number of quasi-identifiers used for an equivalence class as well as the number of response categories in each of them be dynamic. This is necessary if the application is to be applicable for more than a single data set. This, however, poses a problem for a C++ implementation using arrays as the dimensionality and and dimension sizes must be known at compile time. An alternative is to nest an STL container such as vectors to allow for variable dimension sizes. The problem of knowing the number of dimensions a-priori, however, remains.

Due to these difficulties, a single dimensional indexing approach using bit manipulation is adopted. At run time, each quasi-identifier is read from an input file along with its number of response categories. The total number of bits needed represent all response categories of each quasi-identifier is determined and that number of bits is then reserved for the quasi-identifier. For each subsequent quasi-identifier, its bits are assigned to bit positions shifted left by the number of bits already used up by previous quasi-identifiers. For example, consider a scenario in which 5 quasi-identifiers, A, B, C, D, and E, are used to determine an equivalence class. These quasi-identifiers have 2, 14, 6, 8, and 24 response categories respectively. As such, they require 2, 4, 3, 4, and 5 bits respectively to represent them. Starting with A at the least significant bit and working towards E at the most significant bit, an equivalence class can be

represented through a sequence of bits as follows:

EEEEEDDDDCCCBBBBAA

Thus, all equivalence classes can be stored in a single dimensional STL vector indexed by an unsigned 32-bit integer. In practice, this approach is limited by the number of bits available in the data type used for indexing, namely 32 bits or 64 bits depending on the architecture of the target machine. However, due to the nature of the intended application, it is likely that this will be a suitable approach as an excess of quasi-identifiers and response categories would lead to other problems during anonymization such as an inability to reach a reasonable level of anonymity across such a large number of equivalence classes.

While the bit manipulation indexing uses only a single dimension to store the information, all dimensions are still being implicitly addressed. The time complexity to compute an equivalence class index is still *O(d)*, as the response index from each quasi-identifier must be shifted and combined to form the final index. However, as mentioned, the number of quasi-identifiers used to form an equivalence class is typically expected to be fairly small. The equivalence class index needs only be computed once per record and then it can be stored in its unsigned integer form. The time complexity to initially read in the records is *O(nd)*.

Due to the fact that the equivalence class index of a data record is an alternative representation of the record with all of its information encoded, the full data records are not stored in the regions. This is done to save space as there may be millions of records processed. Once aggregation has been performed, all records of the anonymized data set can be extracted from the stored equivalence class indices using the reverse of process which was used to compute the indices.

Anonymity Storage An additional structure is stored in each region to keep track of their levels of anonymity. This structure is an array which stores at each position a

list of equivalence classes that sit at the level of anonymity which corresponds to the array index. In other words, if an equivalence class has 2 members then it will have a reference in the list at index 2 of the anonymity structure. The list simply stores the indices of the equivalence classes at its level of anonymity. This anonymity structure is not actually populated until all records have been loaded into the region. As the records are being loaded, a set is used to keep track of which equivalence classes have become populated. Once all of the records are in, the set of populated equivalence classes is processed. In constant time per equivalence class, the cardinality of the class can be looked up in the equivalence class structure and then inserted into the anonymity structure. This takes $O(n)$ time in total across all regions.

With the storage defined, the approach for carrying out the main operations on the regions remains to be addressed. The operations of note are checking the anonymity of a region and merging regions. In order to efficiently check anonymity, the anonymity structure can be used. The regional anonymity is defined by the lowest anonymity index with a non-zero value. Thus by walking up the indices of the anonymity structure, the first index which a list that has more than 0 entries defines the current level of anonymity.

Finally, the merging of regions must be addressed. This operation is used when aggregation occurs. Since it is necessary to maintain the original regions for analysis done on the aggregation, a new region is created for each aggregated region. The aggregated region then amalgamates each one of its sub-regions. During the amalgamation of a sub-region, the aggregated region first adds a reference of the sub-region to a list which it maintains. Next, each one of the records of the sub-region is added to the aggregated region. Since the records are stored in a region as their equivalence class indices, there is no need to compute the index for each record again so each addition can be done in constant time. The total time complexity across all merge operations is thus $O(n)$.

4.4 Site Number Approximation

4.4.1 Naïve Anonymity-Based

To approximate the number of sites with this approach, the average expected anonymity across the initial regions must first be calculated. Since the population of a region can be checked in constant time and the total number of possible equivalence classes is already know from loading the records into the regions, the expected anonymity of a region can be computed in constant time. The average across all regions can then be computed in *O(r)* time.

Using the average expected anonymity, the site number approximation can then be made in constant time. The ceiling function is applied to the final result since the approximation must be an integer value. The total complexity for this approach is *O(r)*.

4.4.2 Dynamic GAPS

There is little work needed for the implementation of the GAPS approach as the hard work of computing the models is already done and the models are used as is. For the MaxCombs model, the maximum combinations value is calculated in *O(d)* as the product of the number of response categories across each of the quasi-identifiers. The product is then plugged in to the model to compute the cutoff size.

The entropy model requires iterating over the equivalence class cardinalities of each initial region. This is handled using the anonymity structure stored in the regions which keeps track of the number of equivalence classes at each cardinality. As there will often be many equivalence classes sitting at the same cardinalities, this should be fairly quick in practice. In the worst case, however, each equivalence class may have a different cardinality. As such the time complexity of the entropy

calculation is *O(re)*.

For both models, if the data set is restricted to a particular region of Canada then the regional model for that particular region is used. If the data set pertains to the entire country then the largest cutoff size of the three regions is used. Once the cutoff size is computed, the number of sites is approximated as the global population across all regions over the cutoff size. As with the naïve anonymity-based approach, the ceiling function must be applied to the result in order to ensure that an integer is obtained. The additional model computations after calculating the MaxCombs or entropy value take constant time.

4.5 Site Location Selection

4.5.1 Naïve Anonymity-Based

For this approach, the initial regions are categorized based on their current level of anonymity. An array of containers is used to store the categorized regions. This array is given a size equal to the desired level of anonymity, which is typically a value from 3 to 20. Any region with an anonymity level above the required level of anonymity is placed aside in a separate container. As the anonymity can be queried in constant time using the anonymity structure, this takes *O(r)* time to categorize them. There is also technically *O(k)* time spent to set up the array of containers, where k is the desired level of anonymity, however $k << r$.

The lowest level anonymity regions are picked as sites until the required number of sites is met. To do this, while the lowest level anonymity category has a number of regions less than or equal to the remaining number of required regions, all regions from that category are taken as sites. When a point is reached at which the lowest level anonymity category has more regions than the remaining number of required sites,

the regions in that category are further categorized. The sub categories correspond to the cardinality of the equivalence level for a region, in other words, the regions are sorted by how many equivalence classes they have which sit at this particular level of anonymity. This sub categorization is done using the multimap of the C++ STL which is a map that allows for duplicates of keys and handles insertions in logarithmic time. Once the map is constructed, the regions are in sorted order by their cardinalities and the regions from the back of the map with the highest cardinality are taken until the required number of sites is reached. If a scenario occurs in which all categorized regions have been selected and more sites are needed, the remaining sites are arbitrarily chosen from the higher anonymity regions which were placed in the extra container. Using this approach, the time complexity is *O(r log r)*.

4.5.2 Naïve Density-Based

The naïve density-based approach is implemented using the CGAL arrangement structure to create a grid over top of the regional points. First, a bounding rectangle is computed by iterating over each region to find the minimum and maximum x and y coordinates. This takes *O(r)* time. Using the calculations outlined in Chapter 3, the numbers of rows and columns needed for the grid, as well as their height and width, respectively, can be computed in constant time.

Based on the numbers of rows and columns computed in combination with the dimensions of the bounding rectangle, a CGAL arrangement is constructed to represent the grid. The insertion time of an x-monotone segment in a CGAL arrangement takes logarithmic time proportional to the complexity of the arrangement to locate an endpoint of the segment to insert and then linear time proportional to the number of other segments intersected. Based on the number of grid cells being created which is linear in s, the number of sites, the complexity of the arrangement and the total number of segment intersections will be linear in s. The total time complexity for

construction is thus *O(s log s)*. At this point, the number of initial region points in each grid cell must be determined. This is done by iterating over each initial region and using CGAL's RIC point location structure to find which cell the regional point falls in. The point location is done in logarithmic time so the complexity for this step is *O(r log s)*.

Next, sites are assigned to each cell based on the density of the cells. The expected number of regional points in each cell under an even distribution is calculated as the total number of points over the number of cells. The cells are then placed into an STL multiset which is given a custom comparator to sort the cells in descending order by the number of points they contain. As each cell is added to the multiset, it is also assigned a density equal to the floor of its number of points over the expected number of points. An iterative process is run where each cell in the multiset starting from the first cell up to the last one which has the same density is assigned a single site. Each time a cell is assigned a site, its density is decreased by 1. When the last cell of the same density is reached, the process starts once more from the beginning of the multiset, now assigning sites to the cells of the next highest density as the values have all been decreased by 1. This continues until a number of sites equal to the required number of sites have been assigned. In this way, cells with higher density receive more sites and priority is always given to the cells with the most points. Placing the cells into the multiset takes *O(s log s)* time and assigning the sites takes *O(s)* time.

Finally, once all of the sites have been assigned, the remaining work is to position the sites within the cells they have been assigned to. In order to avoid placing the sites close to each other, they are placed with a roughly even distribution inside the cells by using the same technique of subdividing a rectangle into a rectangle into a grid with a certain number of cells, this time setting the cell as the bounding rectangle and the using the number of sites to be placed within the cell as the number of sub-grid cells to create. In order to account for the fact that this time the exact number

of of cells is needed, some rows or columns are supplemented with an additional cell. Each site is then placed at the center of one of the sub-grid cells. This final step takes *O(s)* time. The total time complexity for this approach is bounded by the *O(r log s)* during the point location step.

4.5.3 Balanced Density

In the balanced density approach, the initial approximations for the number of rows and the cells per row can be computed in constant time by taking the floor of the square root of the number of sites and preforming a simple check to see if the number of rows can be increased by 1. Since the creation of the rows requires the initial regions points to be sorted, *O(r log r)* time must be spent to sort them by their y coordinates. This can be done using an STL multimap in C++ with the y coordinates as the keys.

The ideal population per row can be computed in constant time. The points assigned to each row are determined by walking across the sorted points to determine where one row ends and the next begins. As each time a row is determined, it is assigned the points which were walked across in the map. The total time spent during this stage is *O(r)* to walk across all of the points in the map.

Next, each row can be addressed. The computation for the number of cells which are alloted to the row can be done in constant time. The points which were assigned to the row are inserted into another multimap, this time using the x coordinates as the key to re-sort the points by their x coordinates. In total, another *O(r log r)* time is needed in the worst case to sort all the points in each row. Another *O(r)* time is needed in total to walk across the points in each row and determine the divisions between the cells. Since the cells must be sorted by their population, this requires another *O(s log s)* time across all rows as there are s cells in total.

Finally, the median of the points in each cell must be computed in order to place

the sites. This step takes *O(r)* time to compute medians across all r points. The total time complexity of the approach is bounded by the *O(r log r)* spent sorting the initial region points.

4.5.4 Anonymity-Driven Clustering

The first step of the clustering process is to place the initial cluster centers. The random seeding is handled by simply using a random number generator to select coordinates within the minimum and maximum x and y coordinates of the initial region points. Determining the bounds takes *O(r)* time and randomly placing the sites takes *O(s)* time. If one of the other site location approaches is used instead for seeding then the time complexity for this step is that of the chosen approach. Since the site location approaches are already set up to be used with the strategy pattern, it is a simple matter to execute a chosen strategy in order to seed the sites. The random site selection seeding is in fact set up as an additional approach as well in order to apply it in the same way.

Next the regional points are assigned to their clusters by creating a CGAL Voronoi diagram for the sites and then using CGAL's point location to determine which Voronoi regions the points fall in. Details on this are deferred to the next section on the construction of the aggregation. This step takes *O((r log s) + n)* time.

At this point, the iterative optimization can begin. The aggregated region with the lowest level of anonymity is found in *O(s)* time. Although a structure such as a min heap could be employed to select the regions of minimum anonymity more efficiently over multiple iterations, that route is not to taken due to the fact that it is possible for a linear number of regions to be updated at each step, making such an approach less efficient. Once the minimum region is chosen, it is necessary to identify its neighborhood. This information can be extracted from CGAL's Voronoi structure. When each aggregated region is created, it is stored in a map along with

a handle to the Voronoi face which it represents. The point representations of the Voronoi sites are used as the keys for the map. By using the face handle of the minimum region, the neighboring Voronoi faces can be found in constant time and the corresponding aggregated region can be looked up in *O(log s)* time per neighbor. By [44], the expected number of neighbors of a Voronoi region is constant. This step therefore has an expected time complexity of *O(log s)*. A CGAL polygon is then formed in constant time from the sites of the adjacent regions and another polygon is made from the vertices of the cell which is being improved. These operations can all be run by CGAL in constant time since there is a constant number of neighbors.

Next, the optimization step is performed for the minimum region. Using the anonymity structure in the minimum region, its equivalence classes of the lowest anonymity are found in constant time. Each of these classes is processed one at a time. When an equivalence class is checked, the neighboring regions must be checked for sub-regions with members in the same class. Since equivalence class lookups can be done in constant time using the equivalence class structure in each region, this takes *O(r)* to go through each neighboring sub-region. Once the neighboring points of the same equivalence class have been found, the subset of these which fall within the neighborhood polygons must be determined. Using CGAL, these polygons can be queried in constant time since they have constant complexity. The time to determine which points fall within the polygon thus takes *O(r)* time. Finally, the weighted median of the neighborhood points which share members in the equivalence class is calculated in *O(r)* time.

Once a new site location has been determined, a temporary aggregation is constructed. As with the first construction of the aggregation, this takes *O((r log s) + n)* time. Each new region must then be checked to ensure that global anonymity has not decreased and to count the number of equivalence classes at the lowest level of anonymity. Using the anonymity structure in each region, this takes *O(s)* time in

total. If these conditions are not met then the aggregation is rejected.

When a new aggregation is accepted, a new minimum region is found and the optimization continues. Optimization will continue until the aggregation is sufficiently anonymous or a state of convergence has been reached. Convergence occurs when the end of the list of regions is reached during an iteration of optimization since this means that no improvements were made. Since the number of iterations of optimization cannot be determined before running the algorithm, the total time complexity is *O(i((r log s) + n))*.

4.6 Construction of Geographic Aggregation

Currently the only approach which is implemented for the diagram construction component is the basic Voronoi diagram. The diagram is constructed using the CGAL Voronoi diagram structure in *O(s log s)* time. Once the Voronoi diagram has been constructed. The aggregated regions are created by using the CGAL RIC point location on each of the points representing an initial region. The regions are grouped together based on the Voronoi region in which they are located. This step takes *O(r log s)* time. With the regions grouped together, they are then merged into the final aggregated regions by creating a new region for each Voronoi region, using the Voronoi site as the region point, and merging each of the grouped regions into their corresponding aggregated region. As explained earlier, the time complexity of a merge is linear in the number of records in the regions being merged. Since all of the regions across the data set are merged at this step, the time complexity is *O(n)*. The total time complexity is thus *O((r log s) + n)*.

4.7 Application of Suppression

4.7.1 Local Suppression

The implementation of suppression at the local level is fairly simple. Each record in the region is iterated over. If the record is part of an equivalence class of cardinality lower than the required level of anonymity, it is suppressed. The records store their equivalence class index so the cardinality of a record's equivalence class can be checked in constant time. Since this must be done for each record in each region, the total time complexity is *O(n)*.

4.7.2 Global Suppression

The implementation of suppression at the global level is achieved in a fashion quite similar to suppression at the local level. At the beginning of the process, a dummy global region is created and all data records are loaded into it. This takes *O(n)* time. With all of the records in a single region, the application of the local suppression described above on this single region achieves the effect of global suppression. Any records which are not part of sufficiently anonymous equivalence classes when all geographic precision has been removed will be suppressed. Once the global suppression has been performed, the remaining records are inserted to their proper initial regions. Suppression of the global region and redistribution of the records both take *O(n)* time.

4.8 Aggregation Rating

4.8.1 General Objectives

The general objectives, which relate to measurements of anonymity, distance and suppression, can all be measured by making a single traversal across the final aggregated regions to query their data structures. Global anonymity is measured as the lowest level of anonymity across the regions, where each region can be queried for its anonymity in constant time using the anonymity structure. Average anonymity is measured as the average across these regional levels of anonymity and the average deviation of anonymity is measure as the average across all regions of the deviation of their anonymity from the required level of anonymity. The anonymity calculations take *O(s)* time to go over each aggregated region.

The distance measurement requires iterating over each sub-region of each aggregated region and calculating the distance between the point representation of the sub-region and the Voronoi site of its aggregated region. The distance measurement is taken as the average across all such distances. Although the sum of the distances would be sufficient for comparing this measurement between different aggregations on the same set of regions, the average is necessary in order to make comparisons between aggregations which use a different set of regions. The distance measurement takes *O(r)* time since each initial region must be checked.

The total amount of suppression is measured as the original size of the data set minus the sum of the populations of the final regions. This measurement accounts for both global and local suppression. Finally, the average suppression is simply taken as the global suppression over the number of final regions. The suppression measurements require *O(s)* time to go over each aggregated region. The total time complexity for all general objective measures is thus bounded by *O(r)*.

4.8.2 Information Loss Metrics

The precision metric can be computed with a traversal over the aggregated regions. The precision loss of an aggregated region is measured as the log base 2 of the number of its sub-regions over the log base 2 of the total number of initial regions. Each final region stores its sub-regions in a list and the number of initial regions is already known, so the calculation can be made in constant time per aggregated region. This takes *O(s)* time.

The discernibility metric requires looking at each equivalence class in each aggregated region. In order to avoid a need to look at each one individually, the anonymity structure of the aggregated region can be used. Since the calculation assigns the same penalty for each equivalence class of the same cardinality, all equivalence classes of the same cardinality can be dealt with at once, thus requiring only a single traversal of the anonymity structure, which is, in a practice, a constant size. In this way, discernibility can also be computed in *O(s)* time across all aggregated regions.

Non-uniform entropy requires calculating a value for each original attribute value being checked. Since the original attribute values in this context refer to the initial regions, it is necessary to make a calculation for each initial region. The initial regions are stored inside the aggregated region into which they have been amalgamated. This means that each sub-region of each aggregating region must be traversed, giving a total time complexity of *O(r)*.

4.8.3 Comparing Aggregations

In order to account for the fact that the aggregation solutions produced by this process are based on approximations, a small range is placed around the number of sites which are approximated. An aggregation is computed for each number of sites within this range. For each solution, a vector is constructed based on the values from the diagram

analysis. Each time a new solution is computed, its solution vector is compared to the vectors of each other solution that has been stored so far. When comparing two solution vectors, each value is compared in order to check for Pareto dominance. If there is no stored solution which dominates the current solution then it is stored, otherwise it is rejected. While the use of such a range means that the location of the Voronoi sites, construction of the diagram, and analysis of the diagram will have their time complexity multiplied by a factor proportional to the range placed around the site number number approximation, this is not included in the time complexity analysis as this is not recommended as part of the process to be applied in practice. This is simply used as a means to provide additional insight into the effectiveness of the approximations.

4.9 Complexity Summary

A summary is provided in Table 5 of the time complexities from each of the component approaches described in the previous sections. All variables used here correspond to Table 4.

Table 5: Approach Time Complexities

Component	**Approach**	**Complexity**
Site Number Approximation	Naïve Anonymity-Based	*O(r)*
	GAPS MaxCombs	*O(d)*
	GAPS Entropy	*O(rc)*
Site Location Approximation	Naïve Anonymity-Based	*O(r log r)*
	Naïve Density-Based	*O(r log s)*
	Balanced Density	*O(r log r)*
	Anonymity-Driven Clustering	*O(i((r log s) + n))*
Diagram Construction	Basic Voronoi	*O((r log s) + n)*
Suppression	Local	*O(n)*
	Global	*O(n)*
Aggregation Rating	General Objectives	*O(r)*
	Information Loss	*O(r)*

Chapter 5

Testing and Discussion

Data sets for different regions of Canada were generated and used to test the implementation and compare its performance with GeoLeader[1] [9], an implementation of another aggregation system. The tests and comparisons involve the measurement and discussion of levels of suppression, the compactness of aggregated regions, information loss metrics and running times. Details are also provided on the generation of the test data that is used. All tests were run a machine using 16 GB of RAM and a 4.01 GHz processor.

5.1 Test Data

5.1.1 Test Data Generation

Data Sources The data sets used for testing were generated from the Statistics Canada public use microdata file for the 2011 National Household Survey [60] (NHS) in combination with Statistics Canada's data set of Canadian dissemination areas [61]. As required by Statistics Canada's data use regulations, it is stated that the results or views expressed here are not those of Statistics Canada.

[1]It should be noted that there are multiple different systems with this name. The system in question can be found here.

The NHS data set contains respondent level data for 2.7% sample of the Canadian population with 133 variables. The wide range of variables in this data set provide an excellent source of demographic information on which to create equivalence classes during testing. Due to the fact that the release of so many variables may compromise respondent confidentiality, the geographic precision in this data set is limited to the level of provinces. Since a much finer degree of geographic precision is needed in order to test the aggregation process, it was necessary to generate semi-dummy data sets based on the NHS data.

The dissemination areas defined by Statistics Canada are ideal geographic regions to use as the initial regions for the aggregation process. A dissemination area is a small geographic unit which is made up of census blocks (even smaller units) with a population targeted to be between 400 and 700 [61]. The dissemination areas therefore satisfy the necessity for finely grained geographic precision and initial regions with comparable populations. Statistics Canada provides a dissemination area data set which lists every dissemination area in Canada along with other variables. Notably, these variables contain the longitude and latitude coordinates of each dissemination area as well as the province in which it exists. Since each of the initial regions must be given a point representation, the longitude and latitude coordinates can serve as this representation.

Synthetic Data Set Generation Given that the dissemination areas are used as the initial regions, it is necessary for the input data set to have a geographic attribute which corresponds to this level of geographic precision. To achieve this, the two data sets are used to create a semi-dummy input data set. Based on the NHS and dissemination area data sets, a new testing data set is generated for the entire population of Canada using a selection of quasi-identifiers commonly found in health care data sets, and geographic precision at the dissemination area level. The chosen

quasi-identifiers are age, gender, marital status, highest level of education, ethnicity, religion, and income. All of these variables can be found in the NHS data set.

The approach used to create the testing data set first processes the NHS data set to make an approximation of the distribution of the chosen variables in each of the provinces. This is done at the level of provinces since this is the finest level of geographic precision available in the NHS data set. The approximations are simply based on the number of records in a province in a particular response category over the total population of the province. For example, if 460,000 records have a gender value of female in a province with 1,000,000 records, then the province is calculated to have 46% of its records as female. Once these values have been computed for each response category of each variable in each province, they are applied to the dissemination areas. Each dissemination area is processed and is randomly assigned a population between 400 to 700, as per the Statistics Canada documentation. A record is created for each member of the population in the dissemination area. For each variable of the record, the value is chosen based on the probabilities recorded for the province in which the dissemination area exists. For example, if a record is being added for a dissemination area in the province used in the previous example, there is a 46% chance the record will be given a gender of female and 54% chance it will be male.

When this generation of test data is complete, the resultant data set has a number of records roughly equivalent to the population of Canada. The number is not exact due to the random selection of dissemination area populations. It should also be noted that the distribution of the response categories in the final data set is not necessarily representative of the true distributions of these values across Canada. The generated data is based off of a 2.7% sample of the Canadian population and thus cannot be used to make such assumptions. For the purpose of the testing that is conducted on the data set, this is inconsequential. The goal is simply to verify the effectiveness of

VBAS on data which has a realistic distribution of quasi-identifier values.

Regional Data Sets In order to provide a selection of different data sets with different sizes, the testing data that has been generated for all of Canada is used to create 3 subsets, one for Western Canada, one for Central Canada and one for Eastern Canada. Eastern Canada is comprised of all records from provinces east of Quebec, Central Canada is comprised of all records from Ontario and Quebec, and Western Canada is comprised of all records from the provinces west of Ontario and includes the territories. This follows the regional models of [11]. Note that these data sets are not generated anew, they are proper subsets of the original testing data set. The point representations of the dissemination areas in each of the 3 regions can be seen in Figures 10, 11, and 12.

