Spatial Analysis of Privacy Measured Through Individual Uniqueness Based on Simple U.S. Demographics Data

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yilun Lin

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Dr. Francis Harvey

May, 2015

Acknowledgements

I want to thank my committee members who were more than generous with their expertise and precious time. Special thanks to Dr. Francis Harvey, my advisor and committee chair for his tremendous help on starting my academic career also his countless hours of reflecting, reading, encouraging, and most of all patience throughout the entire process. I also want to thank Dr. Steven Manson and Dr. Mark Lindberg for their service on my committee. Dr. Manson provided insightful suggestions to the spatial analysis part of the thesis and helped me connect with people with correct expertise. Dr. Lindberg is always my helper on GIS algorithmic and technical questions. Visiting his office and talking to him brings me much joy in addition to the knowledge.

I would like to thank my fellow graduate students friends and colleagues in the Department of Geography and the U-Spatial. Discussions with Melinda Kernik, Dudley Bonsal, Bryan Runck, Brittany Krzyzanowski, Maria Bakhtiyarova and other people from Human Environment Geographic Information Science Lab provided invaluable feedbacks on my research and presentation. Taylor Long, Agata Miszczyk, Yiqun Xie and Ben Liang from U-Spatial took off a lot of my work duty so that I could focus on the thesis writing.

Dedication

I dedicate my thesis work to my loving parents, Junxing Lin and Qing Yang, who have given me unusual freedom and supports on developing my own career and life path.

I also dedicate this thesis to my church family who have mentally supported me throughout the process

Last, I dedicate this thesis to my Lord Jesus Christ, who not only sought me and found me but also promised and provided me with daily mercy and strength.

Abstract

Previous studies reveal that, using U.S. census data, over 60% population of the U.S. could be uniquely identified with a combination of gender, zip code, date of birth attributes in 1990 and 2000. This thesis extends these studies to examine spatial variation of individual uniqueness in 2010 at different scales and regions in the U.S. In this thesis, I use spatial and non-spatial statistics to study the spatial patterns on both global and local scales. Specifically, I provide 1) the comparison of national level uniqueness between 2000 and 2010, 2) the investigation of spatial variation of uniqueness in different regions and at different scales, 3) the identification of local uniqueness clusters outliers and 4) the evaluation of urban-rural divides on individual uniqueness segregation.

On the global scale, the comparison between 2000 and 2010 reveals that, although overall individual uniqueness changes little, the individual uniqueness of middle-age group members has significantly decreased. The study of regional differences finds that low individual uniqueness for college-age population are spatially homogeneous despite that the overall uniqueness are spatially heterogeneous. The analysis at different scales discloses that overall uniqueness decreases, and the differences between age-group uniqueness reduce, when geographical scales focus on the cores of urban area.

On the local scale, the results indicate an urban-rural divides of individual uniqueness segregation. The Clusters and Outliers Analysis find that places where low individual uniqueness cluster the most are also very urbanized area. The average individual uniqueness of urban area is computed as 58.02% whereas that of rural area is computed as 88.43%. This means, if a person is from an urban area, given the zip code, gender and date of birth information, he/she is much less likely to be identified uniquely.

This study offers contributions to geographic information privacy, particularly relevant to reverse geocoding and related spatial aggregation techniques used in census data.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In the information society, personal level locational data are continuously created and stored. On one hand, individuals' physical coordinates are continuously captured voluntarily or involuntarily through the use of digital devices such as mobile phones, car navigation systems and laptop computers. For instance, people frequently send their current locations to location-base services providers to get restaurant recommendations and routing suggestions. Self-reported geotagged Tweets, location field in Facebook profiles, Place Pins on Pinterest shows the past, present and future of their locations. On the other hand, locational references are also easily added to traditionally non-spatial personal data. With geocoding techniques, hospital visits, voting registration records, census data, and credit cards transactions can be easily mapped, adding to the spatial trajectory of each individual.

The abundance of personal level locational data triggers tremendous societal concerns on privacy issues. In 2011, Apple's covert action of recording users' locations in hidden files on iPhones was revealed and caused wide controversy [5]. The similar action was soon found with Google on Android platform [6]. The concerns about locational privacy started to grow fast and ultimately received international attentions through a 2012 TED Talk entitled "Your phone company is watching". In this talk, German politician Malte Spitz challenged European Unions Data Retention Directives by showing how the details of his life trajectory in the past six month can be re-constructed

through the location information stored by his telecommunications provider [7].

Many concerns about locational privacy have been placed on the GPS tracking data as mentioned above, but geo-referenced traditional personal information has received less attention. The location information in these traditional data is not explicitly collected in the form of GPS coordinates, but usually recorded in the form of plain text such as ZIP codes and census tract FIPS. However, the text data can act as a proxy to geospatial entities in order to geo-register the piece of personal information. Given its long history of usage, this data can provide longitudinal information about individual.

In this thesis, I am primarily concerned with personal privacy in the traditional text data with locational references, particularly census data. In census data, questions about individual privacy is a question about anonymity. In other words, people are concerned about being identified uniquely from the geographically aggregated census data. Two prior studies that examined the unique identification of individual in census data informed my research. The risk of unique identification of individual has been assessed based on 1990 [8] census data and 2000 census data [1]. These studies looked at the individual uniqueness in simple U.S. demographics (ZIP code, gender, date of birth) and concluded that 80% of individuals can be uniquely identified in 1990 and 62% in 2000 using the combination of the three demographic attributes above. These results are significant, but they neglected the effect of geospatial variations. In their research, they simply assumed that the computed national average individual uniqueness is the same for everywhere in the country. However, different ZIP codes reflect different geographies and thus are subject to well-known spatial effects including spatial dependence and spatial heterogeneity. As a result, the spatial questions such as how individual uniqueness varies spatially and if there exists any particular spatial clusters of high low uniqueness remains unanwered in their works.

This thesis helps fill in this gap. It uses widely used statistical and spatial statistical techniques to assess the anonymity of a group size k, or k-anonymity in various regions and age groups. Specifically, 2010 census data are analyzed to see how unique individuals and groups of individual are in the dataset and if this uniqueness is spatially significant, related, and homogenous.

1.2 Research Questions

In this thesis, the following research questions (RQs) are examined:

- RQ 1: does the individual uniqueness in simple demographics vary spatially across the U.S.?
  - RQ 1.1: regional differences - does the uniqueness distribution present regional differences? To what extent does it vary spatially?
  - RQ 1.2: scale differences - within a region (e.g. state), does the uniqueness vary across different scales of cesus geography?
- RQ 2: what significant spatial distribution patterns can be observed?
  - RQ 2.1: where are the regions with high uniqueness and low uniqueness? Are they spatially clustered?
  - RQ 2.2: what are the socio-economics characteristics of those clusters? Why do they become the high/low uniqueness clusters?

1.3 The Structure of the Thesis

This thesis explores the spatial variations of the privacy measured through individual uniqueness using U.S. census data. The thesis covers diverse concepts and techniques from privacy research, geo-privacy studies and spatial statistics, I first introduce the relevant theories from the three fields of studies and connect them to the topic. These concepts and techniques are then applied in the empirical data analyses on census data.

Specifically, the thesis is organized as follows:

In Chapter 2, I provide the theoretical background, explain the central concepts and relate them to the census data privacy issue. I explain the privacy and privacy protection techniques in general and their theories and applications in geographic information domain. In addition, I introduce the relevant spatial statistics techniques that is used to assess the spatial patterns of individual uniqueness.

In Chapter 3 and Chapter 4, I detail how I answer the Research Questions proposed in previous section. Chapter 3 focuses on answering RQ1, which explores the

existence of regional differences and scale differences in individual uniqueness distribution. Chapter 4 focuses on answering RQ2, which characterizes the specific spatial patterns of individual uniqueness. Dataset, methodology, statistical techniques, results and interpretations are detailed in each chapter.

In Chapter 5, I offer the conclusion of this study. I summarize the results and interpretations from Chapter 3 and 4, and suggest future research directions.

Chapter 2

Background and Central Concepts

In this section, background and central concepts related to this thesis are introduced. First, I contextualize this research against the broader themes found in privacy studies. Second, I survey research on geo-privacy issues. Third, I introduce the spatial statistical methods that are carried out in this thesis.

2.1 Privacy protection and uniqueness

Although this thesis uses uniqueness as a proxy to personal privacy, it's important to provide some background about the development of the idea and techniques of privacy protection in general to help us understand the strengths and weaknesses of using anonymity as a proxy for privacy.

2.1.1 Privacy - A Complicated Concept

A widely cited remark about the challenging concept of privacy states, "privacy is a concept in disarray" [9]. When Samuel Warren and Louis Brandeis first advocated the right to privacy in legal framework in 1890, they might not have envisioned that it has developed into such a complicated idea. Starting from its inception, numerous attempts have been made to exactly define privacy. In the initial definition, Warren and Brandeis

define it as "the right to be left alone" [10]. Not satisfied with its vagueness, philosopher Sissela Bok states that the privacy is "the condition of being protected from unwanted access by others" [11]. Richard Posner concretizes the "unwanted access by others" to be "the right to conceal discredible facts about himself" [12]. Yet quite opposite, Charles Fried [13] argues that "privacy is not simply an absence of information about us in the mind of others; rather it is the control we have over information about ourselves". Contending that all the above definitions are too narrow, some privacy theorists explain privacy as a protection of the integrity of personality, which includes "individuality, dignity, autonomy and intimacy" [14]. Other privacy scholars understand privacy in the context of personal relationship development as a form of intimacy. Even only looking at a small array of definitions, it is easy to realize that the definitions of privacy vary greatly in scope and emphasis.