Figure 10: Eastern Canada Dissemination Areas

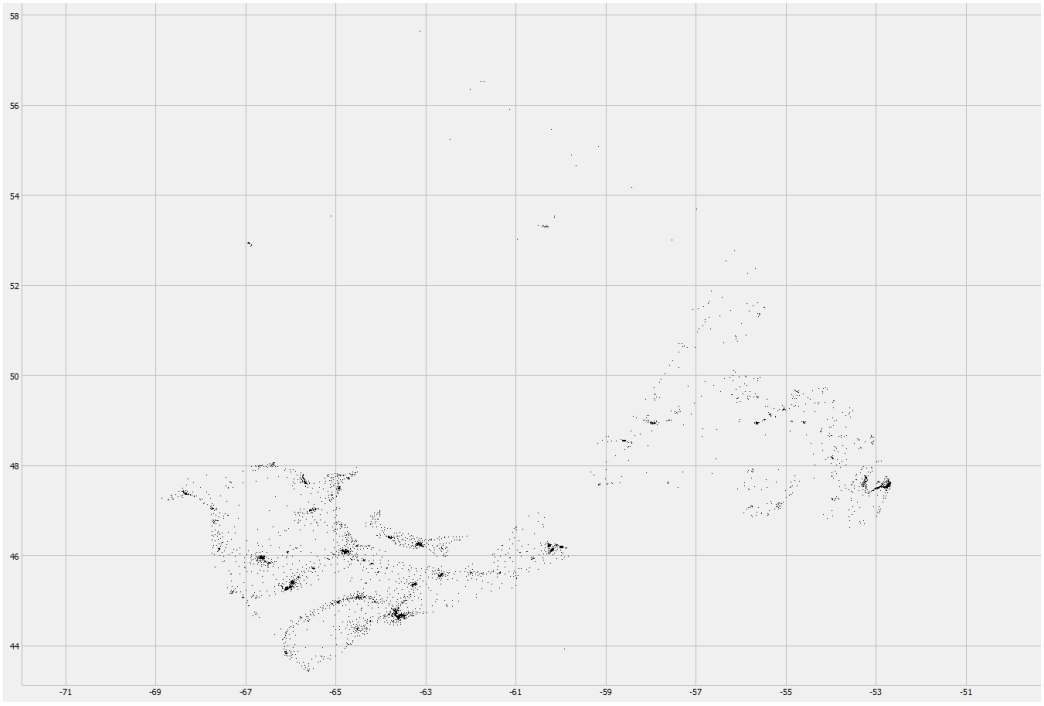


Figure 11: Central Canada Dissemination Areas

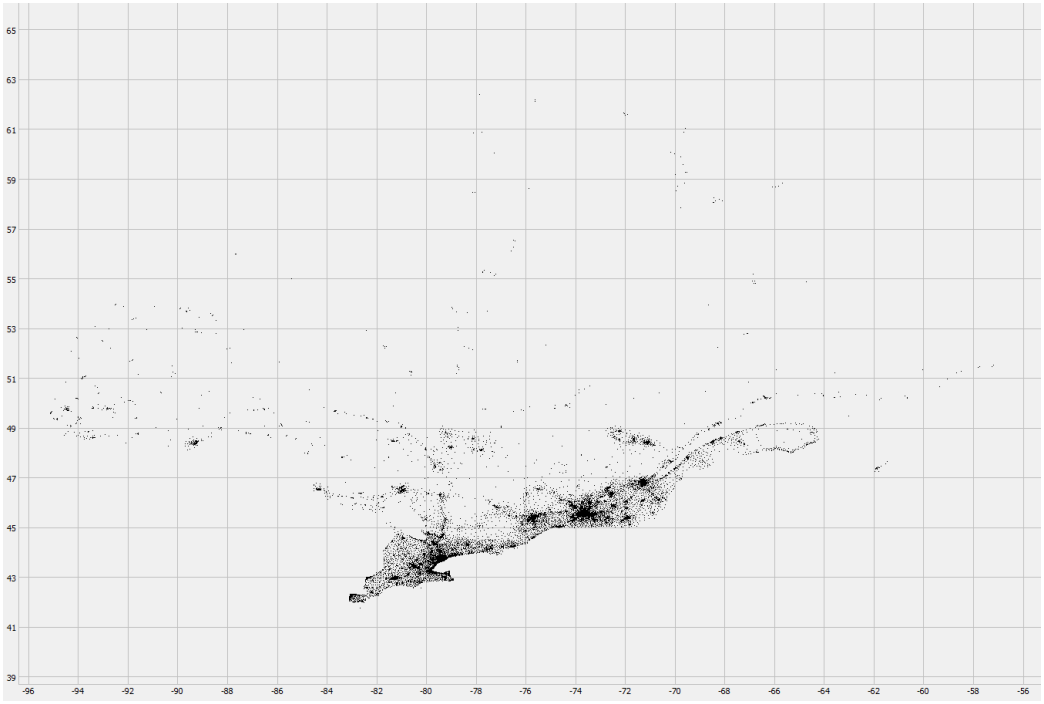
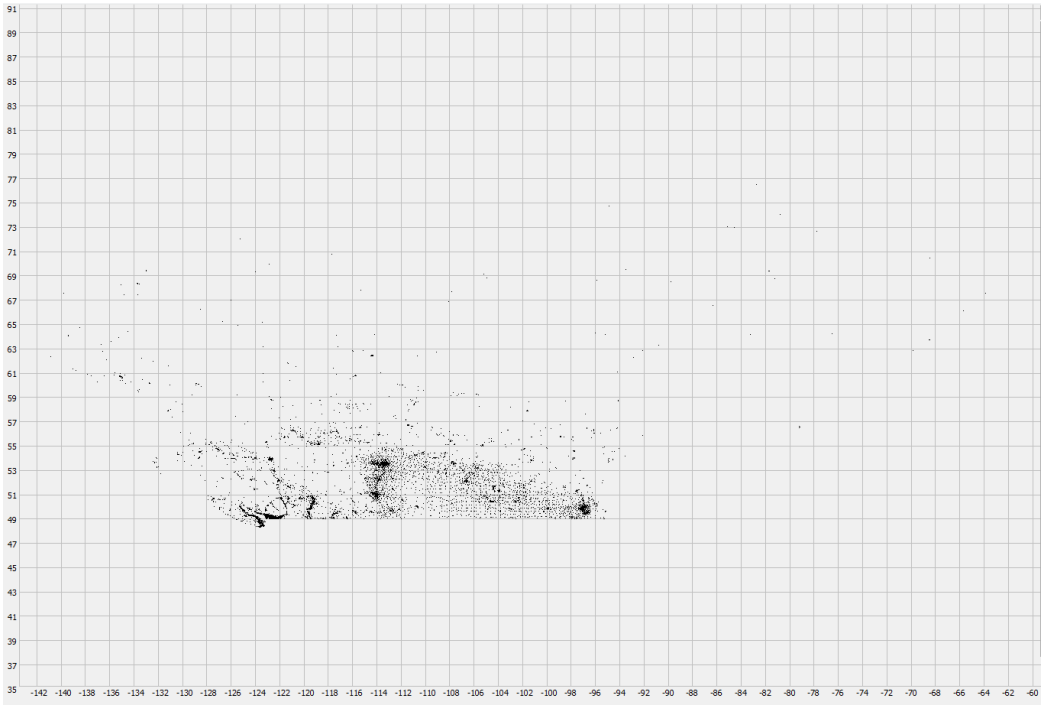


Figure 12: Western Canada Dissemination Areas



5.1.2 Considerations for Generalization

Due to the fact that the original testing data set and its subsets use the chosen quasi-identifiers in their original form, the number of potential equivalence classes may be very high if enough of the quasi-identifiers are selected during the aggregation process. For example, if all quasi-identifiers but marital status are chosen then the total number of possible equivalence classes is 11,753,280. This is roughly $\frac{1}{3}$ of the population of Canada, meaning that achieving k-anonymity using the entire Canadian population for any reasonable k value is highly unlikely without suppressing the majority of the records. For this reason, a generalized version of each data set is also created for testing purposes. Generalization was applied to each of the testing data sets, creating a generalized version of each one. The numbers of response categories for each quasi-identifier before and after generalization are shown in the Table 6.

Table 6: Data Set Generalization

Quasi-Identifier	**Number of Response Categories Before**	**Number of Response Categories After**
Age	22	11
Gender	2	2
Marital Status	6	6
Highest Level of Education	15	8
Ethnicity	53	53
Religion	16	16
Income	21	7

The generalization applied to this data is not ideal in many ways. It was done ad hoc and does not address some of the quasi-identifiers which have difficult response categories to generalize. The main goal of this generalization was simply to reduce the number of possible equivalence classes in order to better understand the performance

of the aggregation process across a greater number of quasi-identifiers. Further details on the benefits of generalization and how to properly achieve it can be found in the section on future work.

5.2 Implementation Testing

In order to compare the different approaches, each combination of approaches has been applied to 3 different selections of attributes on both the regular and generalized versions of the data sets for the Eastern and Western regions. This provides 12 different scenarios in which the combinations can be tested. In addition to running the aggregation for the approximated number of sites, aggregations are run for numbers of sites at 10% and 20% above the approximation as well as 10% and 20% below the approximation. This is done in order to determine how appropriate the site number approximation is in comparison to the other aggregations.

All test results are presented in the tables of Appendix A in Tables 7 - 18. Each table shows the results of all approach combinations for a single scenario. The removal of Pareto dominated solutions was disabled when logging these results in order to provide a complete listing of all solutions produced. It should be noted that only Tables 7, 8, 10, 12, and 14 contain solutions which use ADC. This is due to the fact that in the other scenarios, the higher number of equivalence classes and records caused prohibitively long running times when ADC was employed. Further details are provided on this in the discussion.

It should be noted that there is a limitation in terms of the memory allocation for some of the data structures used in the implementation. The current implementation cannot handle certain cases where there is a very large number of records in the data set or a large number of potential equivalence classes. Although this did not occur in any of the test scenarios presented, it could manifest when more attributes are

selected for anonymization or a larger data set is used. As the application is a 32-bit implementation, there are limitations imposed on the maximum sizes of the containers used as well as the maximum amount of contiguous memory that can be allocated. Part of the future work will entail making a 64-bit implementation to address this issue.

In Appendix B, Graphs 20 - 37 are included to visually compare the results from the different combinations of approaches used. The graphs pertain only to the tests run for Scenario 1, the results of which can be found in Table 7. Graphs were not included for the other scenarios as this would require in excess of 200 graphs to show all comparisons. The first 6 graphs in Appendix B each compare a different measure across all different site location approaches using the naive anonymity-based site number approximation approach. The second set of 6 compare the same measures, this time using the MaxCombs site number approximation approach. The final set of 6 once again compare the same measures, using the entropy site number approximation approach. Although graphs are not included for the other scenarios, the trends which are discussed for the graphs of Appendix B are representative of the trends found in the data for the other scenarios.

The graphical representations of certain testing aggregations can be found in Appendix C. Figures 38, 39, and 40 show the different aggregations created by each of the 3 site location approaches for the Eastern data set with the age and marital status attributes and the use of the anonymity site number approximation approach. Figures 41 and 42 show the differences in the aggregations between the regular and generalized versions of the Western data set using the age, gender, and income attributes with the MaxCombs site number approximation approach and the balanced density site location.

5.2.1 Discussion of Results

There are 4 main factors to be observed in the results: the level of suppression, the distance measurement, the information loss metrics, and the running time. The relationships between these factors and approaches employed serve as an indication of the quality of the approaches.

Site Location Approaches

The comparisons discussed here are restricted for now to the naïve anonymity-based, naïve density-based and balanced density-based site location approaches. All observations related to the ADC variants are deferred to a later section. In general, it can be seen that the balanced density site location approach provided superior results to the naïve anonymity-based and density-based site location approaches in terms of suppression, discernibility and non-uniform entropy. For all 3 suppression graphs in Appendix B, the balanced density outperformed the other 2 approaches by a large margin. It tends, however to produce a higher level of precision loss. Once again, it can be seen that in all 3 graphs, the balanced density had the highest levels of precision loss out of the 3 approaches. Its distance measures are consistently superior to those of the naïve density-based approach however in comparison to those of the naïve-anonymity based approach, in some cases they are superior and in other cases they are inferior.

Suppression and Information Loss Comparison The lower level of suppression produced by the balanced density-based approach is due to the fact that the distribution of sites matches the distribution of initial region points more closely than the other approaches, leading towards regions with more balanced levels of anonymity. This also accounts for the lower measures of information loss in discernibility and non-uniform entropy. It is likely for this same reason that the balanced density-based

approach has a higher level of precision loss. Recall that precision loss of an aggregated region is a function of how far up it has traveled in a generalization hierarchy structured as a binary tree. The measured value is the average of the precision loss across all aggregated regions. If a small number of regions are very high up on the hierarchy then they have a high level of precision loss, however since they are high up, it means that they have taken many of the initial regions into their aggregation. This implies that there are few initial regions left to be distributed among the other aggregated regions, meaning that the other aggregated regions will be rather low on the hierarchy. In such a scenario, the more numerous aggregated regions which are low on the hierarchy pull down the average precision loss causing the resultant value to be lower than the precision loss which would be measured in a case where the distribution were even. This suggests that the modified precision loss measure may not be appropriate for this context.

Average Distance Comparison The cases in which the naïve anonymity-based approach beats the balanced density-based approach in terms of the distance measure can be attributed to the fact that the balanced density-based approach will place most of its sites close together in high density areas, leaving the sparsely populated areas with fewer sites and thus longer distances between the points in the sparsely populated areas and the nearest site. The naïve anonymity-based approach places its sites based solely on the lowest levels of anonymity of the initial regions. This means that it takes no considerations for any global aspects of the data. Due to this, in some cases it may happen to place more sites in the sparely populated areas, significantly decreasing the distance measures for the points in these areas. While this does decrease the global distance measure, it is not consistent and is detrimental for other aspects of the aggregation, making the balanced density-based approach a more desirable choice.

Running Time Comparison It can be seen that the running times of these 3 site location approaches are all very similar. As they all have similar time complexities, this is to be expected. In the graphs of Appendix B, the 3 approaches are almost completely level with each other. The vast majority of these times is actually spent during the loading of the data into the initial regions to fill the equivalence class and anonymity data structures. The time spent on the rest of the anonymization process is in fact a small fraction of the total running time. In the Eastern region, the loading of the records took roughly 10.9 seconds on average and in the Western region, roughly 45.6 seconds.

Based on these observations, the balanced density-approach is recommended for site location from among these 3 options as it consistently performs the best in terms of information loss reduction, has a comparable run time to the others and creates reasonably compact regions, sacrificing a better distance measure as necessary in order to keep information loss lower.

ADC The discussion so far has addressed only 3 of the 4 site location approaches. The reason for this is that the quality of the results produced by ADC is heavily dependent on the method of seeding which is supplied. As such, ADC acts more as an augmentation to an existing site location approach than as a standalone approach. When comparing the results of a site location approach from the tests it can be seen that the ADC version tends to show improvement on suppression, distance, discernibility, and non-uniform entropy values. In each of the relevant graphs in Appendix B, this is easily observable as the lighter shaded ADC bars are all lower than their corresponding darker shaded non-ADC counterparts. The most noticeable reductions are in the levels of suppression. This is due to the fact that the ADC algorithm optimizes only for anonymity, which is expected to cause a reduction in the number of records that must be suppressed. Improvements in information loss

may occur as a product of this but they are not an objective of the algorithm. As expected, by the previous discussion, the precision loss values rise rather than fall. Once again, this is reflected in the graphs as well.

A significant difference in the ADC results is that running time can be much higher. In some scenarios there was only a difference of a few seconds or less whereas in others, ADC took in excess of an extra minute. Due to the high time complexity of this approach and the potential for a great number of iterations of optimization, it has the potential to take a much longer time to complete. In many cases these running times were prohibitively long, thus tests were simply not run for the instances in which the process was running in excess of a few minutes. In the graphs of Appendix B, while all other approach combinations sit nearly even with each other, many of the bars for the ADC approaches are much higher. It is clear from the results that this running time has a strong relationship to the number of sites to be placed and the number of equivalence classes in that data set. The graphs showing the running times for the MaxCombs site number approximation approach show ADC running times which are closer to the other running times than in the other site number approximation comparisons. This is due to the fact that MaxCombs makes the lowest approximations for the site numbers, allowing ADC to run more quickly. As such, this approach is recommended only for cases where the number of sites and the number of equivalence classes are sufficiently low.

Site Number Approximation Approaches

There is a clear relationship between the number of sites used and each of the measured results. The level of suppression rises as the number of sites rises. This is due to the fact that a higher number of sites means that the records are distributed across a larger number of aggregated region. With the records spread more thinly, levels of anonymity will be lower, requiring more suppression in order to reach the

desired level of anonymity. The distance measurement, however, drops as the number of sites increases. Trivially, if more sites are added then some points will be closer to the newly added sites than their previous closest site. Similarly, the information loss values drop as the number of sites increases. Since a greater number of sites implies a greater number of aggregated regions, the aggregated regions will also be smaller. These more numerous and smaller regions are more similar to the initial regions than a smaller number of larger aggregated regions would be. All of these observations are reflected in the graphs of Appendix B as the bars in each graph clearly show an increase with the number of sites for comparisons on suppression, whereas they show a decrease for comparisons on distance and information loss. Finally, the running times are once again very comparable across the different numbers of sites. A small increase in the amount of time taken can be seen as the number of sites increases however this is a very small fraction of the total time. The exception to this is for the cases in which ADC is employed as the number of sites then significantly increases the running time.

Across the 3 site number approximation approaches, the MaxCombs approach tends to make the lowest approximations, followed by entropy, and the naïve anonymity-based approximation is the highest of the 3. The naïve anonymity-based approximation however is more sensitive to high numbers of equivalence classes, causing its approximation to drop much more quickly as the number of equivalence classes rises.

The selection of the best approach is largely dependent on the requirements of the end user for the data set. If the reduction of suppression is the most important factor then MaxCombs should be used to provide a low approximation. If the user wants to reduce information loss or maintain a higher number of more compact regions then the entropy or naïve-anonymity approximations are a better choice.

Scalability

As can be seen when comparing the tables for different scenarios in the same region, the number sites and the number of equivalence classes has a very small effect on the running times of all approaches except for those using ADC. With the ADC approaches, it is clear that both the number of sites and the number of equivalence classes have a large effect on the running time which makes ADC infeasible for instances where the variables are too high.

The factor which has the greatest influence on the running time of the non-ADC approaches is the total number of records in the data set. As can be seen when comparing the running times of the test from the Eastern region and the Western region, the running times differ significantly. The number of initial regions also plays a part in this as can be seen from the time complexities of the approaches, however when using standardized initial regions such as dissemination areas, the number of initial regions is essentially a function of the number of records. Considering the time complexities of the approaches and the running times presented here, it can be concluded that all non-ADC approaches can easily handle scaling for scenarios with large data sets.

5.3 GeoLeader Comparison

A comparison is conducted between the results produced by VBAS and those of another aggregation system called GeoLeader [9]. The GeoLeader system employs the seeded growth algorithm of [29] to grow polygonal districts until a certain criterion is satisfied. Growth of the districts occurs by merging adjacent regions or districts together. In GeoLeader, the seeded growth begins with a set of postal code areas which act as the initial regions. The criterion for when a district no longer needs to be grown in when its population passes a cutoff size. This population cutoff is

determined by a calculation using the MaxCombs models of [11]. Once all regions are merged into districts which have reached the required cutoff size, the process stops.

This process requires a maximum combinations value and information about the initial regions as input. The maximum combinations value is the total number of possible equivalence classes in the data set which is being anonymized. This value is found by calculating the product of the numbers of response categories across all quasi identifiers in the data set. The regional information required constitutes a coordinate representation of each initial region, a radius for each region and adjacency information. The output of the process is a set of lists representing the final aggregation. Each list is an aggregated region consisting of the IDs of the initial regions which make up the aggregated region.

To compare the results of GeoLeader, we have written a program to process the GeoLeader output and make measurements on the same criteria as presented in the previous test results. Since GeoLeader does not guarantee anonymity, it was necessary to first apply local suppression to the aggregated regions before the measurements could be made. It should also be noted that the precision loss metric was not included in these tests; in addition to the previously mentioned flaw with the adaptation of this metric, populations can differ greatly from one postal code area to another, making the binary tree hierarchy inappropriate for this regionalization. An additional distance measurement is also added in for this comparison. Since the regional point representation of an aggregated region in GeoLeader is set to the centroid of the aggregated region, the regular distance measure is likely to be much worse for GeoLeader than for VBAS which places its sites in densely populated areas. This creates a bias towards VBAS since higher distance measures in GeoLeader would not necessarily indicate less compact regions. An alternative distance measurement is therefore used which computes a new regional point for each aggregated region at the median of the initial region points in the region. The average is then computed for the distances

between each initial region point and the alternative point representation of their aggregated regions. This measurement is applied to both systems in order to conduct a more fair comparison.

5.3.1 Test Data

The input data for GeoLeader requires each initial region to be supplied with a population, coordinates, and a radius. Since such data could not be readily obtained, the test was restricted to a data set of the postal code areas of Prince-Edward Island (PEI), Canada, which was available with GeoLeader. The point representations of the PEI postal code areas are shown in Figure 13. To accommodate for this, we have created a PEI subset of the Eastern Canada testing set. Since the GeoLeader regions are postal code areas rather than dissemination areas, it was necessary to use mapping of postal codes to dissemination areas in order to populate the postal codes areas with records. While only VBAS requires records to be associated with the initial regions in order to run, it was necessary to place the records into the postal code regions in order to later analyze of the GeoLeader results. This mapping process was only needed for the purpose of running this comparison; normally, neither process on its own would require this step.

To handle the mapping, the Postal Code Conversion File [62] maintained by Statistics Canada was used. Unfortunately, the mapping between these areas is not a simple one. A postal code may straddle multiple dissemination areas and similarly, a dissemination area may straddle multiple postal codes. An assumption was thus made that the records of a dissemination area would be evenly distributed among the postal codes mapped to that dissemination area. Given the nature of the mapping and the fact that postal code population data is not readily available, this was the only approach which seemed feasible.

Figure 13: PEI Postal Code Areas

5.3.2 Discussion of Results

The results from the GeoLeader tests can be found in Appendix D in Tables 19 and 20. Tests were run for 9 different combinations of attributes on both the regular and generalized version of the data set to give a total of 18 different scenarios tested. The results of VBAS tested on the same 18 scenarios can be found in Appendix E in Tables 21 and 22. Following the findings from the previous testing done on VBAS, only the balanced density-based site location approach and its ADC variant were used as these were found to be the most effective of the approaches. Additionally, the number of sites used in VBAS was matched to the number of final regions in the GeoLeader results in order to focus the comparisons on how the system handles anonymization for the same number of aggregated regions. While the VBAS results for other numbers of sites could be compared as well, such a comparison reveals little information since it is already known from the other tests that higher numbers of

sites will produce greater suppression but lower information loss and more compact regions whereas lower numbers of regions will produce the opposite effects.

From the graph in Figure 14, it can be seen that the levels of suppression in both systems are quite comparable. In some scenarios VBAS has a slightly lower level of suppression and in others it is slightly higher. Both systems aim to create aggregated regions which satisfy a size criterion based on the MaxCombs calculation. Since they achieve these aggregated regions through different means, the composition of initial regions in each aggregated region differs between the two aggregations. This means that the local suppression which is run after aggregation will suppress different records. This accounts for the very similar levels of suppression and the small differences between the two systems. In Scenario 6, it can be seen that the ADC approach has produced a significantly lower level of suppression. This is likely due to the fact it was able to make a greater number of moves before reaching convergence and thus reduced the level of suppression further.