One reason that leads to the complexity of privacy is that when privacy is most frequently discussed in legal and policy cases, people usually seek actionable definitions rather than the actual meaning in that case [15]. This results in various case-based definitions to a point that legal privacy consists of four or five different species of legal rights which are quite distinct from each other and thus incapable of single definition [16]. In a seminal article, Solove argues that judicial opinions and statutes often depend upon some notions of the definition and value of privacy [14] but not all of them. For example, he mentions the different legal definitions of privacy between Fourth Amendment and a tort of intrusion upon seclusion: the former depends on if person has reasonable expectation of privacy and the latter is only applied to matter concerning the private life. The prominent problem here is that the definition must satisfy the scope of the law and policy. As a result, the definitions are either too big or too small [15], further perplexing the privacy concepts when the law and policies are applied to daily life.

Another reason for the complexity is that when privacy is discussed in daily discourses, it is often context sensitive. For instance, Nissenbaum argues that privacy is not simply restricting the flow of information, but ensure that it flows appropriately, and an account of appropriate flow is given here through the framework of contextual integrity [17]. She further explains that privacy demands information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it [18]. For example, since the context-relative informational norms in

health care system are different from that in social network profiles, peoples expectation for the proper flow of information are different.

2.1.2 Privacy protection

Despite the vagueness of the true definition of privacy, researchers have to operationalized the concept in order to protect it. In practice, researchers usually take one spectrum from the complex definition of privacy and design corresponding protection solutions. In general, these privacy protection approaches can be categorized into 4 types:

- **Privacy as confidentiality** [19]

  In this approach, privacy protection relies on simple assumptions that the exposure of information leads to a loss of privacy. This ensures that only the necessary data are collected from the users, and all traffics and communication of the data are confidential.

- **Privacy as anonymity** [2]

  In contrast with privacy as confidentiality, this approach does not keep the information secret, but instead, makes the information anonymous. This means even if this piece of information is accessible, it cannot be linked to its owner. The simplest way to achieve anonymity is to delete the universal identifying key, such as SSN, name etc. However, it could be seen in the following session how this anonymity requires more sophisticated precautions. **According to this definition, the concept of individual uniqueness falls into this category.**

- **Privacy as control** [19]

  The presumption of this approach is that the individual level data have to be revealed to the public with linkage to the owner. Most protection techniques of this kind focus on helping users understand when, how and what extent information will be revealed and make informed decisions about privacy control themselves. This approach deliberately involves users in the privacy protection process.

- **Privacy as dynamics**

This approach, instead of treating privacy protection as a deterministic process, considers it as a subtle and iterative negotiation between the system and the users. It takes into account dynamics such as peer-pressures (e.g. the fact that your friends made their birthdays visible implicitly nudges you to do so) and users opinion (e.g. users different actions after they understand how recommender systems use their data).

2.1.3 Anonymization protection and the breach

As mentioned before, anonymity is one implementation of privacy protection. It deserves further explanation since the idea "uniqueness" used in this thesis falls under this category. Because of its simplicity, the idea of "anonymity" is widely used by government and commercial data holders to generate privacy-compliant data released. To anonymize the dataset, the most straightforward way is to delete sensitive attributes in the dataset. Here, the sensitive attributes refers to all explicit identifiers such as name, address and telephone number [2]. For example, considering a table with records of hospital visits, which includes the patients name, contact number, birth dates, gender, zip code where the patient lives, and the illness/disease (see Fig. 2.1). Assuming most patients have distinct names and contact number, the possibility to uniquely locating a patient in the table is thus high. A straightforward anonymization protection would be to drop the name attribute and contact number attribute. The rest of the attributes are kept because 1) in a big table, its assumed that there are a lot of people sharing the same values of these attributes; 2) they are needed for research purposes. In addition, because there is no unique key after the anonymization, it also further prevents the privacy attacks that try to join this table with other table in order to re-identify the sensitive attributes.

However, the above simple anonymization protection is easily breached by linking anonymized dataset to other knowledge. The term knowledge is used because auxiliary information is not limited to other normative data sources but also including contextual knowledge or background information. Consider the same hospital visits table after simple anonymization, although the sensitive attributes are not provided, they can be re-identified by linking with other data sources that contain these attributes. The re-identification entails that the anonymized dataset shares certain combination of

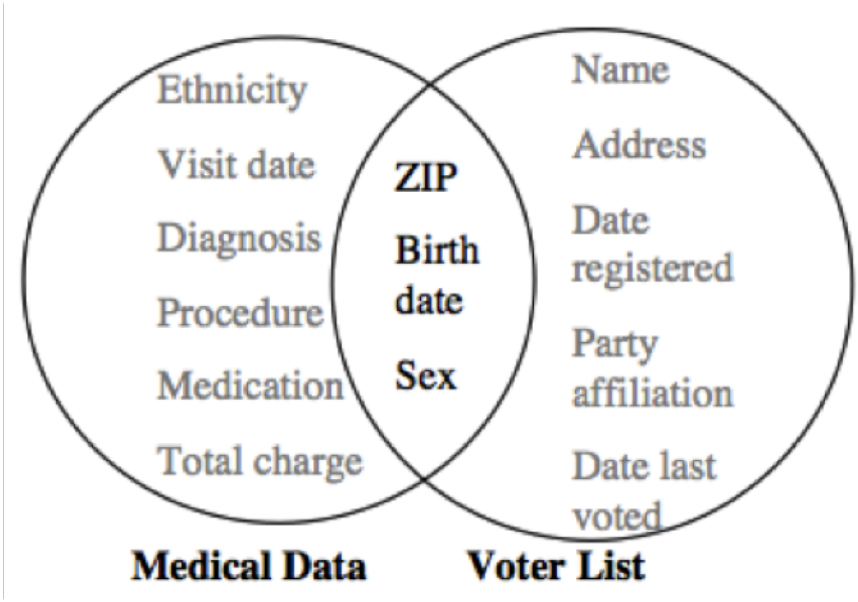

Figure 2.1: Linking to re-identify, excerpted from [2]

attributes with the data containing sensitive attributes that we linked to. One famous example is linking an anonymized hospital visits table with the voter registration table and uniquely identifying Governor of Massachusettss record in the hospital visits table [2]. Fig. 2.1 illustrates how the linkage could be achieved by using a combination of attributes.

2.1.4 Homogenization process: k-anonymity and its derivatives

K-anonymity is a privacy protection measurement invented by Sweeney in her work [2] to measure the uniqueness of attributes combination in the data release. Its purpose is to counteract the above anonymization breach. K-anonymity ensures that each person, represented as a tuple in the release of a dataset, cannot be distinguished from at least other k-1 individuals in the same release. To consider the hospital visits example in the previous section, a k-anonymized release of this dataset would mean that each row, which is a combination of ZIP, birth date and sex, cannot be distinguished from at least k-1 other rows. When the k-anonymized hospital visit dataset is linked to the

voter registration dataset, the sensitive attributes cannot be re-connected to the identity since there are multiple records in the hospital visits dataset that can be linked to the same record in the voter dataset.

In general, k-anonymization uses two techniques: generalization and suppression [20]. Generalization replaces the current values with less specific values but under the same semantics. This typically involves using the the superset of the value to substitute the original value. For example, the zip code 55455 might be generalized to 554**, which includes all zip code prefixed with 554. Another example would be changing nationality from Chinese to Asian. After generalized into superset, the values are more likely to be indistinguishable with values from other records. Suppression, the other k-anonymization technique, is more conservative. It keeps the values completely confidential. Using suppression, the zip code 55455 would be completely excluded from the dataset, or substituted with *. Various algorithms [20, 21, 22, 23] have been invented to achieve k-anonymity.

Numerous researchers questioned if k-anonymity is enough for privacy protection and proposed additional privacy measurements. Machanavajjhala and others [3] observed that k-anonymity fails when 1) the diversity of a sensitive attribute is small and 2) an adversary has background knowledge about the sensitive attribute. Examples can be seen from the Fig. 2.2. Record 9-12 explain situation 1), where (Zip Code, Age, Nationality) are under k-anonymity but the sesitive attribute "Condition" contains only "Cancer". So the adversary would know directly which disease the target gets even after k-anonymization. Records 1-4 explain the situation 2), in which, if the adversary has background knowledge that the target's nationality is Japanese, he can quickly infer the target gets viral infection since it's well-known that Japanese has low rate of heart disease. To tackle this, Machanavajjhala and others proposed the l-diversity that measures the diversity of the non-k-anonymized attributes.

Another improvement to k-anonymity and l-diversity is t-closeness [24]. This privacy measurement argued that k-anonymity and l-diversity are neither sufficient nor necessary since 1) the distribution of the values in the l-diversity attributes might be skew 2) there might be semantic relationships among the values in the sensitive attribute (e.g. all values are diseases about stomach). For the situation 1), the skewness leads to a high probablity that adversary correctly estimate the targets value. For situation 2),

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2.2: Examples for l-diversity, excerpted from [3]

it means that the adversary would gain additional information about the target just by looking at the generalized data release. In [24], Li et al. developed t-closeness that measured the distance t between the distributions of the sensitive attributes and the overall distribution of the whole table. The anonymization process focuses on ensuring the distance t is greater than the threshold.

2.1.5 Existing research of census data privacy

Privacy has been one of the top priorities in the release of the U.S. census data [25]. How privacy is operationalized in census data is based on how census data is organized. In census data, demographic information are aggregated according to various levels of census geographic enumeration areas, ranging from the smallest size - census block to the biggest size- national level. Using this aggregation method, the concern of privacy in census data is de facto the concern for anonymity.

In existing body of research, the anonymity of individual in the aggregated census data is measured by k-anonymity [20, 1]. In the context of census data, k-anonymity

ensures that in census data at least k people share the same demographic characteristics. A k value of 5 means that a group of 5 people with the same demographic characteristics is identifiable from the population. Using this measurement, researches based on the 1990 census and 2000 census reveal that, in the United States, 87% of the population in 1990, and 63% of the population in 2000 can be uniquely identified (i.e. under 1-anonymity) given the simple demographics (gender, zip code, and date of birth) [1, 8]. The practical meaning of these two uniqueness rates is that, the combination of (zip code, gender, and date of birth) can be used as the unique identifier for more than half of the individual in U.S. Considering these three attributes are commonly collected by online account registration, insurance quote, customer service tracking, its really easy to be associated with other dataset for further privacy attack. These studies have influenced privacy protection guidelines for numerous public releases of microdata [26, 25].