Figure 14: Suppression Comparison

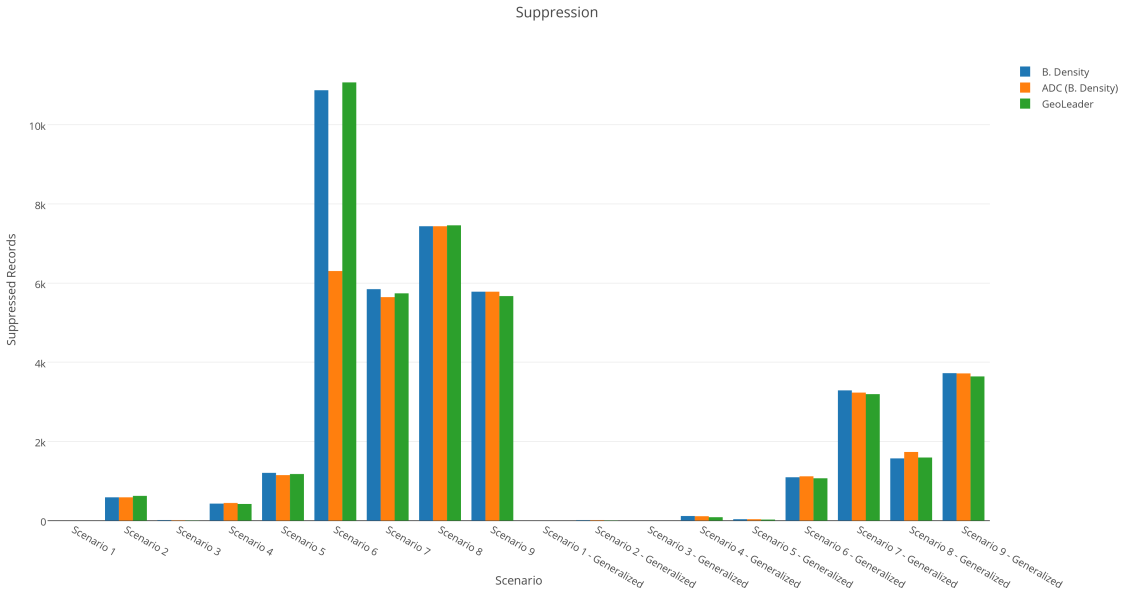


Figure 15 shows a comparison using the regular distance measure. The balanced

density approach shows results which are consistently superior to GeoLeader. Since the balanced density approach places its sites in areas densely populated by the initial region points, it produces a very low average distance measurement for PEI. As can be seen from Figure 13, the majority of the points are grouped very closely together. In Figure 16, a comparison using the alternative distance measure can be seen. This graph shows measurements which are much closer together although in most scenarios, VBAS still has a slightly better results. It is likely that the regions produced by GeoLeader are slightly less compact due to the fact that it does not guarantee the aggregated regions to be convex as VBAS does. In some scenarios, the ADC distance can be seen to be much higher than the balanced density version. This is likely caused by the acceptance of optimization moves which improved the levels of anonymity but also resulted in one or more sites being moved quite far from their starting locations. Such a shift away from the densely populated area would cause an increase in the average distance measurement.

Figure 15: Average Distance Comparison

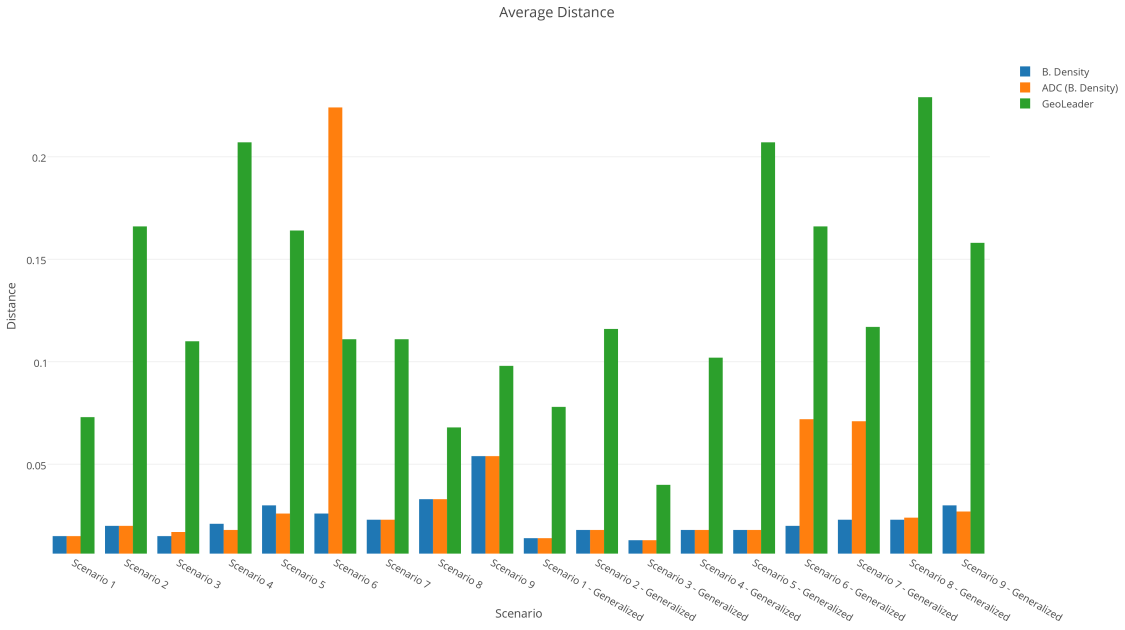
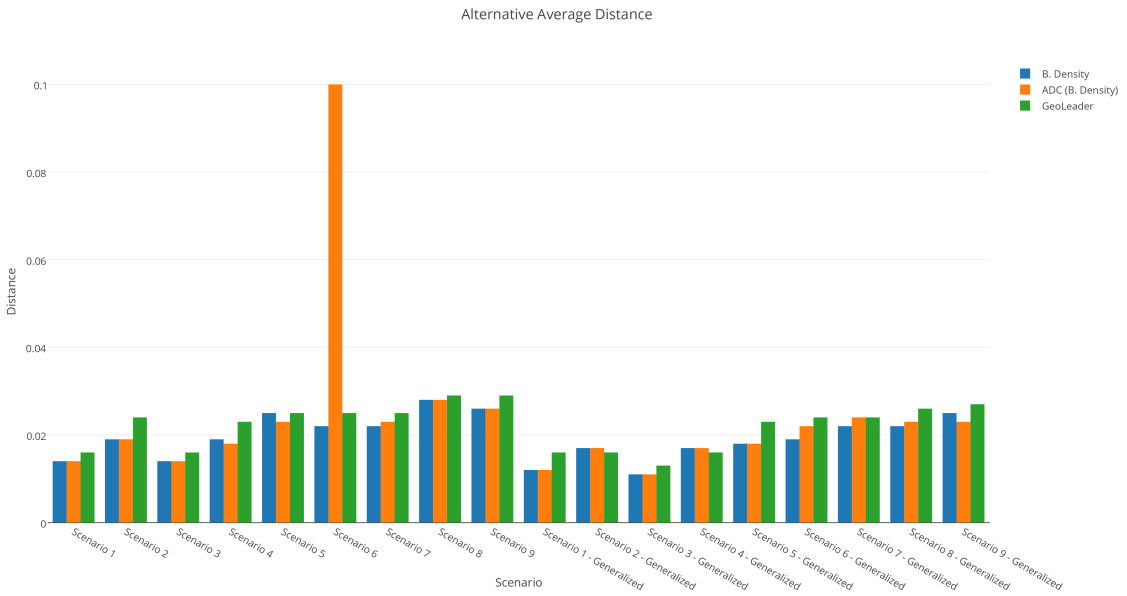


Figure 16: Alternative Average Distance Comparison



Figures 17 and 18 compare the discernibility and non-uniform entropy measurements. For both measures, the two systems are fairly evenly matched. As with the suppression measurements, in some scenarios VBAS is marginally better, and in others, GeoLeader is marginally better. Once again, this can likely be attributed to slightly different composition of initial regions across the final aggregations. A more even distribution of the equivalence classes across the final regions will produce lower levels of discernibility. Regions consisting of fewer initial region points, or non-uniform distributions of their population across the points will produce lower non-uniform entropy values. The two systems seem to have relatively even capabilities in terms of information loss reduction. It should be noted that in Scenario 6, the ADC value for both information loss measures is noticeably higher than the other values. Scenario 6 was also the scenario in which ADC produced a much lower level of suppression. This seems to indicate that as it made additional moves to reduce the level of suppression, it also incurred a greater amount of information loss.

With respect to the running times compared in Figure 19, VBAS has much faster

Figure 17: Discernibility Comparison

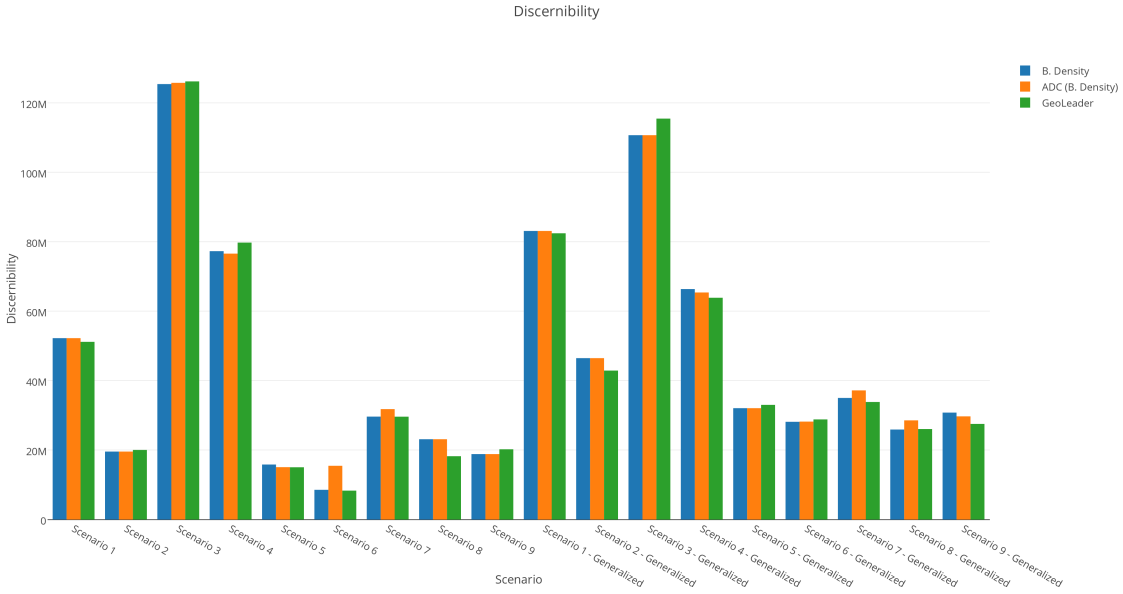
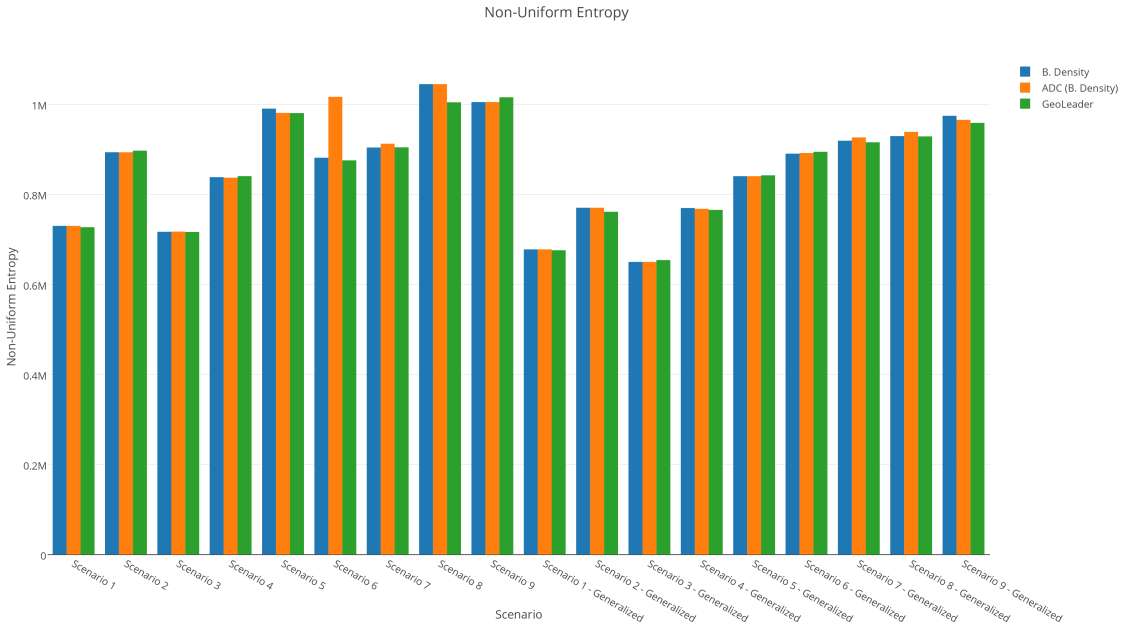


Figure 18: Non-Uniform Entropy Comparison

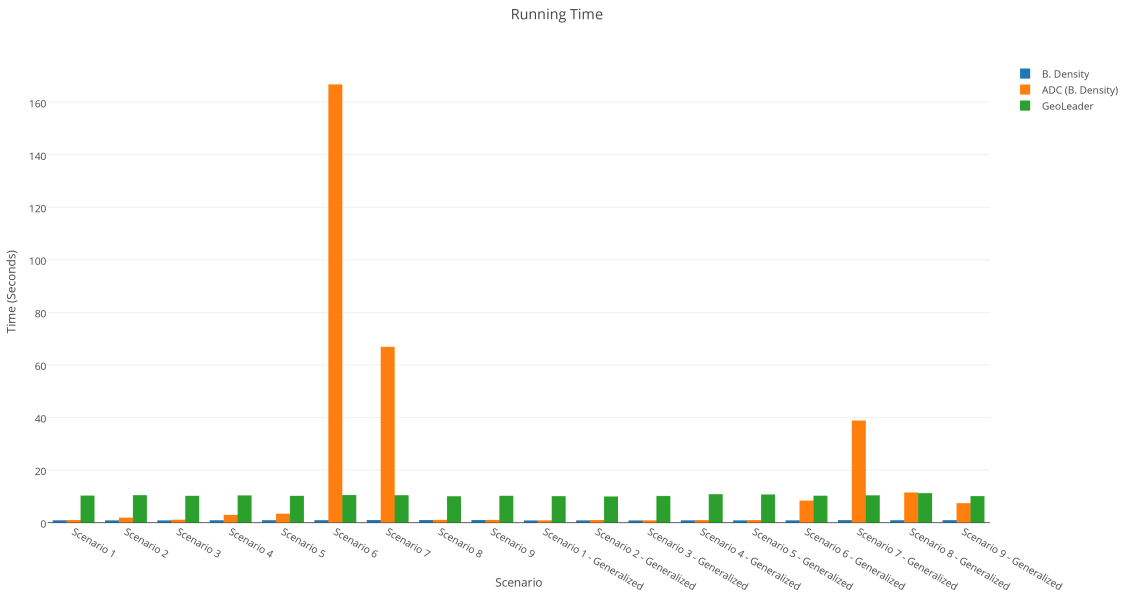


times. This is due to the fact that the majority of the lengthy work is done when loading in the records which takes linear time. As the records are loaded, the equivalence and anonymity structures are set up, which allows for the component approaches to run very efficiently. With GeoLeader, each aggregated region must be grown, starting from the initial district seeds. At each iteration, GeoLeader must check the neighboring regions of the districts and apply certain criteria to select a region to merge. Performing this operation for a large number of regions causes GeoLeader to take a much longer amount of time than VBAS. While both systems are theoretically scalable for scenarios with a larger number of initial regions, the running times for GeoLeader would become quite lengthy for large areas. For example, in Eastern Canada, there are roughly 110,000 postal code areas [62] whereas in PEI there are roughly 3,000. With a linear increase in running time, GeoLeader would take just over 6 minutes to perform aggregation in Eastern Canada whereas the testing results from VBAS in Appendix A show running times for the balanced density approach around 12 seconds on average. In some scenarios, the ADC approach shows a much higher running time. Once again, this can be attributed to a greater number of moves that it ended up checking during its iterative optimization. As expected, in Scenario 6, the running time is much longer. Other scenarios with longer running times may indicate that additional moves were considered or even committed but without any noticeable gain in the other measures.

One final observation between the two processes is that the number of aggregated regions produced by GeoLeader is much lower than the numbers of regions which would be produced by the VBAS had the number of sites not been adjusted. As preciously discussed, the appropriateness of the final number of regions depends upon the criteria in which the end user is interested. As such, this difference is not but not further analyzed.

The similarities in the levels of suppression suggest that both GeoLeader and

Figure 19: Running Time Comparison



VBAS have similar capabilities in terms of creating aggregations with sufficient levels of anonymity and in terms of the reduction of information loss. VBAS has been shown to create slightly more compact aggregated regions. In terms of running times, VBAS is much faster, taking roughly one tenth of the time to run. Based on these observations, VBAS is the preferable choice as it can produce a solutions with more compact regions and which are comparable in all other regards with the GeoLeader solutions, however it can do so much more quickly.

When comparing the input required for the two systems, both require an initial regionalization which has coordinates associated with each of the regions. In order to adjust the geographic attributes of the original data set to their aggregated values, both systems require a mapping of the data records to the initial regions. Additionally, both require information about the response categories of the attributes in the data set. In GeoLeader, this is the MaxCombs value and for VBAS, this is the second line of the data set CSV. While the user may make additional specifications in VBAS, no further information is actually required. For GeoLeader, however, there is also a need

for a radius to be associated with each initial region as well as a list of neighbors for each region or an adjacency matrix. These extra pieces of input may not always be available and thus could also be a deciding factor in which process to use. This shows another advantage of VBAS in terms of its low requirements for input.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions and Results

In this thesis we have presented VBAS, a novel system to achieve anonymity in health care data sets through the aggregation of geographic regions guided by a Voronoi diagram. We have implemented this system with various approaches for its components to conduct tests on different combinations of approaches. In order to conduct these tests, synthetic data sets were created using the National Household Survey and dissemination areas data sets from Statistics Canada.

Of the different approaches tested, the balanced density site location was found to produce the best results. ADC, a novel clustering method was found to be able to improve the results of site location approaches when the numbers of equivalence classes and records are sufficiently low. The various site number approximation approaches were found to have different merits and are thus each suited to different user requirements.

A comparison with another system of geographic aggregation, GeoLeader, has shown that VBAS can produce solutions of comparable quality with respect to the levels of suppression and information loss reduction. Additionally, the aggregated regions show slightly higher levels of compactness and the running time is much

faster. Another advantage of VBAS is that it has lower requirements in terms of the input which must be provided. The main benefits of the system are that it provides a fast, easy to use, and scalable means to anonymize data sets while incurring low levels of information loss.

The modular design of the system and its implementation serve as a testing framework which can be used to compare different selections of approach combinations and equivalence class configurations in order to analyze the solutions and discern information about the strengths and weaknesses of each approach. This design also allows for additional approaches to be added in to conduct further testing. This framework is intended to serve as a tool for testing and analysis by domain experts.

6.2 Future Work

6.2.1 Approach Improvements

Based on the fact that the balanced density site location approach proved the most effective during testing, it seems to be the case that a density-based approach has the potential to work well as long as it is able to match the distribution of the population. In the current approach, however, although each site is associated with roughly an even population when placed, this association is simply based on the initial region points which fall within its grid cell during placement and does not actually reflect the true population of the aggregated region which will be formed. To this end, the location of the sites could potentially be improved by designing a distribution function to match the distribution of the initial region points. Since the point locations are defined by geographic coordinates, such a function would have to be a continuous multivariate distribution function. Alternatively, a form of density-based clustering may serve as a reasonable substitute.

With the ADC approach, potential improvement may be found in allowing the algorithm to modify the number of sites during optimization. Since the number of sites given as input is simply an approximation, there is no need to strictly adhere to it. One way in which this might be achieved is to watch for cases where the current minimum region on which optimization is being run is surrounded by neighboring regions which also have very low levels of anonymity. In such an instance, shifting the site locations may be able to do little in terms of increasing anonymity whereas the removal of the minimum region's site would then allow for its adjacent regions to take the points from that region and increase their levels of anonymity.

With regards to the site number approximation approaches, as explained, each one provides different benefits. Due to this, a potential improvement may be to include an additional input parameter which represents some type of user requirement with respect to the results. Based on the requirement, the site number approximation can be tuned to make an approximation which will better achieve the user's desired solution. For example, if the user indicates a preference of lower suppression over lower information loss, then a smaller site number approximation would be made.

As mentioned in the literature review of Chapter 2, there is a generalization of the Voronoi diagram which handles the presence of polygonal obstacles in the plane. Since the modeling of geographic obstacles can provide additional insights during the analysis of data sets and their related regionalizations, such a diagram would be useful. Through the modular design of the process, the construction of the Voronoi diagram is also a component which can be supplied with different approaches; currently, only one approach is implemented. This generalization of the Voronoi diagram can serve as an additional approach for this component.

6.2.2 Application Improvements

While the application currently allows for the comparison of different solutions within the range of site numbers placed around the initial approximation, the comparison of solutions from different approaches requires the results to be logged in another area such as external tables or graphs. To improve the ability to compare these solutions, the implementation can be augmented to include an alternate results comparison view in which the user can add and manage the results of previously run aggregations. In this view, the user could then make direct comparisons between a wide range of solutions and apply Pareto dominance to only view non-dominated solutions.

In order to allow for a more finely grained analysis of solutions, the interface can also be augmented to provide the ability to view information about each aggregated region. This would allow users to determine if a higher degree of suppression or information loss is occurring in certain areas, such as high density of low density areas and gain further insight into why certain approaches are effective or ineffective.

Due to the fact that there are certain limitations on the data structures used in the application, a 64-bit implementation will also be developed to address these problems.

6.2.3 Generalization and Suppression

As seen in the test results when comparing the solutions produced for a data set with the solutions produced for a generalized version of the data set, the numbers of response categories for the selected equivalence class attributes play a large part in determining the quality of the solutions. Various techniques exist for anonymization based on the generalization of all equivalence class attributes [15, 58, 64]. The incorporation of such techniques into this process can allow for scenarios with larger numbers of response categories to be addressed while still maintaining sufficiently low

levels of suppression and information loss.

In some cases, users may have different degrees of suppression which they consider to be acceptable during anonymization. To accommodate for this, the allowable level of suppression could potentially be incorporated in the site number approximation approaches. If a user should allow for a higher degree of suppression then a greater number of regions could be created, allowing for a higher degree of geographic precision to be maintained. On the other hand, if the user requires that very little suppression occurs, then a much smaller number of regions must be created to achieve this. Alternatively, this can be combined with the incorporation of generalization on the other non-geographic attributes to allow for greater sacrifices in the precision of other attributes in order to keep a low level of suppression and a high level of geographic precision. Essentially, the combination of these concepts would allow for a process which is much more configurable towards user requirements.

6.3 Concluding Remarks

With a demand for the release of health data to researchers and the necessity of high precision geographic information in order to conduct meaningful research, there is a need for anonymization processes which are able to satisfy these demands. The work in this thesis presents a new system to achieve this and provides a framework for the testing and expansion of the ideas presented here. Through testing, it has been shown that VBAS is both scalable and comparable in quality to another similar approach while providing much faster running times. Through the use of the framework that has been created, it is now possible to expand upon the approaches used in the process as well as test new ones in order to make improvements on the results.

References

[1] M. P. Armstrong, G. Rushton, D. L. Zimmerman, *Geographically Masking Health Data to Preserve Confidentiality.* Stat. Med., vol. 18, pp. 497-525, 1999.

[2] P. Arzberger, P. Schroeder, A. Bealieu, et al., *Promoting Access to Public Research Data for Scientific, Economic, and Social Development.* Data Sci. J., vol. 3, pp. 135-152, 2004.

[3] K. S. Babu, N. Reddy, N. Kumar,et al., *Achieving k-anonymity Using Improved Greedy Heuristics for Very Large Relational Databases.* Transactions on Data Privacy, vol. 6, pp. 1-17, Apr. 2013.

[4] K. Benitez, B. Malin, *Evaluating Re-identication Risks with Respect to the HIPAA Privacy Rule.* J. Am. Med. Inform. Assoc., vol. 17, pp. 169-177, Mar. 2010.

[5] M. Boulos, *Towards Evidence-Based, GIS-driven National Spatial Health Information Infrastructure and Surveillance Services in the United Kingdom.* Int. J. Health Geogr., vol. 3, pp. 1, Jan. 2004.

[6] Canadian Institutes of Health Research. (2005, Spt.). *CIHR Best Practices for Protecting Privacy in Health Research (September 2005).* [Online]. Available: `http://www.cihr-irsc.gc.ca/e/29072.html`

[7] CGAL. *Computational Geometry Algorithms Library.* [Online]. Available: `http://www.cgal.org/`

[8] J. Durham, *k-Center Problems.* Graphs, Combinatorics and Convex Optimization Reading Group, 2008.

[9] K. E. Emam, L. Arbuckle *Disclosing Small Geographic Areas while Protecting PrivacyGeoLeader.* TOPHC '13, Mar. 2013.

[10] R. J. Bayardo, R. Agrawal, *Data Privacy Through Optimal k-Anonymization.* Proc. 21st ICDE '05, pp. 217-228, Jan. 2005.

[11] K. E. Emam, A. Brown, P. AbdelMalik, *Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk.* J. Am. Med. Inform. Assoc., vol. 16, pp. 256-266, Apr. 2009.

[12] K. E. Emam, A. Brown, P. AbdelMalik, et al., *A Method for Managing Re-Identification Risk from Small Geographic Areas in Canada.* BMC Med. Inform. Decis., vol. 10, pp. 18, Apr. 2010.

[13] K. E. Emam, D. Buckeridge, R. Tamblyn, et al., *The re-identification risk of Canadians from longitudinal demographics.* BMC Med. Inform. Decis, vol. 11, pp. 46, Jun. 2011.

[14] K. E. Emam, F. K. Dankar, *Protecting Privacy Using k-Anonymity.* J. Am. Med. Inform. Assoc.,vol. 15, pp. 627-637, Oct 2008.

[15] K. E. Emam, F. K. Dankar, R. Issa, et al., *A Globally Optimal k-Anonymity Method for the De-Identification of Health Data.* J. Am. Med. Inform. Assoc., vol. 16, pp. 670-682, Sep. 2009.

[16] K. E. Emam, F. K. Dankar, A. Neisa, et al., *Evaluating the Risk of Patient Re-identification from Adverse Drug Event Reports.* BMC Med. Inform. Decis., vol. 13, pp. 114, Oct. 2013.

[17] K. E. Emam, F. K. Dankar, R. Vaillancourt, et al., *Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records.* Can. J. Hosp. Pharm., vol. 62, pp. 307-319, Aug. 2009.

[18] N. H. Fefferman, E. A. O'Neil, E. N. Naumova, *Confidentiality and Confidence: Is Data Aggregation a Means to Achieve Both?.* J. Public Health Pol., vol. 26, pp. 430-449, 2005.

[19] S. Fortune, *A Sweepline Algorithm for Voronoi Diagrams.* Proc. 2nd Annu. SOGC, 1986, pp. 313-322.

[20] E. Friedman, *Packing Unit Squares in Squares: A Survey and New Results.* Electron. J. Comb. (2009), DS#7, 2009.

[21] E. Gamma, R. Helm, R. Johnson, J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley, 1994.

[22] A. Gionis, T. Tassa, *k-Anonymization with Minimal Loss of Information.* IEEE Trans. Knowl. Data Eng., vol. 21, pp. 206-219, Jul. 2008.

[23] B. Greenberg, L. Voshell, *The Geographic Component of Disclosure Risk for Microdata.* Bureau of the Census Stat. Research Division Report Series SRD Research Report Number: Census/SRD/RR-90/13, Oct. 1990.

[24] B. Greenberg, L. Voshell, *Relating Risk of Disclosure for Microdata and Geographic Area Size.* Proc. SRMS, Am. Stat. Assoc., 1990, pp.450-455.

[25] S. Hawala, *Enhancing the "100,000 Rule" on the Variation of the Per Cent of Uniques in a Microdata Sample and the Geographic Area Size Identified on the File.* Proc. Annu. Meeting Am. Stat. Assoc., 2001.

[26] J. Hershberger, S. Suri, *An Optimal Algorithm for Euclidean Shortest Paths in the Plane.* SIAM J. Comput., vol. 28, pp. 2215-2256, Dec. 1999.

[27] Y. Jafer, S. Matwin, M. Sokolova, *Privacy-aware Filter-based Feature Selection.* IEEE Int. Conf. Big Data '14, Oct. 2014.

[28] Y. Jafer, S. Matwin, M. Sokolova, *Task Oriented Privacy Preserving Data Publishing Using Feature Selection.* Lec. Notes Comp. Sci., Adv. Artif. Intell., vol. 8436, pp. 143-154, 2014.

[29] D. Joshi, L.-K. Soh, *Redistricting Using Constrained Polygonal Clustering.* IEEE Trans. Knowl. Data Eng. vol 24, pp. 2065-2079, Nov. 2012.

[30] H.-W. Jung, K. E. Emam, *A Linear Programming Model for Preserving Privacy when Disclosing Patient Spatial Information for Secondary Purposes.* Int. J. Health Geogr., vol. 13, pp. 16, May 2014.

[31] C.-L. Kuo, H. Fukui, *Geographical Structures and the Cholera Epidemic in Modern Japan: Fukushima Prefecture in 1882 and 1895.* Int. J. Health Geogr., vol. 6, pp. 25, 2007.

[32] A. K. Lyseen, C. Nohr, E. M. Sorensen, et al., *A Review and Framework for Categorizing Current Research and Development in Health Related Geographical Information Systems (GIS) Studies.* Yearb. Med. Inform., vol. 9 , pp. 110-124, Aug. 2014.

[33] A. Machanavajjhala, D. Kifer, J. Gehrke,et al., *L-Diversity: Privacy Beyond k-Anonymity.* ACM Trans. Knowl. Discov. Data, vol. 1, article 3, Mar. 2007.

[34] N. Mohammed, B. C. M. Fung, P. C. K. Hung, et al., *Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service.* Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data, 2009, pp. 1285-1294.

[35] B. Malin, *k-Unlinkability: A Privacy Protection Model for Distributed Data.* Data Knowl. Eng., vol. 64, pp. 294-311, Jan 2008.

[36] B. Malin, *Secure Construction of k-Unlinkable Patient Records from Distributed Providers.* Artif. Intell. Med., vol. 48, pp. 29-41, Oct 2009.

[37] C. Marsh, A. Dale, C. Skinner, *Safe Data Versus Safe Settings: Access to Microdata from the British Census.* Int. Stat. Rev., vol. 62, pp. 35-53, Apr. 1994.

[38] H. Nagamochi, *Packing Unit Squares in a Rectangle.* Electron. J. Comb. (2005), #R37, 2005.

[39] M. H. C. Law, A. P. Topchy, A. K. Jain, *Multiobjective Data Clustering.* IEEE Comput. Soc. Conf. CVPR, vol. 2, pp. 424-430, Jul. 2004.

[40] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, *Incognito: Efficient Full-Domain K-Anonymity.* Proc. ACM SIGMOD Int. Conf. Manage. Data, 2005, pp. 49-60.

[41] W. Lowrance, *Access to Collections of Data and Materials for Health Research: A Report to the Medical Research Council and the Wellcome Trust.* Medical Research Council and the Wellcome Trust, Mar. 2006.

[42] S. L. McLafferty, *GIS and Health Care.* Annu. Rev. Public Health, vol. 24, pp. 25-42, 2003.

[43] D. A. Moore, T. E. Carpenter, *Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology.* Epidemiologic Review, vol. 21, pp. 143-161, 1999.

[44] A. Okabe, B. Boots, K. Sugihara, S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* Wiley, 2000.

[45] K. L. Olson, S. J. Grannis, K. D. Mandl, *Privacy Protection Versus Cluster Detection In Spatial Epidemiology.* Am. J. Public Health, vol. 96, pp. 11, 2002.

[46] B. Perry, W. Gesler, *Physical Access to Primary Health Care in Andean Bolivia.* Soc. Sci. Med., vol. 50, pp. 1177-1188.

[47] R. Pless. (2011). *Lecture 16: Fortune's Algorithm and Voronoi Diagrams* [Online, Image]. Available: `http://www.cs.wustl.edu/~pless/546/lectures/L11.html`

[48] K. Qing-jiang, W. Xiao-hao, Z. Jun, *The (P, , K) Anonymity Model for Privacy Protection of Personal Information in the Social Networks.* IEEE 6th Joint Int. ITAIC, vol. 2, pp. 420-423, Aug. 2011.

[49] Qt. *Qt-Project.* [Online]. Available: `http://qt-project.org/`

[50] M. Rezaeian, G. Dunn, S. St Leger, et al., *Geographical Epidemiology, Spatial Analysis and Geographical Information Systems: a Multidisciplinary Glossary.* J. Epidemiol. Commun. H., vol. 61, pp. 98-102, Feb. 2007.

[51] T. C. Ricketts, *Geographic Information Systems and Public Health.* Annu. Rev. Public Health, vol. 24, pp. 1-6, 2003.

[52] A. L. Rivas, A. L. Hoogesteyn, S. J. Schwager, et al., *Anticipatory Control Measures: Geographical Information Systems-Based Identification of Topographic Factors Acting as Obstacles or Disseminators in the 2001 Uruguayan FMD Epidemics.* The Global Control of FMD - Tools, Ideas and Ideals, 2008.

[53] G. Rushton, *Public Health, GIS, and Spatial Analytic Tools.* Annu. Rev. Public Health, vol. 24, pp. 43-56, Oct. 2002.

[54] J.-R. Sack, J. Urrutia, *Handbook of Computational Geometry.* Elsevier, 2000.

[55] P. Samarati, *Protecting Respondents Identities in Microdata Release.* IEEE Trans. Knowl. Data Eng., vol. 13, pp. 1010-1027, Nov. 2001.

[56] R. Seidel, *A Simple and Fast Incremental Randomized Algorithm for Computing Trapezoidal Decompositions and for Triangulating Polygons.* Comp. Geom-Theor. Appl., vol. 1, pp. 51-64, Jul. 1991.

[57] Y. M. Seoung, T. Asano, *Facility Location Problems with Obstacles.* Proc. CAS '87, pp. 237-242, Jul., 1987.

[58] J. Soria-Comas, J. Domingo-Ferrer, *Differential Privacy via t-Closeness in Data Publishing.* PST, 2013, pp. 27-35.

[59] Statistics Canada, *Canadian Community Health Survey (CCHS) - Cycle 3.1 (2005) - Public Use Microdata File (PUMF) - User Guide.* June 2006.

[60] Statistics Canada. (2011). *Individuals File, 2011 National Household Survey (Public Use Microdata Files), National Household Survey year 2011.* [Online]. Available: `http://www5.statcan.gc.ca/olc-cel/olc.action?objId=99M0001X2011001&objType=46&lang=en&limit=0`.

[61] Statistics Canada. (2011). *Dissemination Area (DA).* [Online]. Available: `http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm`

[62] Statistics Canada *Postal Code^{OM} Conversion File (PCCF), 2013.* Statistics Canada Catalogue no. 92-154-X.

[63] K. Stokes, V. Torra, *n-Confusion: A Generalization of k-Anonymity.* Proc. Joint EDBT/ICDT Workshops, 2012, pp. 211-215.

[64] L. Sweeney, *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression.* Int. J. Uncertain. Fuzz., vol. 10, pp. 571-588, Oct. 2002.

[65] L. Sweeney, *k-Anonymity: A Model for Protecting Privacy.* Int. J. Uncertain. Fuzz., vol. 10, pp. 557-570, Oct. 2002.

[66] L. Sweeney, *Guaranteeing Anonymity when Sharing Medical Data, the Datafly System.* Proc. AMIA Annu. Fall Symp., 1997, pp. 51-55.

[67] A. Vora, D. S. Burke, D. A. T. Cummings, *The Impact of a Physical Geographic Barrier on the Dynamic of Measles.* Epidemiol. Infect., vol. 136, pp. 713-720, May 2008.

[68] J. Xu, W. Wang, J. Pei, et al., *Utility-Based Anonymization for Privacy Preservation with Less Information Loss.* SIGKDD Explor. Newslett., vol. 8, pp. 21-30, Dec. 2006.

[69] T. Yu, *Epidemic Geography: A Theory of International Trade and Disease Transmission.* Paris School of Economics, 2013.

[70] D. L. Zimmerman, C. Pavlik, *Quantifying the Effects of Mask Metadata Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data.* Geogr. Anal., vol. 40, pp. 52-76, Jan 2008.

Appendix A

Implementation Test Results

Table 7: Eastern Data Set - Age, Gender

Records: 2,433,221 / Initial Regions: 4,426 / Equivalence Classes: 44							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	707	1630	0.06	0.1855	402873819	7361387	11.64
	796	2246	0.05	0.1726	352544000	6916873	11.65
	885	2870	0.05	0.1625	300800526	6461681	11.66
	974	4277	0.04	0.1465	291708394	6258568	11.65
	1063	5328	0.04	0.1354	276223977	5998983	11.66
Anonymity / Density	707	2581	0.10	0.1643	1248018151	10007552	11.79
	796	2978	0.10	0.1603	972998553	9352799	11.78
	885	3027	0.08	0.1561	1059445070	9199478	11.80
	974	3868	0.10	0.1478	814547037	8769830	11.79
	1063	4871	0.08	0.1351	916239100	8767952	11.81
Anonymity / B. Density	707	944	0.06	0.1995	344230039	7123955	11.87
	796	1246	0.06	0.1859	314188738	6771096	11.86
	885	1866	0.05	0.1731	292863814	6482146	11.86
	974	2306	0.05	0.1611	265265184	6132549	11.89
	1063	2760	0.05	0.1540	249689928	5875180	11.88
Anonymity / ADC (Anonymity)	707	1188	0.06	0.1888	395717756	7300702	25.26
	796	1602	0.05	0.1765	345451221	6843224	35.10
	885	2032	0.05	0.1671	290121519	6360478	39.61
	974	3366	0.04	0.1513	280251782	6148677	110.11
	1063	4313	0.04	0.1396	266091650	5895780	184.49
Anonymity / ADC (Density)	707	1997	0.10	0.1725	1222714203	9891430	62.52
	796	2184	0.09	0.1692	958713673	9290589	54.17

	885	2431	0.08	0.1645	1051584496	9167315	61.23
	974	2949	0.09	0.1584	744993265	8609989	75.34
	1063	3664	0.07	0.1443	895927080	8709027	125.10
	707	710	0.06	0.2020	342018136	7101662	15.18
	796	927	0.06	0.1885	310820822	6749792	16.85
Anonymity / ADC (B. Density)	885	1421	0.05	0.1764	285560579	6429082	26.56
	974	1901	0.05	0.1641	260715040	6088685	25.10
	1063	2272	0.05	0.1569	244315585	5833744	34.65
	280	407	0.14	0.2669	1414133716	11317224	11.54
	315	471	0.13	0.2530	1340336487	10988211	11.53
MaxCombs / Anonymity	350	583	0.13	0.2412	1265086669	10660227	11.54
	385	706	0.12	0.2309	1193892878	10348803	11.54
	420	795	0.12	0.2212	1153631497	10092836	11.56
	280	472	0.20	0.2570	2712711427	12728682	11.70
	315	532	0.15	0.2470	2641758274	12439174	11.69
MaxCombs / Density	350	695	0.18	0.2345	1977938460	11990561	11.71
	385	728	0.14	0.2290	1836949070	11529968	11.71
	420	1079	0.14	0.2202	1879773401	11621798	11.70
	280	118	0.12	0.3046	889356342	10462751	11.75
	315	183	0.11	0.2931	784737807	10038643	11.74
MaxCombs / B. Density	350	149	0.11	0.2809	700028026	9633568	11.74
	385	182	0.09	0.2695	646688146	9280946	11.74
	420	266	0.09	0.2605	580771163	8977507	11.73
	280	265	0.14	0.2713	1403012429	11266472	13.47
	315	309	0.13	0.2574	1327177898	10929566	13.61
MaxCombs / ADC (Anonymity)	350	359	0.13	0.2470	1238303299	10567866	14.52
	385	440	0.12	0.2360	1180500199	10277461	14.80
	420	529	0.12	0.2258	1140445518	10022453	15.16
	280	284	0.20	0.2652	2658821191	12655901	14.28
	315	456	0.15	0.2504	2637546715	12399378	13.83
MaxCombs / ADC (Density)	350	522	0.17	0.2398	1960275744	11921079	14.32
	385	623	0.14	0.2320	1832726990	11509123	17.05
	420	821	0.14	0.2278	1862635025	11554967	16.57
	280	23	0.11	0.3082	878723928	10419081	12.69
	315	110	0.11	0.2955	774845068	10002359	12.37
MaxCombs / ADC (B. Density)	350	100	0.11	0.2823	696462872	9615936	12.31
	385	50	0.09	0.2730	637348866	9229336	13.71
	420	197	0.09	0.2616	577566892	8953789	12.97
	596	1204	0.10	0.1975	619700279	8309603	11.74
Entropy / Anonymity							

	671	1344	0.07	0.1919	424390873	7541668	11.74
	746	1904	0.05	0.1797	377648225	7156574	11.76
	821	2354	0.05	0.1702	331587841	6762360	11.78
	896	3000	0.05	0.1606	299211547	6431355	11.77
Entropy / Density	596	1476	0.10	0.1976	1375828126	10345829	11.73
	671	2529	0.10	0.1724	1375687075	10062197	11.74
	746	3153	0.12	0.1590	997558716	9622776	11.73
	821	3134	0.09	0.1564	980264453	9386548	11.78
	896	3104	0.10	0.1565	1003522395	9308373	11.77
Entropy / B. Density	596	578	0.07	0.2192	443876824	7901258	11.78
	671	1006	0.06	0.2034	376033704	7398614	11.78
	746	1236	0.06	0.1922	335401115	7004742	11.80
	821	1577	0.06	0.1807	306846033	6680955	11.78
	896	1914	0.05	0.1716	292758817	6443111	11.78
Entropy / ADC (Anonymity)	596	814	0.10	0.2015	607006030	8232754	17.09
	671	970	0.07	0.1950	417149405	7482890	19.22
	746	1356	0.05	0.1835	370228237	7084411	25.73
	821	1687	0.05	0.1741	324322743	6687205	37.54
	896	2157	0.04	0.1652	288339439	6329070	38.66
Entropy / ADC (Density)	596	1222	0.10	0.2016	1368634824	10307942	30.12
	671	1956	0.10	0.1811	1362724337	10011832	50.44
	746	2333	0.12	0.1681	955857056	9493760	88.33
	821	2517	0.09	0.1635	976266359	9360599	116.33
	896	2406	0.10	0.1645	980808754	9248174	91.80
Entropy / ADC (B. Density)	596	515	0.07	0.2206	438422362	7872036	15.34
	671	758	0.06	0.2062	367947216	7351020	19.41
	746	990	0.06	0.1944	330935750	6961736	17.28
	821	1188	0.06	0.1840	301304066	6634378	21.19
	896	1310	0.05	0.1757	284912260	6383865	26.06

Table 8: Eastern Data Set Generalized - Age, Gender

Records: 2,433,221 / Initial Regions: 4,426 / Equivalence Classes: 22							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	707	286	0.06	0.1793	852299253	7537469	11.48
	796	395	0.05	0.1639	801681509	7201516	11.50
	885	479	0.05	0.1525	709227109	6803609	11.47
	974	536	0.05	0.1422	661148206	6476822	11.48
	1063	569	0.04	0.1351	584735167	6080604	11.48

Anonymity / Density	707	61	0.10	0.1643	2431402984	10008011	11.43
	796	99	0.10	0.1603	1895161179	9353288	11.44
	885	87	0.09	0.1561	2063126951	9200028	11.45
	974	110	0.10	0.1478	1586720195	8770510	11.44
	1063	112	0.08	0.1351	1784215824	8768711	11.46
Anonymity / B. Density	707	12	0.06	0.1995	669033345	7124311	11.48
	796	17	0.06	0.1859	610412863	6771665	11.48
	885	48	0.05	0.1731	568786235	6482766	11.48
	974	68	0.05	0.1611	515073663	6133444	11.48
	1063	50	0.05	0.1540	484760162	5876106	11.48
Anonymity / ADC (Anonymity)	707	243	0.06	0.1800	850692934	7526941	13.47
	796	342	0.05	0.1645	800655255	7192557	14.25
	885	457	0.05	0.1526	708935683	6800712	12.34
	974	513	0.05	0.1423	660828409	6473710	12.50
	1063	546	0.04	0.1352	584415370	6077492	12.53
Anonymity / ADC (Density)	707	52	0.10	0.1652	2423196191	9989264	12.93
	796	61	0.10	0.1629	1892478779	9342331	13.62
	885	84	0.09	0.1569	2063011065	9201082	12.46
	974	106	0.10	0.1488	1586822985	8773755	12.41
	1063	107	0.08	0.1355	1784007852	8767910	12.19
Anonymity / ADC (B. Density)	707	0	0.06	0.1998	669127623	7123335	11.83
	796	4	0.06	0.1864	610337225	6770317	12.20
	885	46	0.05	0.1732	568626621	6481330	11.95
	974	46	0.05	0.1616	514156431	6129295	12.83
	1063	50	0.05	0.1540	484760162	5876106	11.65
MaxCombs / Anonymity	344	70	0.12	0.2526	2153305076	10428079	11.34
	388	80	0.11	0.2361	2062538276	10120525	11.33
	432	102	0.10	0.2256	1864647754	9722959	11.35
	476	130	0.10	0.2133	1798889438	9462876	11.35
	520	143	0.10	0.2029	1738946019	9214201	11.37
MaxCombs / Density	344	16	0.18	0.2372	3863948859	12021915	11.47
	388	6	0.14	0.2281	3480681762	11521147	11.46
	432	4	0.13	0.2180	3242108127	11241979	11.48
	476	6	0.13	0.2154	2973993145	11093550	11.46
	520	32	0.12	0.2057	3410317898	11128434	11.48
MaxCombs / B. Density	344	6	0.11	0.2814	1378993915	9686525	11.49
	388	0	0.09	0.2678	1239154742	9253312	11.51
	432	10	0.09	0.2572	1112897557	8898662	11.50

	476	0	0.08	0.2435	999872571	8552415	11.49
	520	10	0.08	0.2348	941825660	8296330	11.50
MaxCombs / ADC (Anonymity)	344	39	0.12	0.2555	2147616089	10401937	12.39
	388	47	0.11	0.2391	2055219281	10089227	12.54
	432	69	0.10	0.2278	1859658363	9699161	12.55
	476	97	0.10	0.2154	1793900047	9439078	12.57
	520	110	0.10	0.2048	1733956628	9190403	12.59
MaxCombs / ADC (Density)	344	4	0.18	0.2380	3862990049	12014201	12.08
	388	6	0.14	0.2281	3480681762	11521147	11.55
	432	0	0.13	0.2189	3241827403	11241792	11.53
	476	0	0.13	0.2166	2972081555	11088863	11.66
	520	8	0.12	0.2077	3407839898	11122686	12.33
MaxCombs / ADC (B. Density)	344	3	0.11	0.2822	1372939086	9674225	11.83
	388	0	0.09	0.2678	1239154742	9253312	11.58
	432	0	0.09	0.2575	1112222797	8895049	11.84
	476	0	0.08	0.2435	999872571	8552415	11.60
	520	5	0.08	0.2350	940516821	8293041	11.94
Entropy / Anonymity	633	164	0.06	0.1968	885563417	7786280	11.55
	713	301	0.06	0.1781	846696987	7511930	11.56
	793	395	0.05	0.1644	802612967	7211089	11.57
	873	479	0.05	0.1537	716949725	6855865	11.55
	953	511	0.05	0.1445	671232022	6551552	11.55
Entropy / Density	633	62	0.13	0.1765	2731950362	10304655	11.22
	713	58	0.10	0.1637	2422042595	9975950	11.22
	793	99	0.10	0.1612	1897691555	9362963	11.24
	873	87	0.09	0.1563	2068778868	9231552	11.24
	953	120	0.09	0.1478	1588270493	8786422	11.26
Entropy / B.lt Density	633	22	0.07	0.2120	760759296	7581746	11.41
	713	11	0.06	0.1989	674167641	7120671	11.43
	793	18	0.06	0.1860	610250003	6772580	11.43
	873	30	0.05	0.1738	578994441	6529694	11.41
	953	62	0.05	0.1622	522983484	6189497	11.44
Entropy / ADC (Anonymity)	633	135	0.06	0.1978	883360881	7773225	12.31
	713	258	0.06	0.1788	845090668	7501402	13.10
	793	342	0.05	0.1650	801586713	7202129	13.94
	873	457	0.05	0.1538	716658299	6852967	12.07
	953	488	0.05	0.1446	670912225	6548440	12.19
Entropy / ADC (Density)	633	60	0.13	0.1767	2731435909	10301920	12.07
	713	52	0.10	0.1642	2419890381	9968791	12.23