However, previous studies of the individual uniqueness fail to take into account the datas spatial nature. Census data are aggregated by geographic enumeration areas at different spatial scales. As a result, the spatially heterogeneous population distributions would impact the distribution of uniqueness. For instance, as shown in the 2010 population distribution map [4] (Fig. 2.3), the population density in the eastern U.S. is prominently higher than that in the western U.S. Furthermore, the state population pyramids also show disparities. As a result, I am interested to examine how privacy in the census data, measured through individual uniqueness, varies spatially.

2.2 Geographic information privacy and protection

2.2.1 Geographic data in privacy protection

Privacy has been a integeral part of geographic information science research agenda [27]. Compared to other privacy protection studies, research in geographic information privacy has to deal with diverse data types which are more complex than just the tabular data. In geography, they data types can be broadly classified into:

- **Textual format**

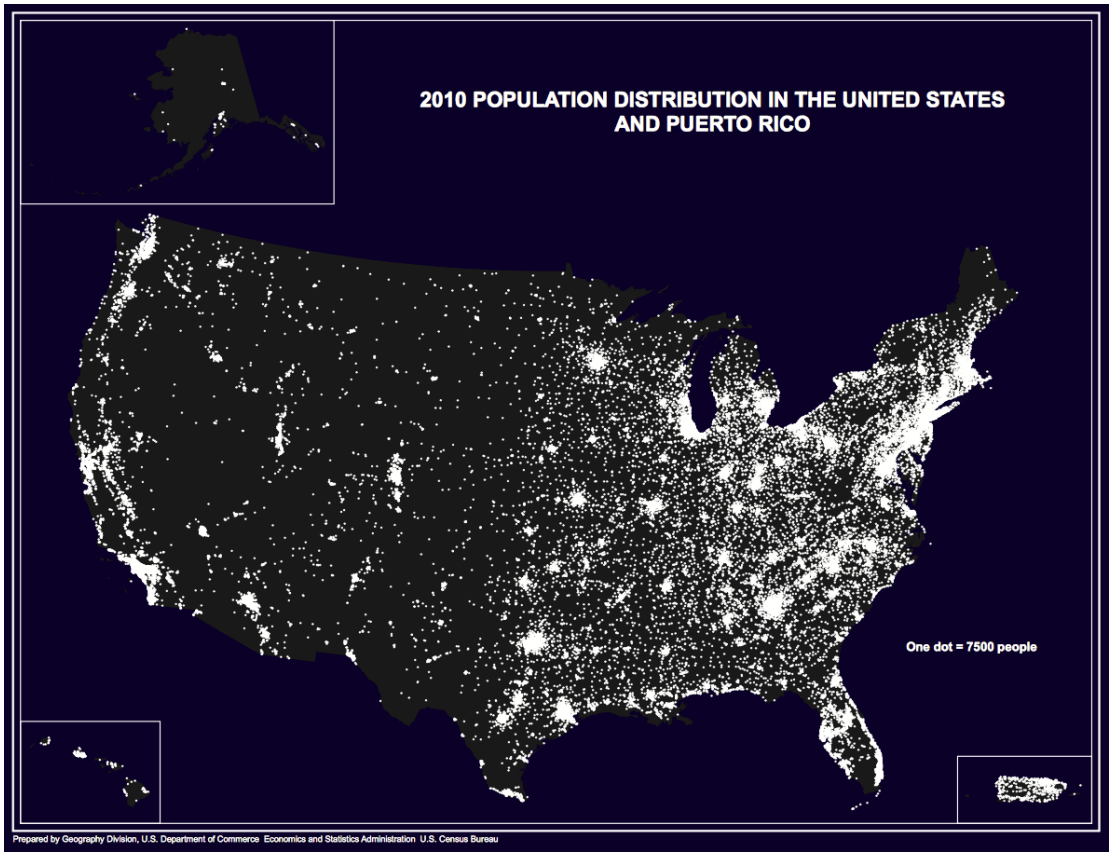

Figure 2.3: 2010 population distribution in the United States and Puerto Rico [4]

Textual format data primarily refers to the descriptive element of geographic information, including the address and the geographic enumeration area. It can also refer to some indirect contextual information such as the service units an individual belongs to. Specifically, census data uses textual geographic information in the forms of various levels of census geography including census block, zip code tabulation area (ZCTAs), county and state.

- **Geometrical format**

  This refers to the geometrical element of geographic information, including the individuals geodetic coordinates and the boundary of the areal object that contains the individual.

Census data are geographically aggregated tabular data. The columns of the table are the census attributes and the rows represent geographic aggregation units. The fact that they are geographically aggregated has two important implications: 1) aggregation is the privacy protection method implemented, which will be discussed in the next section; 2) the data adds spatial reference in textual format to the data.