	793	61	0.10	0.1636	1895354205	9354674	13.55
	873	84	0.09	0.1571	2068662982	9232605	12.29
	953	120	0.09	0.1478	1588270493	8786422	11.54
Entropy / ADC (B. Density)	633	20	0.07	0.2121	760231930	7580151	11.80
	713	0	0.06	0.1991	674087176	7118941	11.84
	793	4	0.06	0.1864	610337413	6771613	12.23
	873	28	0.05	0.1739	578657039	6527427	11.94
	953	45	0.05	0.1627	520871237	6177550	12.85

Table 9: Eastern Data Set - Age, Martial Status

Records: 2,433,221 / Initial Regions: 4,426 / Equivalence Classes: 132							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	352	18613	0.16	0.2665	595409930	9840376	11.79
	397	22728	0.14	0.2565	462128113	9245980	11.79
	442	27186	0.14	0.2436	410004007	8847570	11.79
	487	32358	0.13	0.2320	380945784	8517522	11.81
	532	37964	0.13	0.2202	361362663	8248463	11.81
Anonymity / Density	352	19685	0.16	0.2346	1443345607	11883949	11.86
	397	22525	0.13	0.2267	1174525804	11428820	11.85
	442	24283	0.14	0.2155	1212228160	11522331	11.84
	487	25679	0.12	0.2162	1253986914	11261599	11.86
	532	30018	0.11	0.2037	1243648222	10996776	11.84
Anonymity / B. Density	352	14943	0.11	0.2797	462772129	9579623	11.73
	397	20018	0.09	0.2646	415454939	9138527	11.75
	442	23024	0.09	0.2567	368409395	8742276	11.75
	487	28659	0.08	0.2411	335425111	8418090	11.74
	532	32786	0.08	0.2316	311806154	8135142	11.76
MaxCombs / Anonymity	200	5842	0.22	0.3332	996712635	11818885	11.61
	225	7897	0.21	0.3168	951452123	11508163	11.61
	250	9574	0.17	0.3081	766575720	10988034	11.62
	275	11738	0.17	0.2959	704037135	10665299	11.62
	300	13644	0.17	0.2863	660569564	10359327	11.62
MaxCombs / Density	200	10064	0.26	0.2766	2852939098	13997334	11.70
	225	10121	0.22	0.2808	2054326967	13364498	11.71
	250	12619	0.22	0.2668	1915465539	13109054	11.71
	275	14348	0.20	0.2604	1647394762	12564075	11.72
	300	15843	0.16	0.2470	1817334461	12571479	11.71

MaxCombs / B. Density	200	5182	0.15	0.3444	773031793	11487992	11.53
	225	6121	0.13	0.3332	671153329	11020750	11.54
	250	7404	0.12	0.3228	607297646	10670627	11.54
	275	10136	0.12	0.3050	590642204	10463339	11.56
	300	11934	0.11	0.2965	563204728	10208830	11.56
Entropy / Anonymity	567	41855	0.08	0.2144	317907884	7939330	11.37
	639	51070	0.08	0.2000	289870711	7535859	11.39
	711	60244	0.08	0.1880	264678139	7144105	11.39
	783	69141	0.07	0.1773	230182333	6751622	11.41
	855	78176	0.06	0.1664	214535059	6454449	11.42
Entropy / Density	567	31483	0.12	0.2053	1126412490	10718959	11.46
	639	41453	0.12	0.1755	931231018	10224832	11.49
	711	47640	0.10	0.1640	822013278	9926877	11.46
	783	52303	0.10	0.1581	694359204	9504399	11.49
	855	57189	0.08	0.1555	701276869	9192642	11.50
Entropy / B. Density	567	37320	0.08	0.2221	308244864	7974332	11.72
	639	44869	0.07	0.2104	260577017	7480792	11.71
	711	54340	0.06	0.1980	232770182	7035192	11.72
	783	62705	0.06	0.1882	209583221	6677222	11.73
	855	70801	0.06	0.1767	193962439	6407413	11.73

Table 10: Eastern Data Set Generalized - Age, Martial Status

Records: 2,433,221 / Initial Regions: 4,426 / Equivalence Classes: 66							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	707	17029	0.06	0.1909	463998941	7138164	11.52
	796	21701	0.05	0.1760	417923722	6744638	11.53
	885	26739	0.05	0.1634	376841486	6378735	11.54
	974	31666	0.05	0.1531	345091547	6029874	11.54
	1063	36250	0.04	0.1453	298958543	5630785	11.57
Anonymity / Density	707	17865	0.10	0.1643	1605298024	9992346	11.73
	796	19828	0.10	0.1603	1252971072	9336042	11.73
	885	21677	0.08	0.1561	1365113405	9179880	11.76
	974	24031	0.10	0.1478	1049096512	8750400	11.75
	1063	28591	0.08	0.1351	1181027348	8745930	11.75
Anonymity / B. Density	707	14709	0.06	0.1995	443007159	7099665	11.52
	796	18635	0.06	0.1859	404217555	6742052	11.53
	885	22704	0.05	0.1731	376559371	6450133	11.51

	974	27837	0.05	0.1611	341214450	6095266	11.53
	1063	31250	0.05	0.1540	321090749	5834224	11.53
Anonymity / ADC (Anonymity)	-	-	-	-	-	-	-
Anonymity / ADC (Density)	-	-	-	-	-	-	-
Anonymity / ADC (B. Density)	-	-	-	-	-	-	-
MaxCombs / Anonymity	247	1848	0.12	0.3065	1535385786	11100953	10.94
	278	2667	0.11	0.2941	1258698499	10606911	10.97
	309	3447	0.11	0.2807	1161685725	10273059	10.97
	340	3999	0.11	0.2709	1054330337	9909961	10.97
	371	4803	0.10	0.2606	989713355	9626241	10.97
MaxCombs / Density	247	3729	0.22	0.2658	3768181529	13194909	11.47
	278	4561	0.20	0.2581	3495724835	12738896	11.45
	309	5340	0.16	0.2464	3427273426	12470684	11.47
	340	6158	0.18	0.2347	3241529638	12292146	11.47
	371	6722	0.15	0.2316	2337269477	11594346	11.47
MaxCombs / B. Density	247	1248	0.12	0.3246	1204213950	10735901	11.13
	278	1980	0.12	0.3066	1135421290	10441997	11.13
	309	2409	0.11	0.2936	1030803088	10093215	11.15
	340	2505	0.11	0.2877	898672523	9641740	11.13
	371	3630	0.10	0.2726	878459825	9443886	11.13
MaxCombs / ADC (Anonymity)	247	1151	0.12	0.3174	1413485200	10892257	26.17
	278	1892	0.11	0.3046	1167356076	10405485	45.81
	309	2527	0.11	0.2911	1076677401	10072579	67.88
	340	2960	0.10	0.2814	976907960	9706693	89.33
	371	3459	0.10	0.2729	862827662	9327031	134.84
MaxCombs / ADC (Density)	247	2350	0.21	0.2936	3275873707	12792592	81.08
	278	2849	0.19	0.2837	3364630289	12490409	124.05
	309	3641	0.15	0.2714	3093348552	12142656	214.26
	340	3970	0.17	0.2667	2702676763	11738238	213.24
	371	4282	0.14	0.2623	2136853253	11238790	222.11
MaxCombs / ADC (B. Density)	247	830	0.12	0.3325	1127503023	10589628	19.71
	278	1396	0.12	0.3155	1101138806	10361235	27.35
	309	1716	0.11	0.3028	988580110	9956850	31.45
	340	2043	0.10	0.2937	878766547	9587128	58.54
	371	2718	0.10	0.2829	800069215	9261809	57.34
Entropy / Anonymity	598	13067	0.07	0.2064	573496285	7844859	11.77
	673	15607	0.06	0.1961	486493492	7322512	11.77
	748	19230	0.06	0.1836	444822053	6960768	11.78
	823	23186	0.05	0.1724	400563312	6617031	11.78

	898	27439	0.05	0.1621	370344854	6317637	11.77
Entropy / Density	598	12950	0.11	0.1855	1846612491	10432527	11.25
	673	16067	0.10	0.1723	1772136813	10031599	11.25
	748	19605	0.12	0.1585	1289205702	9604755	11.27
	823	20495	0.09	0.1563	1258129449	9360038	11.27
	898	21046	0.10	0.1588	1108944332	8958661	11.27
Entropy / B. Density	598	10381	0.07	0.2188	565596885	7852820	11.23
	673	13709	0.06	0.2031	482816330	7371442	11.21
	748	16983	0.06	0.1909	441519920	6993985	11.23
	823	20338	0.06	0.1781	403034010	6684326	11.23
	898	23150	0.05	0.1717	374019629	6388046	11.21
Entropy / ADC (Anonymity)	-	-	-	-	-	-	-
Entropy / ADC (Density)	-	-	-	-	-	-	-
Entropy / ADC (B. Density)	-	-	-	-	-	-	-

Table 11: Eastern Data Set - Age, Gender, Income

Records: 2,433,221 / Initial Regions: 4,426 / Equivalence Classes: 924							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	78	34131	0.39	0.4329	406588528	15341152	12.62
	88	42441	0.38	0.4161	393028735	15028966	12.61
	98	48659	0.37	0.4086	314552166	14415660	12.61
	108	56534	0.35	0.3981	282283889	14009155	12.61
	118	65496	0.35	0.3865	274871440	13756965	12.64
Anonymity / Density	78	36769	0.41	0.4067	555037734	16305629	12.53
	88	39236	0.37	0.3992	443578448	15752821	12.55
	98	43427	0.34	0.4001	472767561	15549370	12.55
	108	55043	0.33	0.3824	457639020	15207963	12.58
	118	57046	0.35	0.3615	480568727	15344713	12.55
Anonymity / B. Density	78	29918	0.29	0.4515	275405102	14695368	12.42
	88	38044	0.28	0.4336	258450012	14362163	12.44
	98	41866	0.22	0.4360	197793041	13659503	12.45
	108	50002	0.21	0.4225	190384824	13362829	12.45
	118	58651	0.21	0.4111	164778900	12975253	12.44
MaxCombs / Anonymity	110	58225	0.35	0.3960	280208909	13951934	12.52
	124	71538	0.31	0.3786	263403970	13609898	12.52
	138	83259	0.31	0.3684	245108284	13180223	12.53
	152	95498	0.30	0.3600	228559247	12756982	12.56
	166	110115	0.30	0.3449	223062377	12533998	12.56

MaxCombs / Density	110	51990	0.36	0.3736	452316324	15381823	12.46
	124	64842	0.36	0.3627	445775599	15113641	12.49
	138	69476	0.34	0.3525	435662449	14809346	12.49
	152	75703	0.32	0.3411	374662040	14407435	12.49
	166	99714	0.28	0.3157	332759246	13893579	12.51
MaxCombs / B. Density	110	52486	0.22	0.4140	179953557	13286937	12.40
	124	63594	0.20	0.4034	161914499	12832768	12.42
	138	73523	0.19	0.3932	145046661	12414095	12.41
	152	85930	0.19	0.3789	134285139	12072816	12.43
	166	97530	0.17	0.3676	121090115	11698385	12.45
Entropy / Anonymity	504	433859	0.09	0.2191	57217631	7617758	12.84
	567	492522	0.09	0.2056	51419649	7119444	12.87
	630	546990	0.09	0.1924	47792969	6718477	12.95
	693	603939	0.08	0.1827	42863888	6269912	12.99
	756	661119	0.07	0.1755	36475540	5765288	13.04
Entropy / Density	504	281485	0.12	0.2130	159132754	10519637	12.65
	567	324713	0.12	0.2053	148421390	9990106	12.71
	630	379102	0.10	0.1796	125269561	9386356	12.76
	693	404084	0.10	0.1600	141115806	9500800	12.79
	756	431779	0.10	0.1652	87349299	8620725	12.82
Entropy / B. Density	504	414980	0.08	0.2355	43791715	7219462	12.84
	567	466952	0.08	0.2221	40640332	6803598	12.88
	630	524517	0.07	0.2122	34420117	6261556	12.92
	693	577310	0.06	0.1987	32147038	5921151	12.95
	756	646358	0.06	0.1926	28171914	5397309	13.00

Table 12: Eastern Data Set Generalized - Age, Gender, Income

Records: 2,433,221 / Initial Regions: 4,426 / Equivalence Classes: 154							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	352	12408	0.09	0.2716	279673475	9677725	11.75
	397	15457	0.09	0.2599	237098651	9173408	11.75
	442	19793	0.08	0.2462	217453021	8817860	11.77
	487	24548	0.07	0.2337	193694344	8466160	11.77
	532	28938	0.11	0.2243	176841430	8128784	11.77
Anonymity / Density	352	18546	0.16	0.2346	788448773	11895092	12.00
	397	20980	0.13	0.2267	644209533	11441813	11.97
	442	23567	0.14	0.2155	663922412	11535393	12.00

	487	24558	0.12	0.2162	686310404	11275109	11.99
	532	28787	0.11	0.2037	684193448	11012312	12.00
Anonymity / B. Density	352	10308	0.11	0.2797	254057332	9597293	11.86
	397	13964	0.09	0.2646	228666008	9160368	11.86
	442	15993	0.09	0.2567	202705194	8766551	11.87
	487	21492	0.08	0.2411	184535650	8444189	11.86
	532	24930	0.08	0.2316	171599854	8162367	11.87
Anonymity / ADC (Anonymity)	-	-	-	-	-	-	-
Anonymity / ADC (Density)	-	-	-	-	-	-	-
Anonymity / ADC (B. Density)	-	-	-	-	-	-	-
MaxCombs / Anonymity	191	3151	0.14	0.3387	612530145	12053606	11.68
	215	4012	0.13	0.3266	515555930	11565567	11.69
	239	5399	0.12	0.3136	443658737	11153922	11.69
	263	7115	0.12	0.3010	415719268	10861099	11.69
	287	8670	0.11	0.2920	345672619	10468413	11.69
MaxCombs / Density	191	7364	0.25	0.2948	1098035190	13560118	11.66
	215	9733	0.26	0.2752	1339159075	13705484	11.70
	239	10637	0.21	0.2683	1037453065	13170516	11.68
	263	12933	0.21	0.2575	951850056	12772391	11.69
	287	14494	0.17	0.2477	1082501936	12687945	11.69
MaxCombs / B. Density	191	2355	0.15	0.3525	443734773	11665629	11.70
	215	3173	0.13	0.3384	389889104	11217625	11.70
	239	4104	0.12	0.3260	349099263	10838354	11.70
	263	5235	0.12	0.3136	328311145	10553232	11.70
	287	6845	0.11	0.3033	318933276	10392320	11.73
MaxCombs / ADC (Anonymity)	-	-	-	-	-	-	-
MaxCombs / ADC (Density)	-	-	-	-	-	-	-
MaxCombs / ADC (B. Density)	191	1311	0.14	0.3617	430378284	11563666	39.07
	215	1841	0.12	0.3485	360692897	11021145	39.16
	239	2678	0.11	0.3363	337181421	10709919	54.62
	263	3385	0.11	0.3257	297371691	10330876	151.57
	287	4112	0.11	0.3174	298404426	10193926	157.00
Entropy / Anonymity	550	30990	0.07	0.2199	170944112	8012901	11.79
	619	39135	0.06	0.2061	153726002	7590149	11.80
	688	47968	0.06	0.1931	134894413	7201497	11.83
	757	57003	0.05	0.1829	121488873	6837743	11.82
	826	66188	0.05	0.1731	112901481	6527537	11.82
Entropy / Density	550	28436	0.11	0.2080	657922208	10811192	11.80
	619	39064	0.11	0.1837	517268244	10332288	11.81

	688	52404	0.10	0.1590	589320420	10278041	11.83
	757	52106	0.10	0.1653	362560207	9456206	11.83
	826	58679	0.9	0.1568	354695077	9326861	11.84
Entropy / B. Density	550	26670	0.08	0.2281	165106525	8033279	11.83
	619	33598	0.07	0.2142	148365210	7602817	11.83
	688	42759	0.06	0.1999	138549518	7305362	11.83
	757	48703	0.06	0.1924	119859802	6834902	11.85
	826	60903	0.06	0.1772	113865005	6606574	11.86
Entropy / ADC (Anonymity)	-	-	-	-	-	-	-
Entropy / ADC (Density)	-	-	-	-	-	-	-
Entropy / ADC (B. Density)	-	-	-	-	-	-	-

Table 13: Western Data Set - Age, Gender

Records: 9,647,422 / Initial Regions: 17,613 / Equivalence Classes: 44							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	2816	12327	0.16	0.1365	5353555212	35278029	54.78
	3169	15029	0.16	0.1274	5166821894	33594777	54.82
	3522	20033	0.15	0.1163	5089850272	32486605	54.84
	3875	25278	0.15	0.1068	5031196534	31476220	54.87
	4228	30593	0.15	0.0987	4924660386	30389401	54.90
Anonymity / Density	2816	9564	0.14	0.1411	11023473410	48332488	48.39
	3169	10477	0.19	0.1425	6967477150	45459820	48.43
	3522	12645	0.12	0.1289	7220332397	44542884	48.44
	3875	12917	0.14	0.1319	7249748530	43038021	48.47
	4228	14552	0.11	0.1271	5158870690	40945436	48.50
Anonymity / B. Density	2816	7421	0.08	0.1624	1581275195	29470118	47.15
	3169	9398	0.08	0.1514	1461997689	28108965	47.17
	3522	11513	0.07	0.1431	1263966930	26462962	47.18
	3875	13742	0.06	0.1346	1115608080	24878131	47.22
	4228	15897	0.06	0.1277	1052038615	23869881	47.22
MaxCombs / Anonymity	891	1289	0.79	0.2435	32521398272	54370080	47.56
	1003	1852	0.78	0.2287	31587661643	53283195	47.55
	1115	2322	0.78	0.2163	31449590981	52428746	47.57
	1227	2864	0.78	0.2066	27888315977	51039736	47.60
	1339	3216	0.77	0.1990	27701297033	50082300	47.58
MaxCombs / Density	891	2439	0.27	0.2126	29657986717	61549660	48.09
	1003	2903	0.25	0.1996	30134246601	60735222	48.11
	1115	3250	0.24	0.1972	21245163497	58263716	48.11

	1227	3800	0.23	0.1903	22423283603	57995784	48.12
	1339	4324	0.22	0.1798	25514979696	58211659	48.14
MaxCombs / B. Density	891	539	0.16	0.2728	4752388219	45110966	47.20
	1003	686	0.16	0.2626	4277109398	43568593	47.22
	1115	922	0.15	0.2506	3862196553	42123637	47.20
	1227	1422	0.14	0.2376	3738946730	41269649	47.19
	1339	1315	0.13	0.2337	3250324593	39632472	47.22
Entropy / Anonymity	2536	10612	0.16	0.1448	5503138530	36671302	47.71
	2854	12620	0.16	0.1355	5337623531	35102059	47.75
	3172	15049	0.16	0.1273	5165698498	33581708	47.77
	3490	19557	0.15	0.1173	5095127220	32575854	47.78
	3808	24225	0.15	0.1084	5041266805	31662118	47.81
Entropy / Density	2536	8599	0.21	0.1491	10111986775	48782923	49.13
	2854	10354	0.13	0.1376	9973121586	47814552	49.14
	3172	10489	0.19	0.1425	6964190507	45435458	49.14
	3490	11986	0.12	0.1322	6940241023	44035971	49.16
	3808	12641	0.19	0.1332	7311313229	43364766	49.21
Entropy / B. Density	2536	6034	0.09	0.1716	1753720009	30972443	48.59
	2854	7907	0.08	0.1604	1574250624	29358964	48.60
	3172	9433	0.08	0.1513	1462929197	28109703	48.64
	3490	11466	0.07	0.1429	1254215294	26435508	48.63
	3808	13188	0.06	0.1362	1127258558	25065625	48.67

Table 14: Western Data Set Generalized - Age, Gender

Records: 9,647,422 / Initial Regions: 17,613 / Equivalence Classes: 22							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	2816	2359	0.09	0.1301	5077514964	33520066	46.61
	3169	2785	0.08	0.1250	4112575141	30946780	46.60
	3522	2901	0.07	0.1230	3280402536	28259071	46.65
	3875	4162	0.07	0.1095	3235296866	27599551	46.65
	4228	5463	0.07	0.0982	3189768697	26941885	46.69
Anonymity / Density	2816	606	0.14	0.1411	21316850651	48334620	47.75
	3169	596	0.19	0.1425	13481007617	45461883	47.80
	3522	776	0.12	0.1289	13974977568	44545403	47.82
	3875	825	0.14	0.1319	14039642276	43040557	47.86
	4228	875	0.11	0.1271	9979846274	40948331	47.85
Anonymity / B. Density	2816	346	0.08	0.1624	3053529929	29474569	47.32
	3169	444	0.08	0.1514	2822393811	28114605	47.35

	3522	573	0.07	0.1431	2438879881	26469786	47.34
	3875	632	0.06	0.1346	2151680051	24885748	47.37
	4228	701	0.06	0.1277	2028456731	23879012	47.39
Anonymity / ADC (Anonymity)	-	-	-	-	-	-	-
Anonymity / ADC (Density)	2816	580	0.14	0.1417	21295416534	48302745	56.88
	3169	588	0.19	0.1428	13453629060	45428872	58.57
	3522	773	0.12	0.1290	13978624910	44545832	58.06
	3875	790	0.14	0.1323	14032279304	43029110	58.32
	4228	861	0.10	0.1272	9983171179	40948502	65.58
Anonymity / ADC (B. Density)	2816	336	0.08	0.1625	3049648291	29460084	59.97
	3169	424	0.08	0.1517	2821159037	28111822	72.03
	3522	576	0.07	0.1432	2439615137	26475588	68.14
	3875	603	0.06	0.1347	2151608518	24884847	81.22
	4228	692	0.06	0.1278	2030548251	23885022	75.68
MaxCombs / Anonymity	1192	860	0.29	0.1737	36002954239	53931614	47.59
	1342	933	0.20	0.1760	16235908095	47889377	47.62
	1492	1038	0.16	0.1759	12436265297	44683388	47.62
	1642	1041	0.16	0.1789	9734465548	41546629	47.64
	1792	1212	0.14	0.1694	9147755909	40473110	47.65
MaxCombs / Density	1192	211	0.24	0.1971	42473735512	58165743	48.00
	1342	273	0.22	0.1795	49290011990	58185275	48.00
	1492	264	0.21	0.1777	31328922092	54227213	48.02
	1642	293	0.21	0.1744	26405611238	52262950	48.04
	1792	401	0.21	0.1675	24138938782	51666472	48.06
MaxCombs / B. Density	1192	49	0.14	0.2414	7335687455	41570791	47.85
	1342	48	0.13	0.2331	6258130009	39601029	47.87
	1492	55	0.13	0.2228	5598349654	38099351	47.89
	1642	138	0.11	0.2147	5076418746	36762665	47.87
	1792	123	0.11	0.2057	4627772249	35513127	47.87
MaxCombs / ADC (Anonymity)	-	-	-	-	-	-	-
MaxCombs / ADC (Density)	1192	201	0.24	0.1978	42473789621	58167667	48.68
	1342	257	0.22	0.1803	49243365087	58136607	52.88
	1492	256	0.21	0.1783	31291869831	54194621	50.69
	1642	286	0.21	0.1747	26411704175	52269962	49.23
	1792	401	0.21	0.1675	24149894908	51670280	48.90
MaxCombs / ADC (B. Density)	1192	46	0.14	0.2417	7327766464	41552567	49.80
	1342	49	0.13	0.2334	6253211086	39583428	50.55
	1492	61	0.13	0.2229	5590797748	38087565	51.28
	1642	120	0.11	0.2149	5070292102	36748677	53.20

	1792	120	0.11	0.2058	4627223735	35509064	52.15
Entropy / Anonymity	2760	2186	0.09	0.1328	5090667966	33659015	46.58
	3106	2766	0.08	0.1252	4294559570	31485522	46.62
	3452	2837	0.08	0.1240	3390360974	28677403	46.61
	3798	3904	0.07	0.1122	3245283526	27746397	46.66
	4144	5165	0.07	0.1006	3200517894	27102387	46.64
Entropy / Density	2760	600	0.14	0.1400	25846063471	49728820	47.33
	3106	586	0.18	0.1407	16271069164	46815011	47.35
	3452	709	0.12	0.1341	13452416693	44135023	47.37
	3798	849	0.19	0.1326	12166795890	43130798	47.41
	4144	880	0.12	0.1264	11152173315	41302954	47.43
Entropy / B. Density	2760	338	0.08	0.1631	3147817261	29797652	47.08
	3106	429	0.08	0.1534	2896138096	28415275	47.11
	3452	529	0.07	0.1442	2409073446	26470301	47.11
	3798	590	0.06	0.1364	2176762556	25092200	47.11
	4144	672	0.06	0.1280	2031879091	24043523	47.16
Entropy / ADC (Anonymity)	-	-	-	-	-	-	-
Entropy / ADC (Density)	2760	588	0.14	0.1404	25849751326	49738687	64.43
	3106	586	0.17	0.1410	16230004678	46778764	58.47
	3452	673	0.12	0.1346	13452936022	44141347	76.65
	3798	831	0.18	0.1331	12153697591	43104283	64.59
	4144	876	0.12	0.1265	11152618933	41304902	74.06
Entropy / ADC (B. Density)	2760	329	0.08	0.1632	3145637406	29788820	59.64
	3106	411	0.08	0.1536	2895973998	28417121	66.55
	3452	529	0.07	0.1442	2410542294	26474272	68.41
	3798	571	0.06	0.1365	2176689949	25091333	75.65
	4144	658	0.06	0.1281	2033742704	24049804	81.50

Table 15: Western Data Set - Age, Marital Status

Records: 9,647,422 / Initial Regions: 17,613 / Equivalence Classes: 132							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	1407	82211	0.18	0.2281	2293562004	38764245	48.99
	1584	97683	0.17	0.2209	1861096280	36405435	49.00
	1761	117452	0.16	0.2105	1670450675	34835911	49.04
	1938	138696	0.16	0.1990	1581347493	33724605	49.06
	2115	160735	0.15	0.1895	1471842650	32583357	49.11
Anonymity / Density	1407	71031	0.22	0.1847	15294426768	55636653	48.27
	1584	84790	0.21	0.1716	11753761055	54168344	48.27

	1761	99621	0.20	0.1617	10281872836	52150382	48.31
	1938	102997	0.16	0.1604	9318437007	51235790	48.33
	2115	114289	0.15	0.1541	8564306271	49688280	48.37
	1407	79005	0.13	0.2274	2220185866	38954950	48.57
	1584	95265	0.12	0.2167	1914979605	37223569	48.57
Anonymity / B. Density	1761	113879	0.11	0.2074	1688966662	35500889	48.61
	1938	133704	0.11	0.1948	1614074977	34585368	48.62
	2115	150397	0.10	0.1890	1438218002	33056786	48.63
	560	14217	0.27	0.3165	6311590549	52472163	48.48
	631	17758	0.26	0.3080	5037155736	50173487	48.47
MaxCombs / Anonymity	702	20790	0.23	0.3018	4194655923	48071626	48.49
	773	25711	0.22	0.2912	3932199884	46878936	48.51
	844	31400	0.22	0.2800	3775543598	45955984	48.51
	560	25633	0.32	0.2497	36042659667	68203883	50.35
	631	28157	0.30	0.2471	22235032891	64189887	50.35
MaxCombs / Density	702	32262	0.30	0.2316	33040329317	66395640	50.36
	773	34661	0.28	0.2329	22195209989	63042813	50.38
	844	38753	0.27	0.2225	23388644261	63186906	50.37
	560	13686	0.22	0.3238	5091007828	51228318	51.12
	631	18595	0.21	0.3082	4605673405	49825639	51.17
MaxCombs / B. Density	702	22362	0.19	0.2988	4206603847	48377099	51.17
	773	26502	0.18	0.2893	3699763593	46826336	51.17
	844	32996	0.17	0.2774	3616020044	46031943	51.17
	2384	191614	0.14	0.1806	1226761725	30401156	51.75
	2683	231123	0.11	0.1682	1106116236	28770419	51.79
Entropy / Anonymity	2982	270940	0.11	0.1568	1027879621	27398200	51.78
	3281	307859	0.10	0.1492	899443507	25724698	51.84
	3580	347224	0.09	0.1399	840078700	24543114	51.86
	2384	121895	0.15	0.1547	6245481709	47568498	52.48
	2683	132216	0.19	0.1432	6624601718	47933502	52.52
Entropy / Density	2982	130997	0.20	0.1399	6931522452	48437208	52.51
	3281	156829	0.15	0.1366	4974229296	44873407	52.57
	3580	143035	0.20	0.1390	4681219460	45033773	52.58
	2384	179033	0.09	0.1785	1259084837	31278146	53.63
	2683	212860	0.08	0.1663	1164582959	29896700	53.65
Entropy / B. Density	2982	244128	0.08	0.1574	1050099124	28409738	53.68
	3281	278085	0.07	0.1486	959007792	27064799	53.68
	3580	318054	0.07	0.1424	832442357	25348948	53.73

Table 16: Western Data Set Generalized - Age, Marital Status

Records: 9,647,422 / Initial Regions: 17,613 / Equivalence Classes: 66							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	2816	76338	0.08	0.1642	1877532818	28221953	52.33
	3169	94828	0.07	0.1531	1649858679	26450763	52.33
	3522	114608	0.07	0.1424	1501721212	24996875	52.40
	3875	135246	0.07	0.1329	1365832805	23647585	52.43
	4228	156730	0.07	0.1237	1267218863	22505870	52.46
Anonymity / Density	2816	56788	0.14	0.1411	14800657315	48286107	51.31
	3169	58607	0.19	0.1425	9354411192	45415623	51.32
	3522	72186	0.12	0.1289	9690137054	44488200	51.35
	3875	72198	0.14	0.1319	9770408963	42986441	51.38
	4228	80959	0.11	0.1271	6926100504	40884935	51.45
Anonymity / B. Density	2816	73736	0.08	0.1624	2120744583	29364405	48.91
	3169	90376	0.08	0.1514	1960534416	27987510	48.95
	3522	106094	0.07	0.1431	1694333943	26323786	48.93
	3875	125312	0.06	0.1346	1494965410	24719778	48.95
	4228	141624	0.06	0.1277	1409951258	23697803	48.97
MaxCombs / Anonymity	752	4489	0.21	0.2973	7205301459	46866882	50.14
	846	6738	0.20	0.2825	6712482925	45574871	50.14
	940	7904	0.19	0.2743	5742990396	43741101	50.18
	1034	9607	0.15	0.2649	5157277757	42372286	50.16
	1128	11364	0.15	0.2564	4733918699	41125036	50.19
MaxCombs / Density	752	11889	0.28	0.2316	49623582151	64045602	50.37
	846	13595	0.27	0.2227	45196668978	63207114	50.36
	940	16007	0.27	0.2182	32831642813	59610246	50.39
	1034	18131	0.24	0.2076	35702887458	59221722	50.39
	1128	21676	0.24	0.1955	28655557120	58129453	50.39
MaxCombs / B. Density	752	5399	0.19	0.2913	7583993296	47447046	49.36
	846	7090	0.17	0.2785	6969213663	46075743	49.33
	940	8430	0.17	0.2699	5924842271	44291575	49.37
	1034	11371	0.16	0.2576	5647024604	43238739	49.36
	1128	13812	0.15	0.2481	5332951419	42273836	49.33
Entropy / Anonymity	2563	64068	0.09	0.1735	2069947069	29575660	49.81
	2884	79831	0.08	0.1620	1839633059	27884638	49.85
	3205	96791	0.07	0.1519	1635122859	26304161	49.87
	3526	114803	0.07	0.1424	1499540895	24976479	49.87
	3847	133461	0.07	0.1336	1374122151	23744077	49.94
Entropy / Density	2563	50318	0.19	0.1483	13546774534	48596962	50.11

	2884	61321	0.13	0.1362	13387988574	47683541	50.13
	3205	60363	0.18	0.1411	9333173959	45262731	50.16
	3526	56899	0.21	0.1436	8984549961	45387996	50.18
	3847	71470	0.18	0.1324	9817909644	43119694	50.22
Entropy / B. Density	2563	63434	0.09	0.1710	2320352238	30710164	50.33
	2884	76712	0.08	0.1603	2090671440	29111646	50.36
	3205	92614	0.08	0.1504	1876706819	27670152	50.35
	3526	106395	0.07	0.1428	1694512365	26314982	50.35
	3847	123822	0.06	0.1353	1499173947	24792207	50.39

Table 17: Western Data Set - Age, Gender, Income

Records: 9,647,422 / Initial Regions: 17,613 / Equivalence Classes: 924							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	311	132773	0.40	0.3737	1591876458	60195537	55.92
	351	163211	0.37	0.3624	1380926923	58267743	55.96
	391	188302	0.34	0.3558	1210868662	56218879	56.00
	431	224713	0.32	0.3442	1129102679	54911558	56.08
	471	258744	0.31	0.3361	981010195	53382832	56.11
Anonymity / Density	311	160413	0.47	0.3071	8527041449	74861258	60.98
	351	182076	0.42	0.2842	8980656009	74247943	61.02
	391	202852	0.38	0.2818	5514942308	70631619	60.97
	431	235279	0.39	0.2638	7168456714	71495985	60.95
	471	261525	0.35	0.2642	5554636571	69218853	60.94
Anonymity / B. Density	311	117742	0.30	0.3821	1350493930	59118644	51.13
	351	136731	0.27	0.3765	1102509107	56591036	51.17
	391	172395	0.26	0.3621	1010068849	55237630	51.19
	431	202104	0.25	0.3514	928330933	53846936	51.18
	471	231388	0.23	0.3444	822488795	52236367	51.20
MaxCombs / Anonymity	248	86825	0.44	0.3982	1953950195	63442299	51.17
	279	105421	0.42	0.3883	1678520844	61381252	51.18
	310	132054	0.40	0.3740	1593233805	60230546	51.29
	341	151783	0.37	0.3675	1385318336	58462150	51.32
	372	176157	0.36	0.3591	1287860143	57142365	51.28
MaxCombs / Density	248	129818	0.54	0.3215	10143975519	77687377	51.29
	279	135878	0.48	0.3249	6796619289	74320098	51.29
	310	159793	0.47	0.3072	8530073948	74903558	51.31
	341	173009	0.43	0.2898	8775637940	74156323	51.34
	372	186816	0.41	0.2936	6085437726	71479696	51.32

MaxCombs / B. Density	248	76176	0.35	0.4058	1626539930	62202246	51.15
	279	92671	0.32	0.3963	1410325011	60182763	51.19
	310	115909	0.30	0.3840	1330633805	59009378	51.19
	341	140135	0.28	0.3721	1237194849	57791560	51.20
	372	155415	0.27	0.3683	1074322319	55960593	51.24
Entropy / Anonymity	2015	1858150	0.15	0.1943	218706574	28415266	52.55
	2267	2114049	0.14	0.1830	195401329	26293570	52.77
	2519	2359912	0.14	0.1729	175673736	24373673	52.93
	2771	2594043	0.12	0.1639	156014785	22616853	53.15
	3023	2827514	0.12	0.1557	142272881	21066017	53.36
Entropy / Density	2015	959274	0.20	0.1578	1289673446	48948021	53.13
	2267	1008382	0.20	0.1532	1054859099	47070279	53.17
	2519	1046021	0.21	0.1492	1019217514	46749367	53.22
	2771	1189987	0.13	0.1403	972803970	45403285	53.35
	3023	1124865	0.20	0.1395	994630426	46017324	53.34
Entropy / B. Density	2015	1730561	0.10	0.1942	215195372	29300367	52.63
	2267	1974769	0.09	0.1829	191770079	27152633	52.75
	2519	2189419	0.09	0.1727	174969656	25574971	52.95
	2771	2403510	0.08	0.1629	159807553	23947198	53.10
	3023	2581958	0.08	0.1557	147907205	22738852	53.29

Table 18: Western Data Set Generalized - Age, Gender, Income

Records: 9,647,422 / Initial Regions: 17,613 / Equivalence Classes: 154							
Approaches	**Sites**	**Suppression**	**Avg. Dist.**	**Prec. Loss**	**Discernibility**	**Entropy**	**Time**
Anonymity / Anonymity	1407	46166	0.15	0.2305	1126012331	38611250	49.14
	1584	57125	0.13	0.2213	908741183	36380346	49.16
	1761	73687	0.13	0.2097	843483705	35065486	49.18
	1938	92241	0.12	0.1991	782437666	33844055	49.22
	2115	107960	0.10	0.1918	684355615	32305513	49.25
Anonymity / Density	1407	72635	0.22	0.1847	7886116277	55687944	49.55
	1584	89236	0.21	0.1716	6083843111	54222083	49.57
	1761	108435	0.20	0.1617	5333227768	52207900	49.59
	1938	112261	0.16	0.1604	4805671818	51295258	49.62
	2115	128015	0.15	0.1541	4419498085	49748879	49.64
Anonymity / B. Density	1407	49393	0.13	0.2274	1155119953	39068249	50.68
	1584	63363	0.12	0.2167	997027653	37345762	50.72
	1761	75668	0.11	0.2074	879603000	35640343	50.72

	1938	98092	0.11	0.1948	840795196	34729082	50.73
	2115	109846	0.10	0.1890	749767700	33216616	50.78
MaxCombs / Anonymity	526	4036	0.32	0.3275	3205804805	52888938	49.98
	592	5428	0.24	0.3184	2551669602	50583772	49.98
	658	6713	0.23	0.3078	2370072708	49236371	49.97
	724	8819	0.22	0.2978	2189735700	47989228	50.00
	790	10347	0.21	0.2914	1867876702	46319719	50.00
MaxCombs / Density	526	20197	0.34	0.2545	19339085768	68771484	49.76
	592	20369	0.31	0.2533	17031837399	67461136	49.75
	658	29507	0.29	0.2309	16139852508	66315363	49.77
	724	32031	0.29	0.2282	12570717341	64292960	49.78
	790	35336	0.28	0.2231	12671488442	63475612	49.77
MaxCombs / B. Density	526	6282	0.23	0.3256	2858617661	52364304	50.25
	592	6978	0.22	0.3172	2516041572	50525262	50.25
	658	9409	0.20	0.3053	2268623947	49095268	50.28
	724	11502	0.19	0.2944	2100357798	47954234	50.29
	790	12901	0.18	0.2887	1876475510	46508722	50.31
Entropy / Anonymity	2265	125292	0.10	0.1846	650553523	31414051	50.72
	2549	159326	0.09	0.1726	591450458	29832627	50.77
	2833	195181	0.08	0.1625	521699780	28200021	50.81
	3117	236016	0.08	0.1528	479347097	26884549	50.89
	3401	272279	0.07	0.1459	424297990	25358143	50.90
Entropy / Density	2265	129803	0.20	0.1535	3859695014	49095936	51.16
	2549	141143	0.20	0.1485	3637471026	48643799	51.20
	2833	164276	0.13	0.1390	3588357357	47746233	51.22
	3117	166377	0.18	0.1399	3025722589	46665465	51.27
	3401	191115	0.12	0.1356	2296993575	43505286	51.27
Entropy / B. Density	2265	124722	0.09	0.1828	695019660	32166333	50.77
	2549	156254	0.09	0.1715	632573168	30689626	50.79
	2833	190556	0.08	0.1609	574925661	29222006	50.82
	3117	222494	0.08	0.1531	541081620	28062689	50.85
	3401	254895	0.07	0.1462	478055322	26713544	50.85

Appendix B

Comparison of Approach Combinations from Tests on the *Eastern Canada - Age, Gender* Scenario