2.2.2 Protection techniques

The techniques to protect geographic information privacy are referred to as geographic masking [28, 29, 30, 31]. In practice, some common techniques of geographic masking are summarized as follows by [29]:

- **Perturbation**

  This adds random perturbation, within a specified limit [29] or within both upper and lower limits [32], to the coordinates of the objects. This method is usually applied to the data in geometrical format. The advantage is preserving the independence of each object. In other words, objects in the resulting dataset and the objects in the original dataset are bijective. This preserves the linkability [33] between the geographically masked dataset and the original dataset.

- **Aggregation**

Figure 2.4: Aggregation method is used in census data protection to avoid unique identification

This is one of the most widely used techniques of geographic masking and is used in a variety of ways. The general idea is to aggregate people in a region in order to prevent unique identification. After the aggregation, the features of the regionalized attribute are used to represent the features of the aggregated individuals. This technique can be applied to both textual format and geometrical format geographic data. Compared to perturbation, aggregation techniques do not retain the linkability.

- For textual geographic information, usually all people in the same region are enumerated and the value of their attributes are aggregated. The aggregated attributes are then used to represent that of all people in the region. One typical example is to aggregate specific address to zip code areas. Census data implements this strategy for privacy masking.
- For data in geometrical format, one aggregation technique is to use one type of geometric median (e.g. centroid) to represent all neighboring data [34]. Another technique is spatial cloaking [35], which extends the idea of k-anonymity to spatial domain. It constructs a spatial region, which contains at least k spatial entities, to be used as the aggregation and ensure that the k entities are indistinguishable.

In census data release, aggregation is the primary method for privacy protection, which is clearly stated on the Data Protection page on Census Bureau website (Fig. 2.4). The individual's census survey response are aggregataed on different levels of census geography ranging from census block to the nation.

2.3 Relevant spatial statistics theories

In this section, I introduce the spatial statistical theories and tools that are relevant to this thesis. The focus will be put on their applications in and relevance to the spatial variation of privacy in census data. It should be noted that areal data model is implicit in this section since census data is aggregated and represented in by geographically aggregated area.

2.3.1 Scale in spatial statistics

Scale has been a central issues to geography [36] and geographic information science [27]. Historically, scale carries different meanings among human geographers, physical geographers and GIScientists [36]. Physical geographers consider the scale to be the granularity of the partition to the space. They often use hierarchy theory which partition the space into hierarchical system where horizontal levels consist of equal size units, each of which then was subdivided into smaller units. By contrast, human geographers consider scale to be socially defined. Most intuitively, this involves using the administrative boundaries such as blocks, zip codes, counties, cities and states [36]. In some cases, it also extends to carry more socially constructed meanings using the socially constructed space theory. GIScientists and cartographers consider scale mostly in the production, storage and usage of geographic data [37]. In this thesis, I adopt the human geographers' definition of scale, treating it as different census aggregation units including zip codes, counties, metropolitan statistical areas and states.

A scale issue that is very related to this study is the Modifiable Area Unit Problem (MAUP). Geographers have long been studying the influence of scale on the statistical results. Some phenomena are only observable on certain level of scale but not in the other. Stan Openshaw and Peter Taylor identified this problem as modifiable area unit problem [38]. The optimal scale of analysis is usually decided by 1) finding the operational scale [39] based on data per se or 2) conducting a scale-variance approach that measures the variance of single variable at different scales. In this thesis, I follow the idea of the second approach and examine the scale's effects by computing the uniqueness on various scales, ranging from zip code, county, metropolitan statistical area and state.

2.3.2 Spatial autocorrelation and cluster/outliers analysis

Global spatial autocorrelation and its interpretation

Spatial autocorrelation is the phenomena that the variable correlates with it self according to spatial proximity. It operationalizes the Toblers First Law of Geography that "everything is related to everything else, but near things are more related than distant things" [40]. This is a very common statistical problem in spatial domain that is caused by either physical processes or human activities.

Strictly speaking, spatial autocorrelation describes the spatial patterns from a data-driven perspective [41]. By looking solely at the data per se, it characterizes to what degree similar observation clusters in spatial. It essentially measures the effect of distance to the distribution of observation. Negative spatial autocorrelation indicates that similar values are away from each other, while positive spatial autocorrelation shows that similar values are close to each other. Zero spatial autocorrelation suggests complete spatial randomness. However, it does not explain the underlying spatial processes which generate the spatial patterns. In other words, it doe s not explain why the similar values group together.

By contrast, model-specification perspective, which is mostly popular in spatial econometrics, specifies the spatial processes into two main categories, spatial dependence and spatial heterogeneity [42]. In the model, the spatial dependence is usually expressed by the spatially lagged term and the spatial heterogeneity is usually represented by the heteroskedasticity. Their differences can be also understood from scale perspective. While spatial heterogeneity focuses on global regimes [43], spatial dependence focuses on local non-independence occurrence [44]. However, the challenge of the model-specification perspective is the difficulty to separate spatial dependence with spatial heterogeneity [45, 42], since both of them could lead to a positive and significant spatial autocorrelation [46, 47].