Figure 20: Anonymity - Suppression Comparison

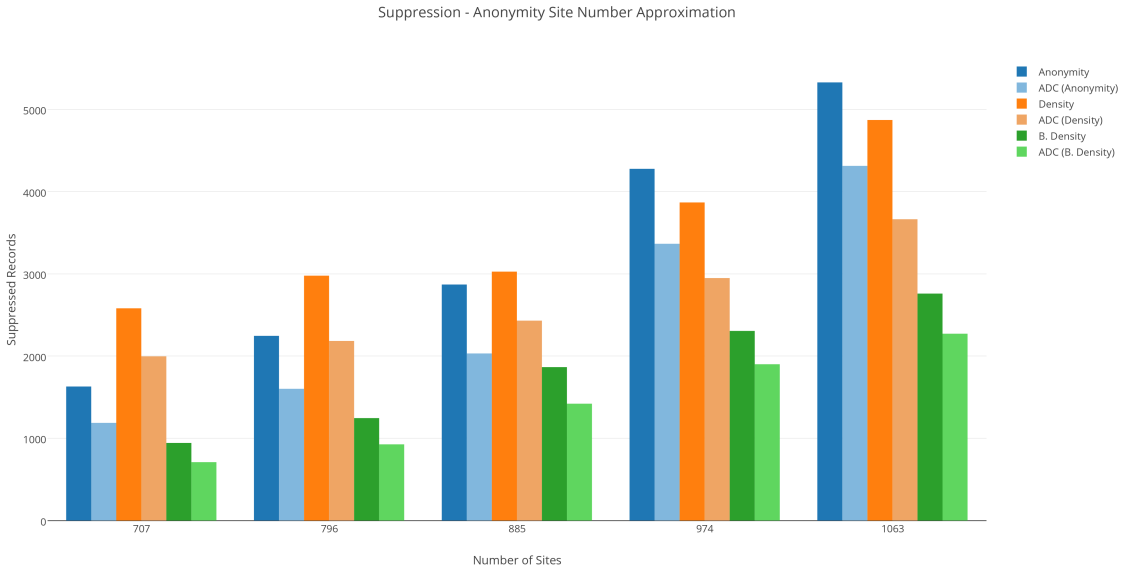


Figure 21: Anonymity - Average Distance Comparison

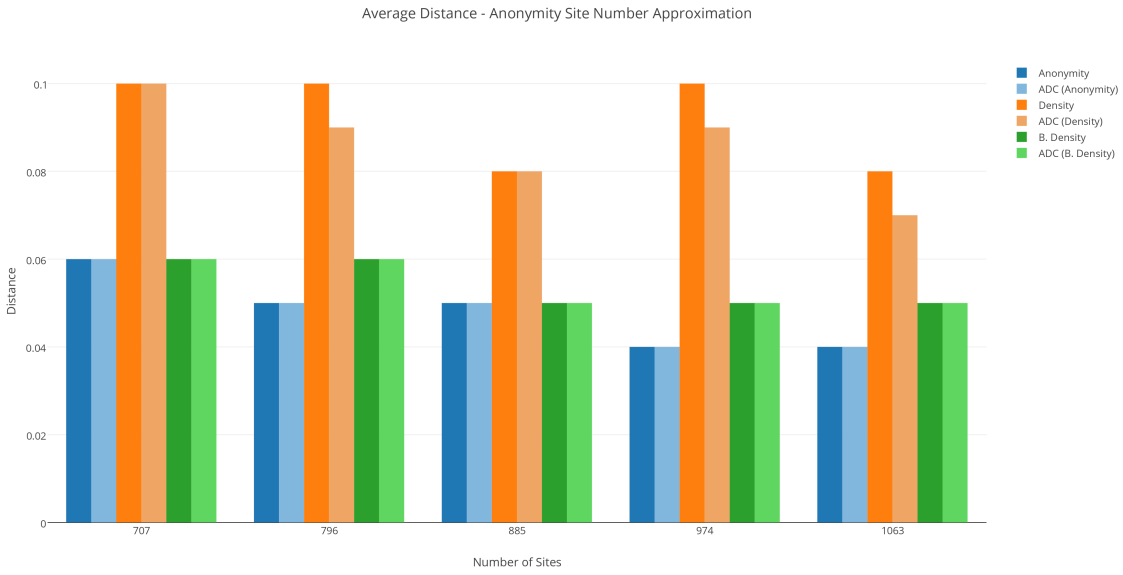


Figure 22: Anonymity - Precision Loss Comparison

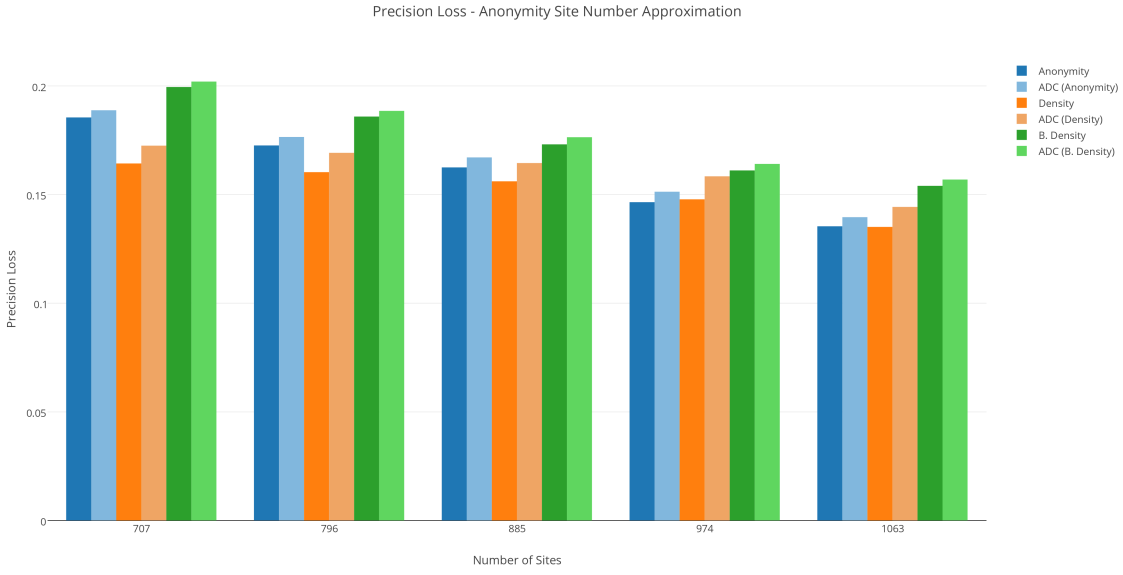


Figure 23: Anonymity - Discernibility Comparison

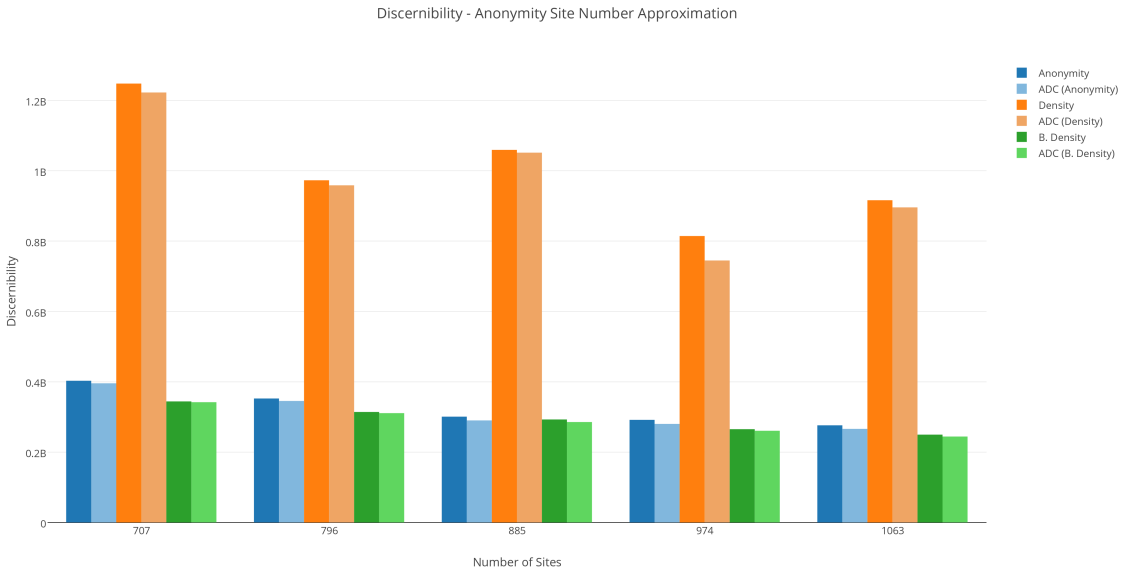


Figure 24: Anonymity - Non-Uniform Entropy Comparison

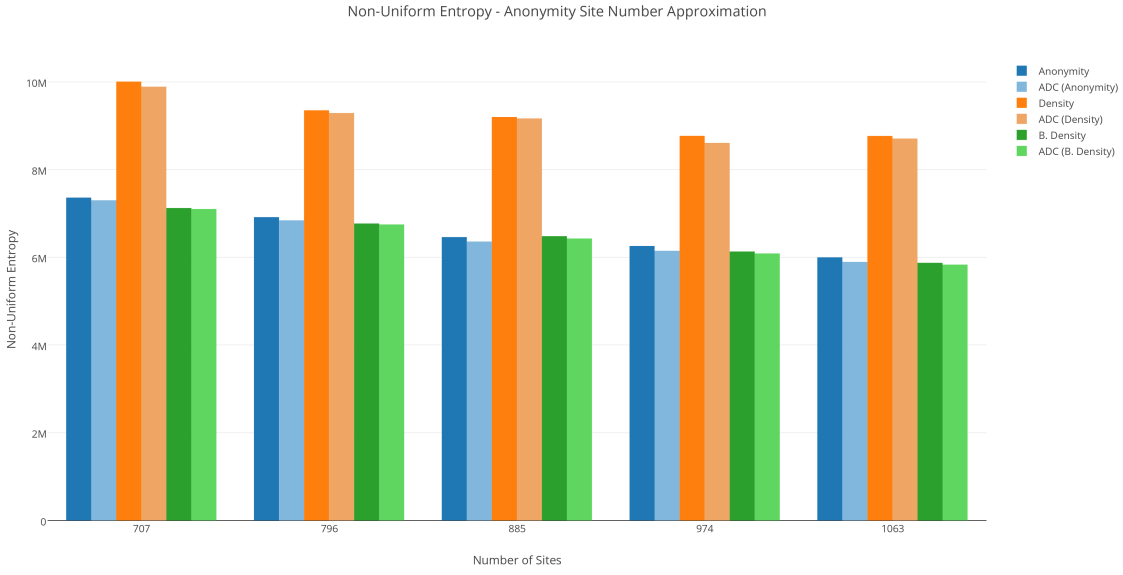


Figure 25: Anonymity - Running Time Comparison

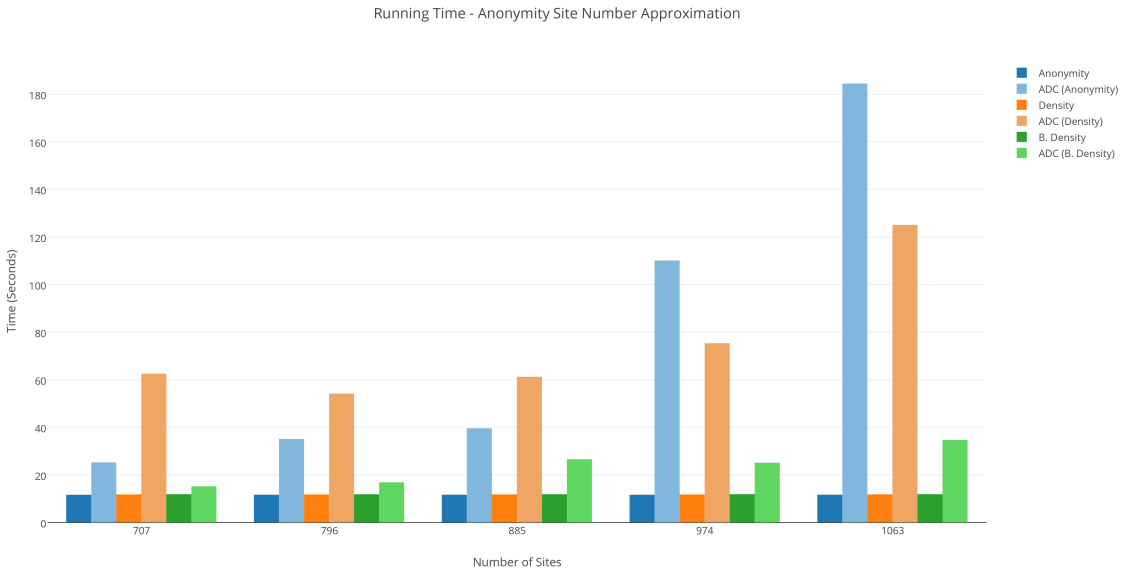


Figure 26: MaxCombs - Suppression Comparison

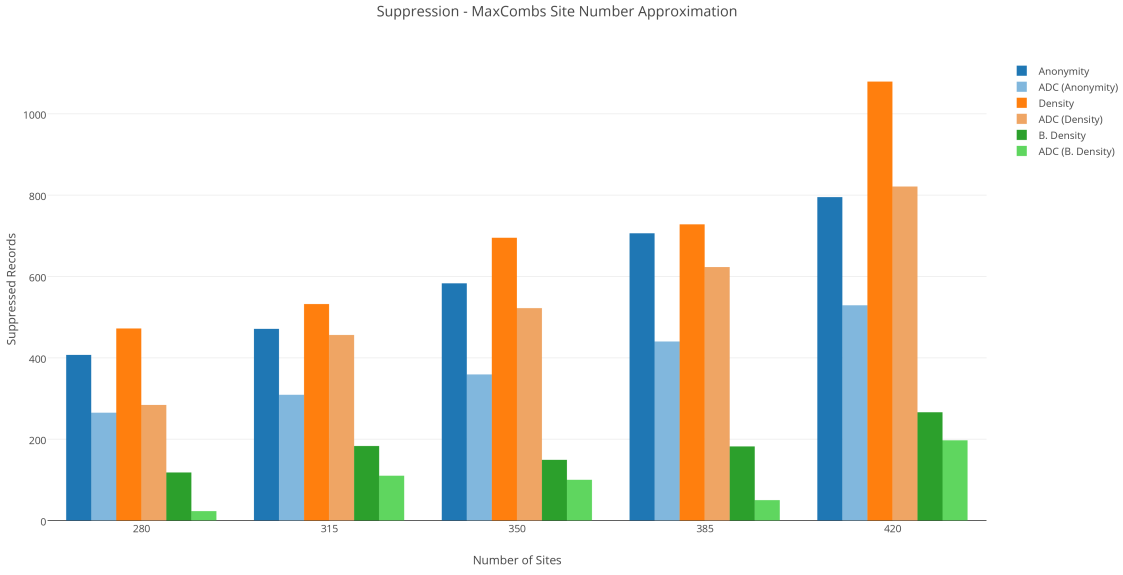


Figure 27: MaxCombs - Average Distance Comparison

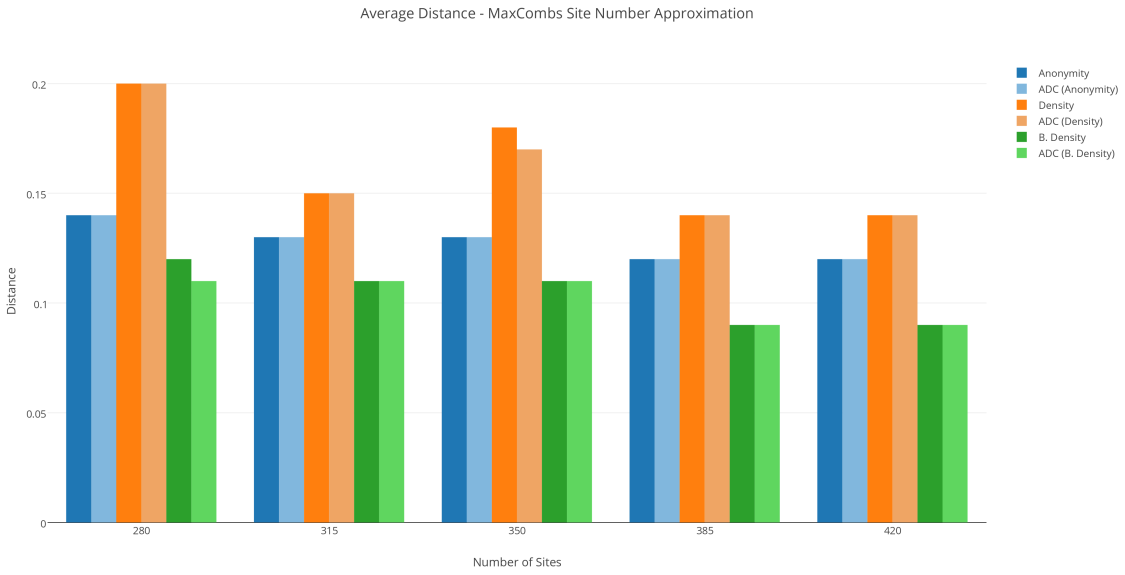


Figure 28: MaxCombs - Precision Loss Comparison

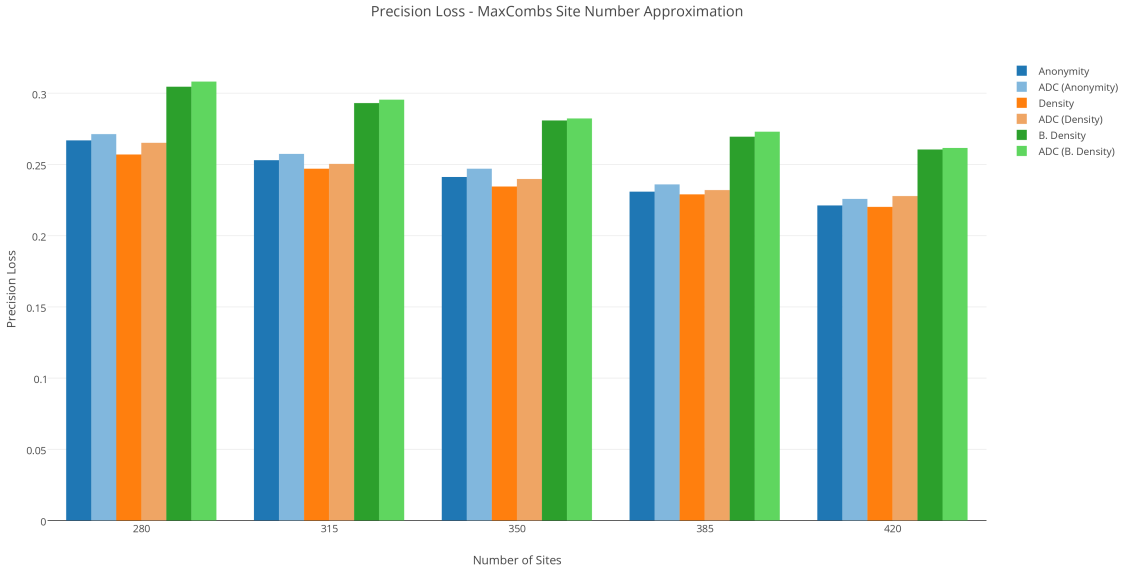


Figure 29: MaxCombs - Discernibility Comparison

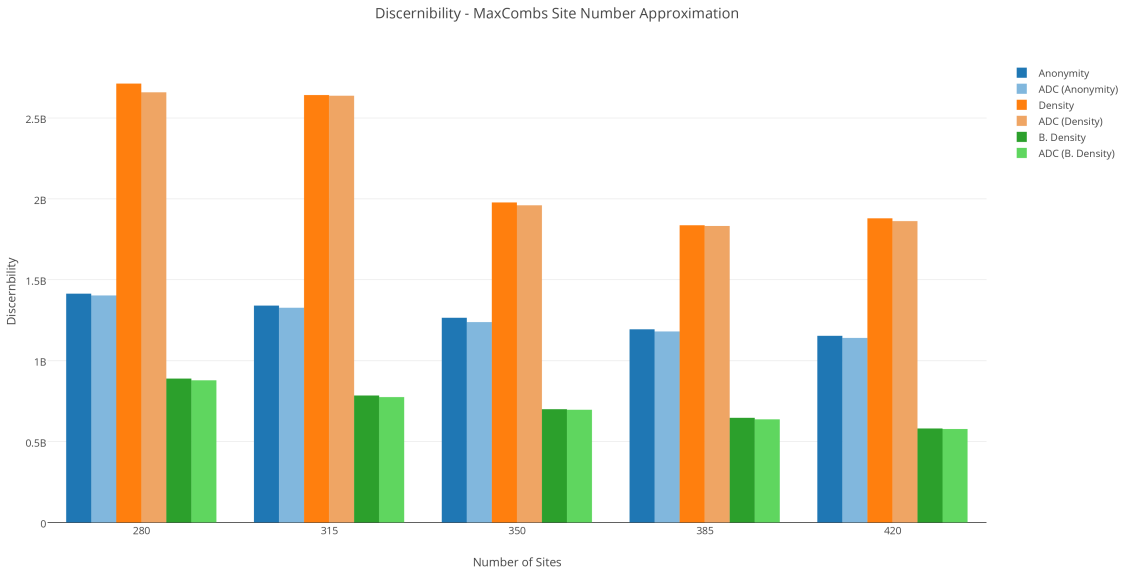


Figure 30: MaxCombs - Non-Uniform Entropy Comparison

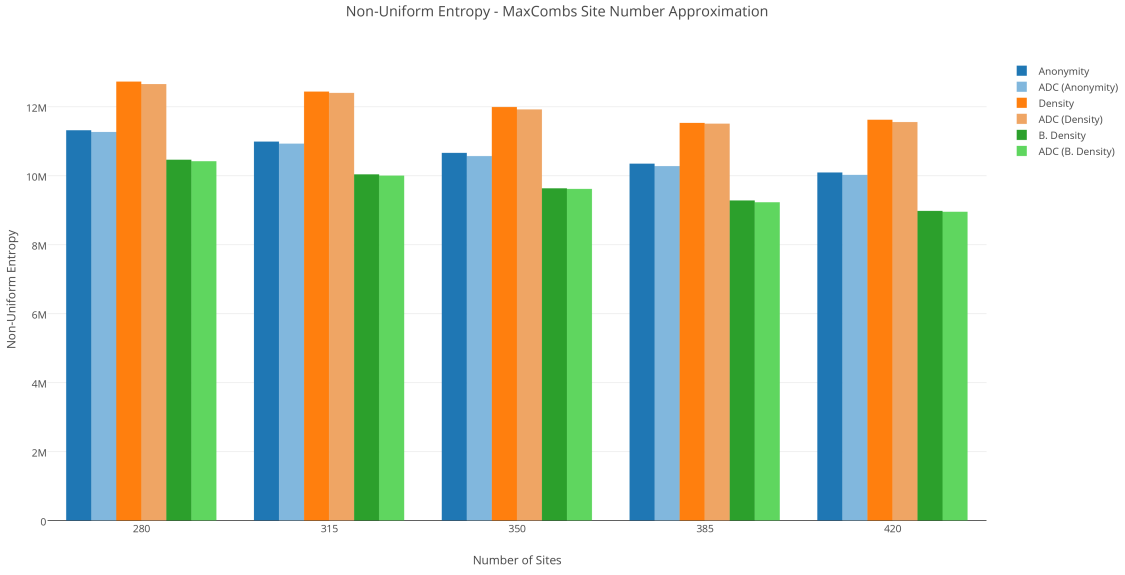


Figure 31: MaxCombs - Running Time Comparison

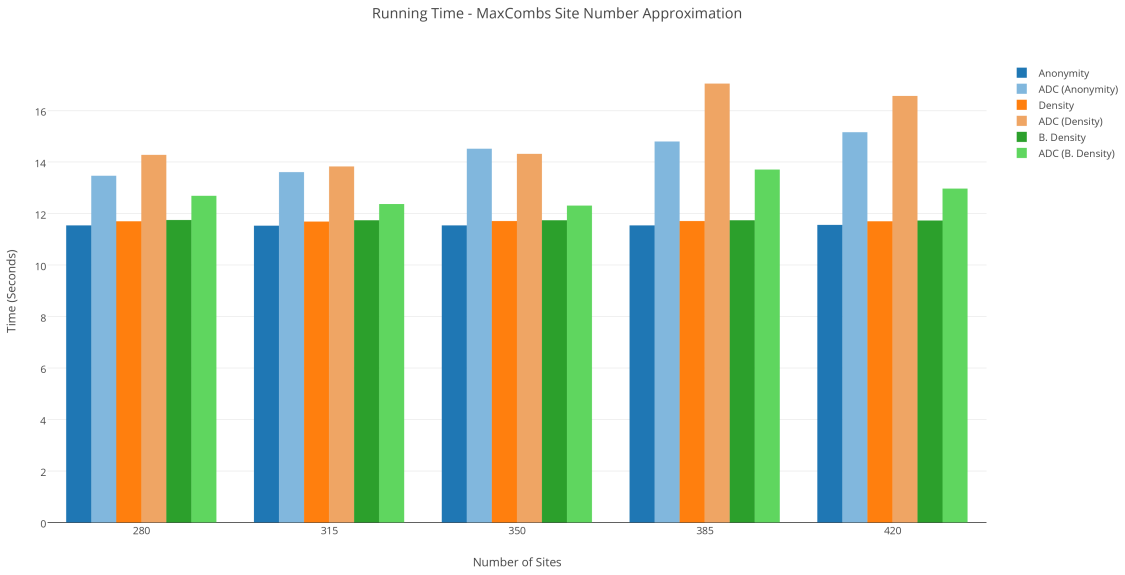


Figure 32: Entropy - Suppression Comparison

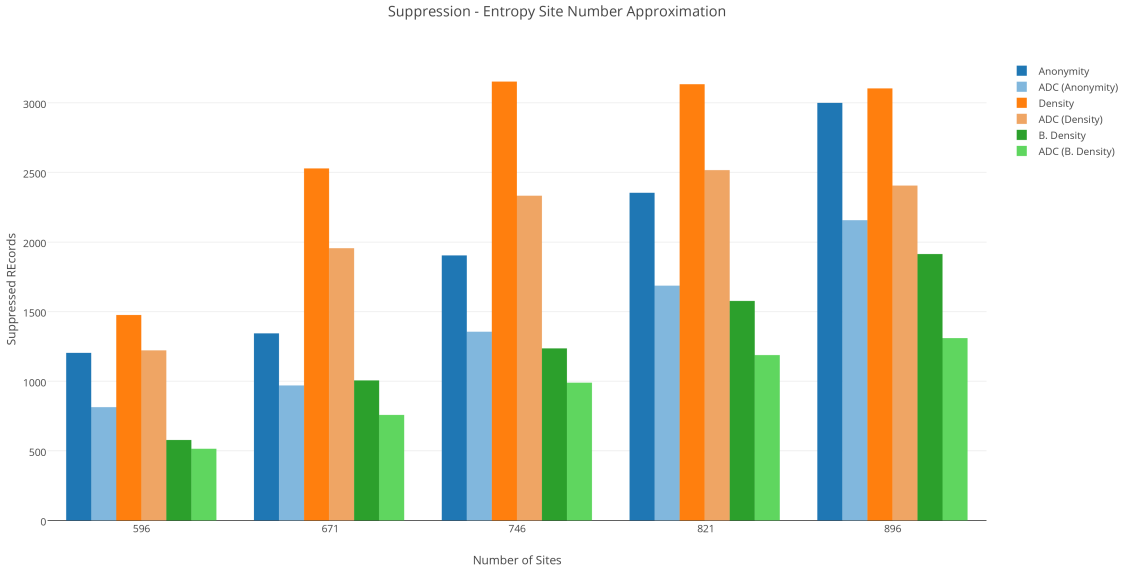


Figure 33: Entropy - Average Distance Comparison

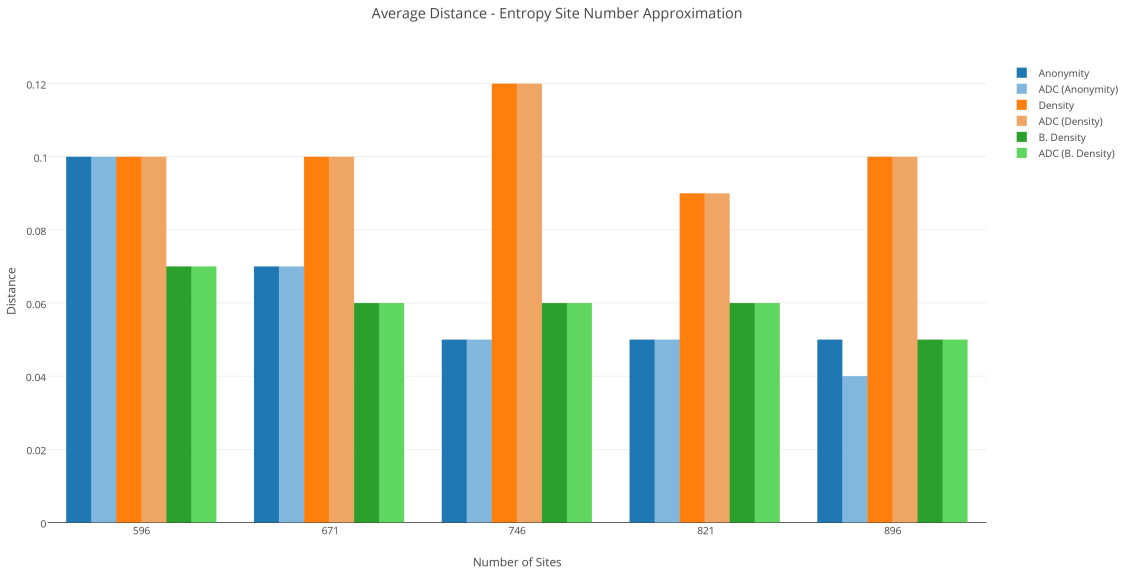


Figure 34: Entropy - Precision Loss Comparison

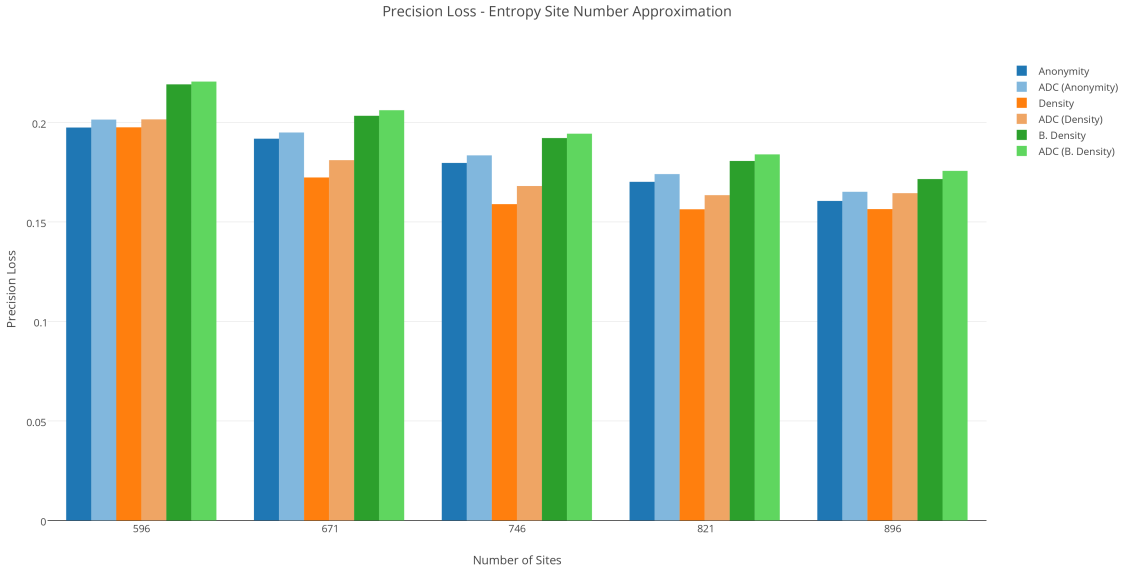


Figure 35: Entropy - Discernibility Comparison

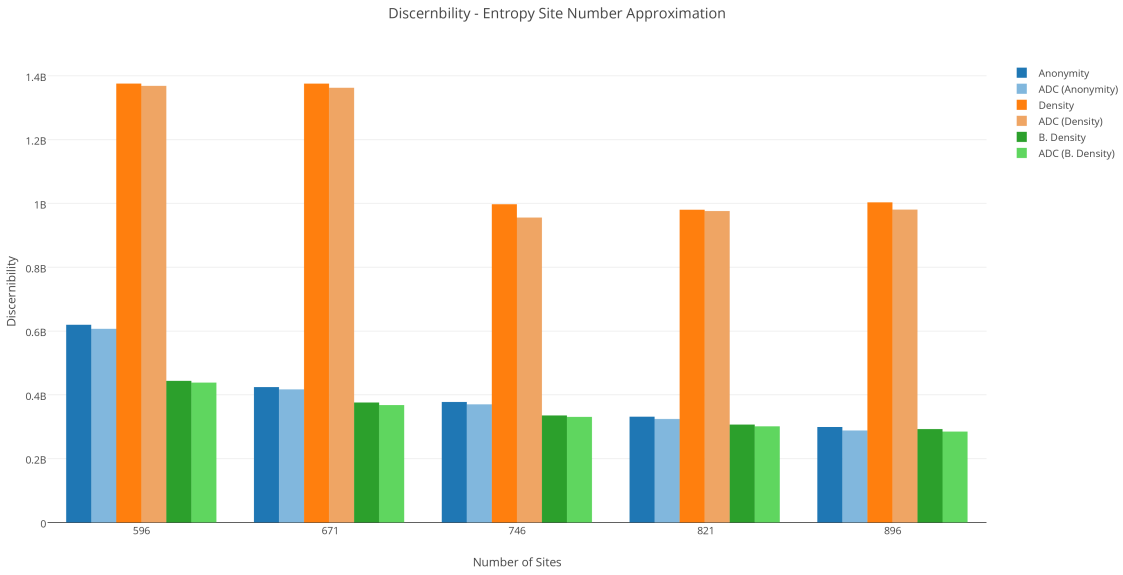


Figure 36: Entropy - Non-Uniform Entropy Comparison

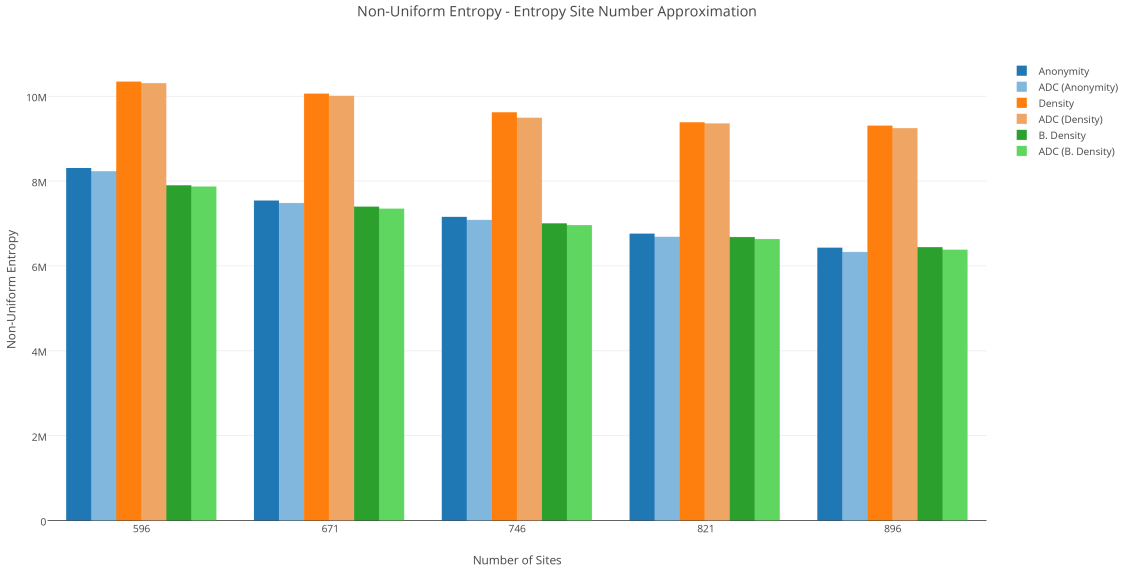
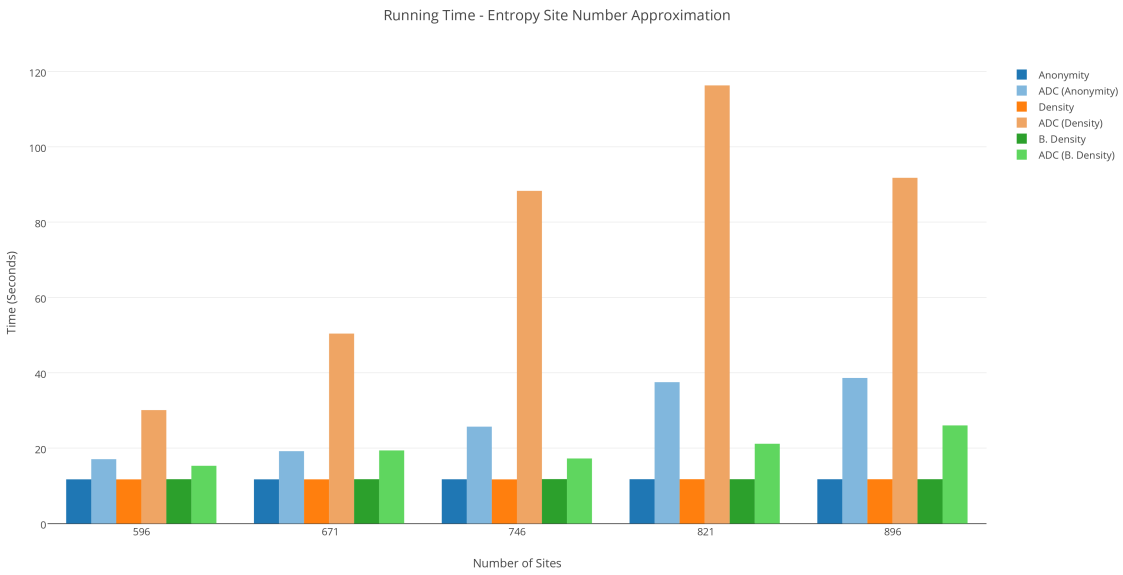


Figure 37: Entropy - Running Time Comparison



Appendix C

Aggregation Screenshots

Figure 38: Eastern Data Set - Age, Marital Status - Anonymity / Anonymity

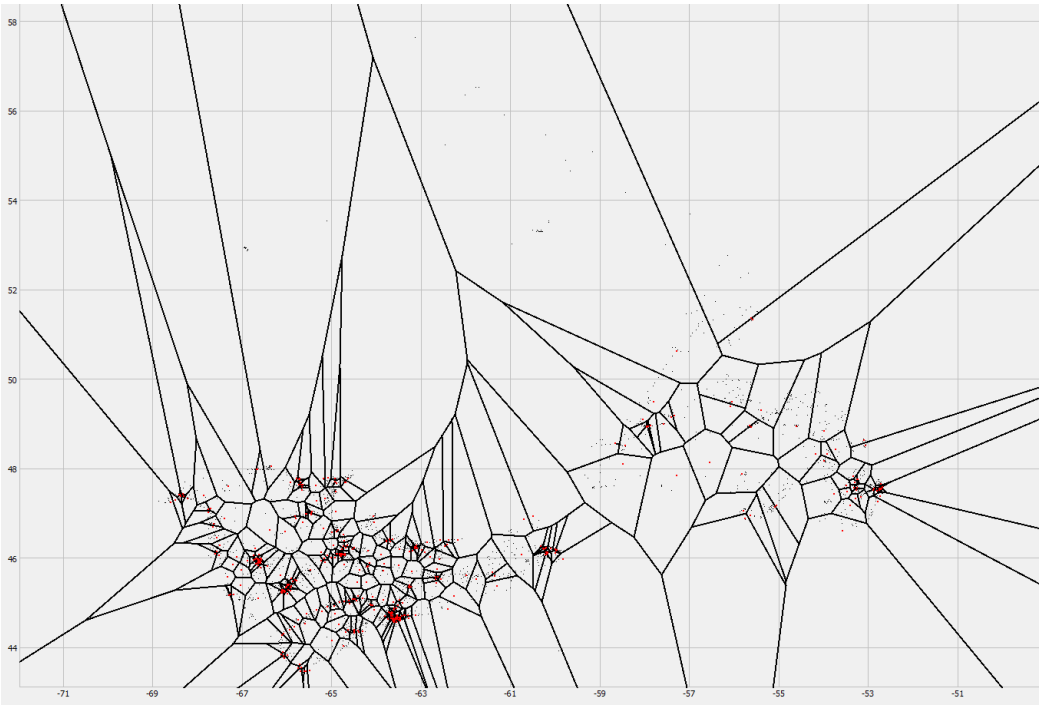


Figure 39: Eastern Data Set - Age, Marital Status - Anonymity / Density

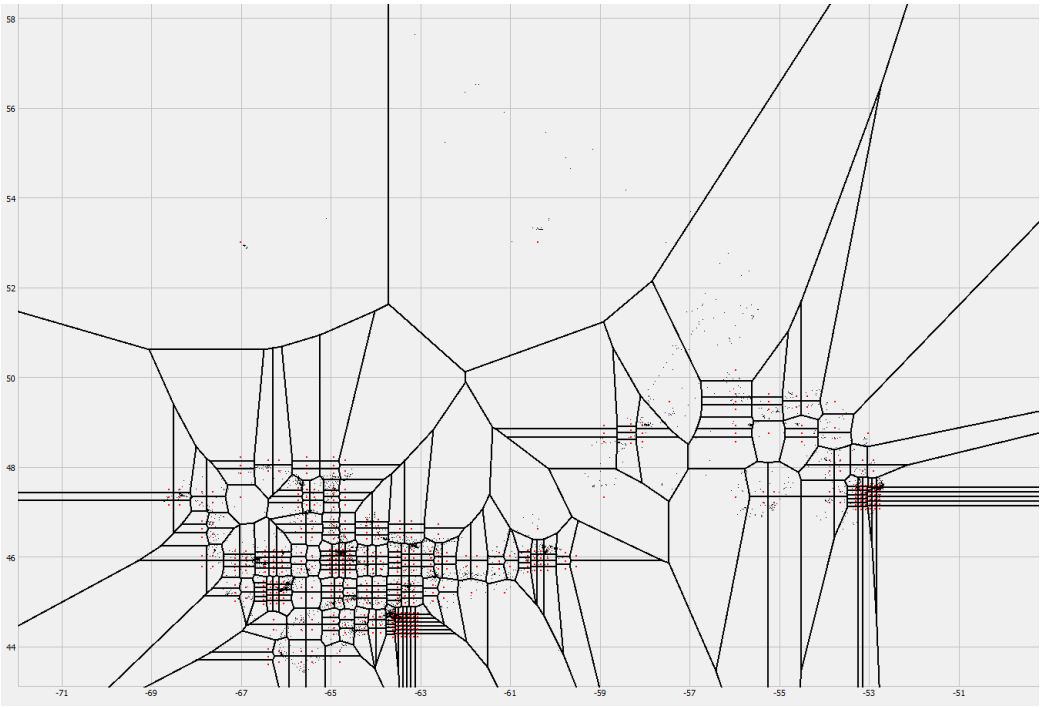


Figure 40: Eastern Data Set - Age, Marital Status - Anonymity / B. Density

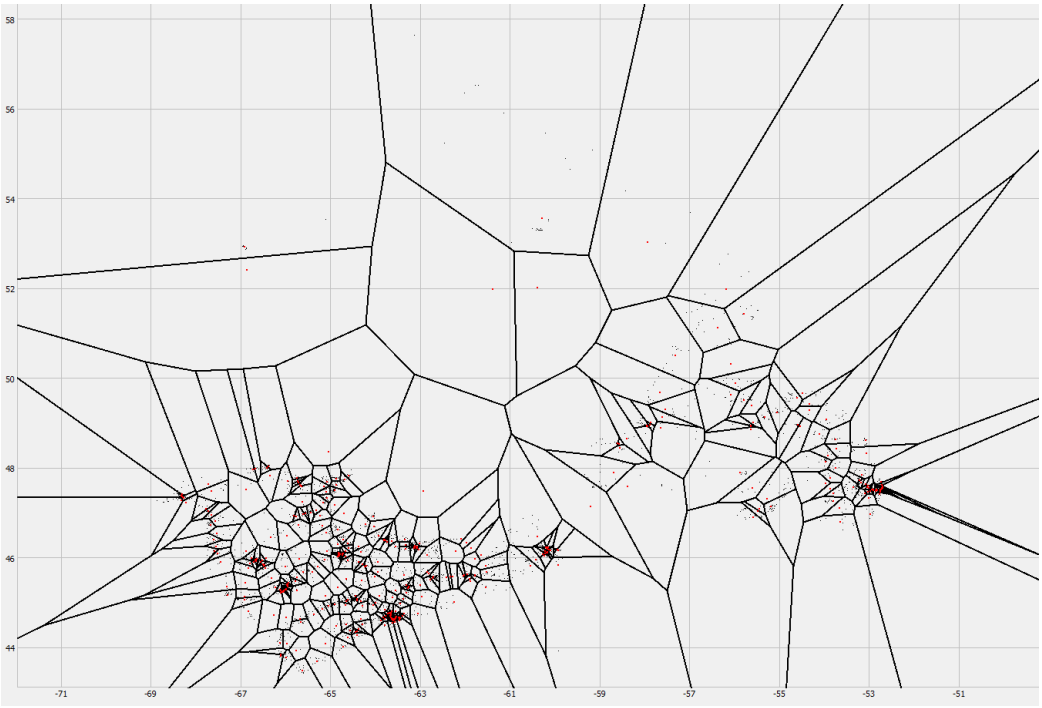


Figure 41: Western Data Set - Age, Gender, Income - MaxCombs / B. Density

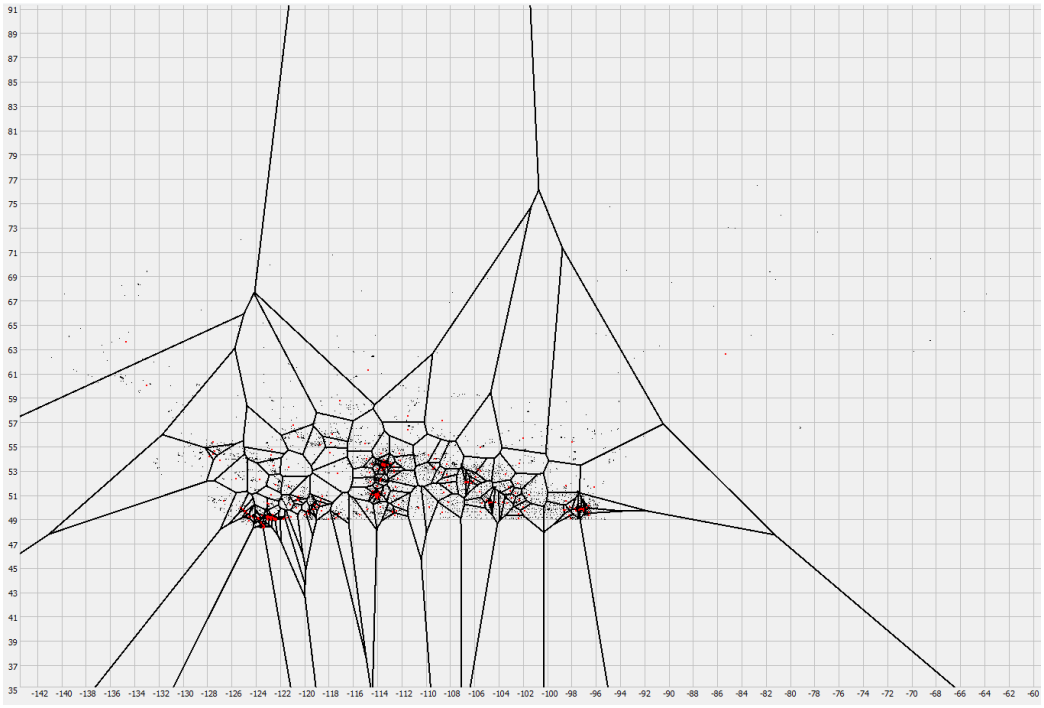
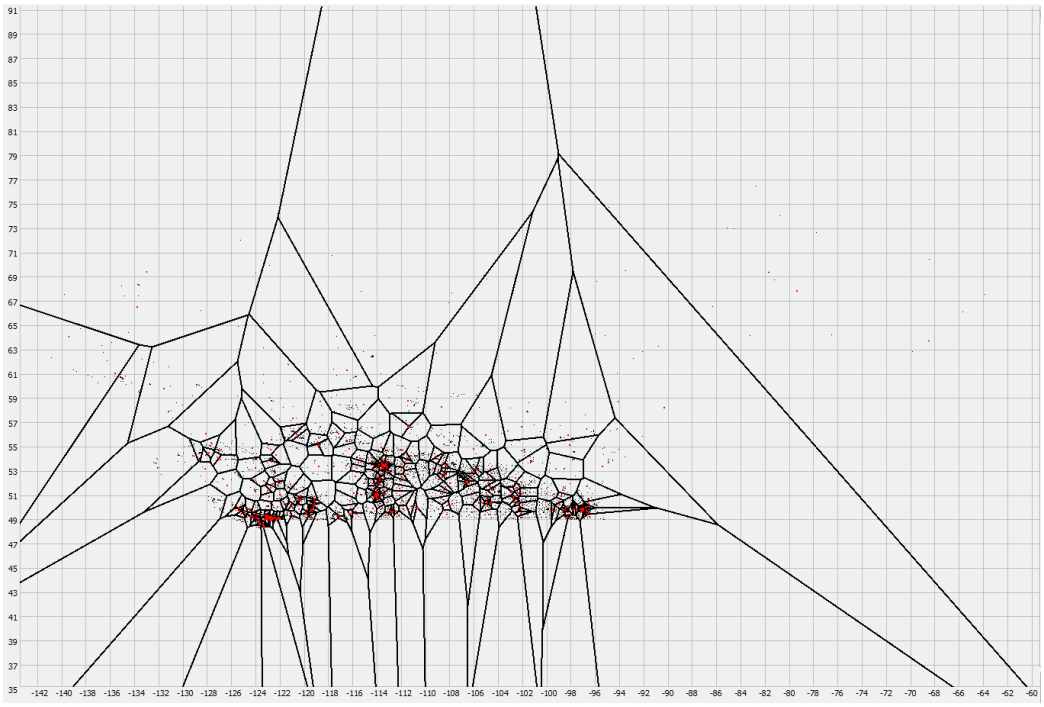


Figure 42: Western Data Set Generalized - Age, Gender, Income - MaxCombs / B. Density



Appendix D

GeoLeader Test Results

Table 19: Prince Edward Island

Scenario	Sites	Suppression	Avg. Dist.	Alt. Avg. Dist.	Discernibility	Entropy	Time
Age, Gender	13	0	0.073	0.016	51189300	727160	10.27
Age, Income	6	624	0.166	0.024	20049500	897295	10.43
Highest Degree, Gender	14	6	0.110	0.016	126210000	716632	10.18
Highest Degree, Marital Status, Gender	8	419	0.207	0.023	79777700	840640	10.34
Age, Gender, Income	4	1176	0.164	0.025	15058400	980723	10.17
Age, Income, Marital Status	5	11072	0.111	0.025	8329780	875663	10.46
Highest Degree, Marital Status, Religion, Gender	5	5742	0.111	0.025	29633300	904632	10.41
Highest Degree, Marital Status, Gender, Income	3	7460	0.068	0.029	18252400	1004700	10.01
Age, Gender, Marital Status, Religion	3	5672	0.098	0.029	20249300	1015910	10.19

Table 20: Prince Edward Island Generalized

Scenario	Sites	Suppression	Avg. Dist.	Alt. Avg. Dist.	Discernibility	Entropy	Time
Age, Gender	17	0	0.078	0.016	82462800	675939	10.05
Age, Income	11	7	0.116	0.016	42910900	761460	9.93
Highest Degree, Gender	19	0	0.040	0.013	115496000	654153	10.13
Highest Degree, Marital Status, Gender	11	85	0.102	0.016	63888200	765664	10.80
Age, Gender, Income	8	25	0.207	0.023	33040800	842553	10.69
Age, Income, Marital Status	6	1068	0.166	0.024	28848100	894739	10.23
Highest Degree, Marital Status, Religion, Gender	5	3195	0.117	0.024	33864100	915866	10.37
Highest Degree, Marital Status, Gender, Income	5	1593	0.229	0.026	26070000	928935	11.20
Age, Gender, Marital Status, Religion	4	3642	0.158	0.027	27554000	958940	10.07

Appendix E

Implementation Comparison Results

Table 21: Prince Edward Island

Scenario	Sites	Suppression	Avg. Dist.	Alt. Avg. Dist.	Discernibility	Entropy	Time
Age, Gender							
B. Density	13	0	0.015	0.014	52253582	730174	0.86
ADC (B. Density)	13	0	0.015	0.014	52253582	730174	0.95
Age, Income							
B. Density	6	587	0.020	0.019	19582477	893750	0.84
ADC (B. Density)	6	587	0.020	0.019	19582477	893750	1.88
Highest Degree, Gender							
B. Density	14	13	0.015	0.014	125434625	717030	0.82
ADC (B. Density)	14	10	0.017	0.014	125791110	717459	1.09
Highest Degree, Marital Status, Gender							
B. Density	8	428	0.021	0.019	77308080	838377	0.88
ADC (B. Density)	8	444	0.018	0.018	76597460	837325	2.92
Age, Gender, Income							
B. Density	4	1205	0.030	0.025	15852885	990625	0.90
ADC (B. Density)	4	1148	0.026	0.023	15093464	981182	3.38
Age, Income, Marital Status							
B. Density	5	10874	0.026	0.022	8566736	881564	0.90
ADC (B. Density)	5	6305	0.224	0.100	15484703	1016857	166.80
Highest Degree, Marital Status, Religion, Gender							
B. Density	5	5847	0.023	0.022	29652893	904255	0.95
ADC (B. Density)	5	5646	0.023	0.023	31809102	912726	66.91
Highest Degree, Marital Status, Gender, Income							
B. Density	3	7437	0.033	0.028	23141741	1045082	0.96
ADC (B. Density)	3	7437	0.033	0.028	23141741	1045082	1.03
Age, Gender, Marital Status, Religion							
B. Density	3	5783	0.054	0.026	18860913	1005247	0.96
ADC (B. Density)	3	5783	0.054	0.026	18860913	1005247	1.01

Table 22: Prince Edward Island Generalized

Scenario	Sites	Suppression	Avg. Dist.	Alt. Avg. Dist.	Discernibility	Entropy	Time
Age, Gender							
B. Density	17	0	0.014	0.012	83133580	677987	0.81
ADC (B. Density)	17	0	0.014	0.012	83133580	677987	0.85
Age, Income							
B. Density	11	12	0.018	0.017	46492510	770432	0.83
ADC (B. Density)	11	12	0.018	0.017	46492510	770432	0.99
Highest Degree, Gender							
B. Density	19	0	0.013	0.011	110716182	650083	0.81
ADC (B. Density)	19	0	0.013	0.011	110716182	650083	0.85
Highest Degree, Marital Status, Gender							
B. Density	11	115	0.018	0.017	66388977	769845	0.85
ADC (B. Density)	11	110	0.018	0.017	65414194	768107	0.94
Age, Gender, Income							
B. Density	8	34	0.018	0.018	32083496	840557	0.84
ADC (B. Density)	8	34	0.018	0.018	32083496	840557	0.94
Age, Income, Marital Status							
B. Density	6	1094	0.020	0.019	28153152	890639	0.85
ADC (B. Density)	6	1118	0.072	0.022	28220382	891992	8.35
Highest Degree, Marital Status, Religion, Gender							
B. Density	5	3290	0.023	0.022	35024612	919411	0.93
ADC (B. Density)	5	3234	0.071	0.024	37205614	926674	38.84
Highest Degree, Marital Status, Gender, Income							
B. Density	5	1571	0.023	0.022	25936575	929568	0.89
ADC (B. Density)	5	1733	0.024	0.023	28571503	938960	11.45
Age, Gender, Marital Status, Religion							
B. Density	4	3725	0.030	0.025	30814037	974701	0.91
ADC (B. Density)	4	3718	0.027	0.023	29722922	965692	7.39