In spatial demography spatial heterogeneity is the norm [44, 48, 49]. This is because that sub-regional patterns dominates people's residency, behavior and other demographic processes. As a result, the positive global spatial autocorrelation in this thesis is interpreted as heterogeneity rather than spatial dependence. The interpretation of local spatial autocorrelation will be further discussed in next section.

Numerous statistics are devised to measure the extent of spatial autocorrelation. Some of the mostly used statistics includes Moran's I and Geary's C. Global Moran's I takes the general form of Eq. 2.1:

$$I = \frac{N \sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2 \sum_i \sum_j wij} \tag{2.1}$$

where X_i and X_j denote the variable, w_{ij} denotes the contiguity matrix of X_i, N denotes the number of X_i. Essentially, Moran's I computes the covariance between the X_i and its neighbor X_j and offsets the value by eliminating the effect of variance of X_i itself and the w_{ij}.

2.3.3 Local spatial autocorrelation

Global Moran's I can also be written in local form as Eq. 2.2

$$I_i = \frac{(N-1) \cdot (X_i - \bar{X}) \cdot \sum_{j \neq i} w_{ij}(X_j - \bar{X})}{\sum_{j \neq i}(X_j - \bar{X})^2} \tag{2.2}$$

The formula in Eq. 2.2 is also known as one form of Local Indicator of Spatial Association (LISA) [51]. LISA computes a statistics for each areal feature that measures the degree of spatial autocorrelation between this feature and the its neighboring features.

One powerful usage of LISA is in the Clusters and Outliers Analysis. Taking Local Morans I for example, by computing its expectation and variance and following a Bonferroni bounds procedure [50], the statistical significance of each Local Morans I statistics can be assessed [51]. Local hot spots and cold spots are then identified as sets of contiguous features of which Local Morans I are positives and significant. On the other hand, local outliers are defined as features of which Local Morans I are negative and significant. In this thesis, LISA is used to find significant hot spots and outliers of the high/low individual uniqueness.

Chapter 3

First Empirical Analysis: Exploring Regional Differences and Scale Differences

This section answers the Research Questions 1 in the Section 1.1 - is the individual uniqueness in simple demographics spatially homogenous across U.S.?

In this analysis, I break down the question to two parts to examine scale differences and regional differences respectively. This analysis is exploratory - to qualitatively assess the existence of scale and regional differences using some exemplary states. The purpose of this section is three-fold: 1) introducing a methodology to measure spatial variation of uniqueness in census data, 2) demonstrating the existence of spatial variations on individual uniqueness, which is missing in previous studies and 3) suggesting interesting spatial pattern of uniqueness distribution, which will be scrutinized in later chapter.

Analysis in this section extends and compares to previous studies by Golle [1] and Sweeney [8]. The studies from Golle and Sweeney revealed that by using solely U.S. census data, over 60% population could be uniquely identified with (gender, zip code, date of birth) in 1990 and 2000. However, their research only looked at the national level individual uniqueness and failed to consider if the uniqueness rate differs from place to place. Specifically, this section accomplishes the following items:

- The national level individual uniqueness is computed based on the 2010 census.

This result is compared to the 2000 census data from previous studies. The comparison shows that the individual uniqueness remains the same (62% in 2010 census and 63% in 2000 census) but more middle-age group people are uniquely identifiable in the 2010 census data.

- Florida, Connecticut, Iowa and Minnesota are chosen as examples to show the regional differences. The results indicate some remarkable differences among the four states. More importantly, besides the differences, the results also suggest some regionally invariant characteristics of uniqueness that deserve further investigation.

- Variations on spatial scales are analyzed using the above four states. In each state, the overall individual uniqueness on state level, metropolitan statistical area level and county level are computed and compared. The results suggest that although general trends persist at different scales, inter-age-group variation subsides and the overall uniqueness decreases as the scales become smaller and more focused on the core of the region.

3.1 Data

In order to be comparable with previous studies of the 1990 and the 2000 census data, the same data sources are used for this study. Specifically, I use table PCT 12 (Sex by Age) from 2010 census data release, which documents the respective number of males and females, of the specific age (1 year interval between 0 and 99 years old, 5 year interval between 100 to 110 years old and over 110 years old) in specific ZIP Code Tabulation Areas (ZCTAs). According to [52], ZCTAs are the generalized areal representations of USPS Zip Code service areas. Their construction involves 1) assigning each census block a zip code that the majority of its containing addresses use, and 2) aggregating census block with the same majority zip code. This is the census geographic identity that most approximates zip code and is the same with ZIP code in most cases [52]. The data is downloaded from National Historical Geographic Information System (NHGIS) project [53]. Other auxiliary data are retrieved to translate zip code areas to larger geopolitical boundaries, including city, metropolitan statistical area and state. Specifically, the Zip

Code Lookup table is downloaded from U.S. Department of Labor [54].

3.2 Method

3.2.1 Examine national level individual uniqueness

To examine the national level uniqueness, I calculate the uniquely identifiable population percentage (individual uniqueness) based on two sets of demographic characteristics: (zip code, gender, date of birth) and (zip code, gender, age). These two sets of the attributes are chosen because they are commonly requested by most online registration process ranging from social media such as Facebook to the financial agency such as auto insurance company.

Since the table PCT12 (Sex by Age) is aggregated by age, it needs an additional transformation to estimate the number of people born on a specific date. This estimation can be reduced into the question of the expected number of days in a year on which k people are born in a given geographic region and calculated by:

$$E(X_n^k) = C_k^n (\frac{1}{365})^{k-1} \cdot (\frac{364}{365})^{n-k} \tag{3.1}$$

Where X_n^k is the number of days in a specific year on which exact k individual are born, given n as the total number of people in the given gender, zip code and year. Golle offers a proof of the solution in his paper [1].

As a result, the generic form of estimation of population **under** k-anonymity of i-anonymity, given (zip code, gender, date of birth) is as follows:

$$P_{i-anonymity} = \sum_{k=1}^{i} (k \cdot E(X_n^k)) \tag{3.2}$$

3.2.2 Exploring spatial variation of uniqueness - using examplary states

I explore the spatial variation of uniqueness in two dimensions: regional differences and scale differences, using examplary states.

Regional differences

I choose these four states, Florida (FL), Iowa (IA), Minnesota (MN) and Connecticut (CT), due to their distinct age-gender population constitution, seen in the respective population pyramids (Fig. 3.1)

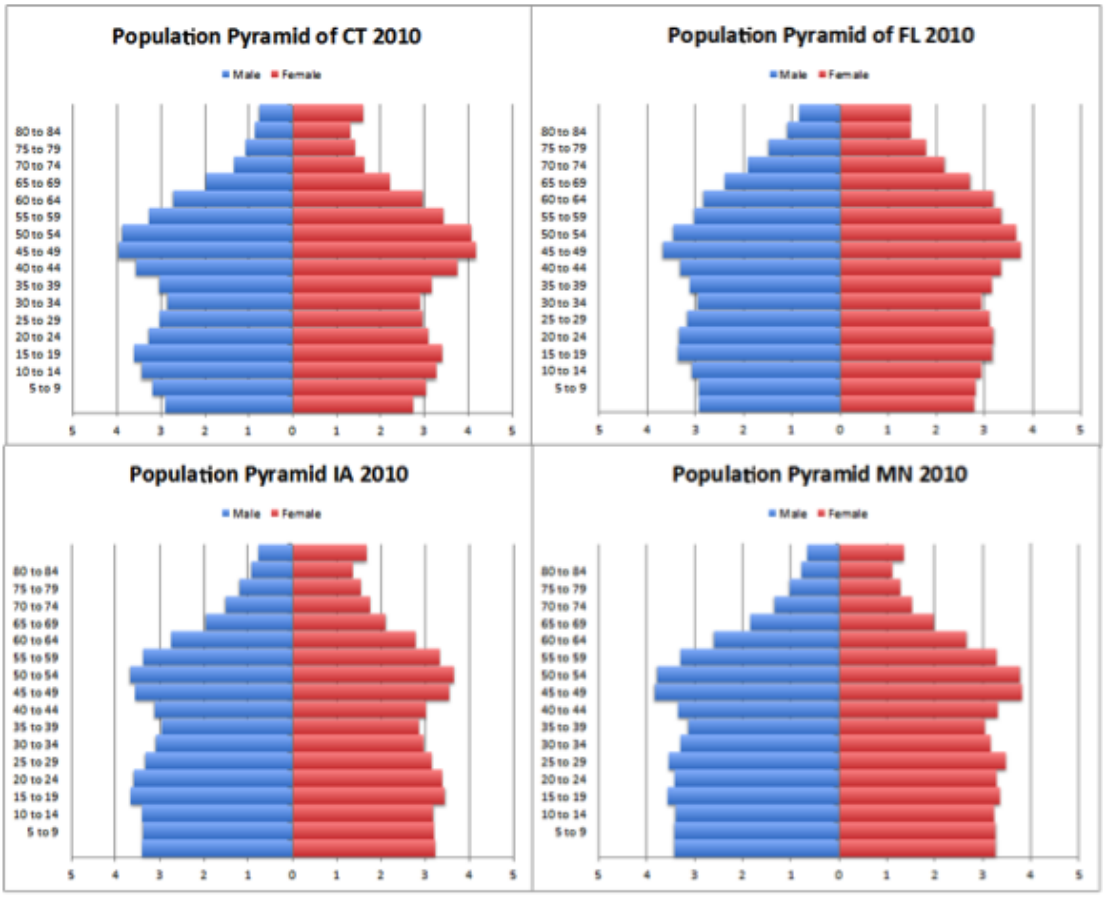

Figure 3.1: Population pyramids of Florida, Minnesota, Iowa and Connecticut

The population pyramids of the four states show the regional differences of the simple demographics used in this paper. For example, Florida stands out with a dominant percentage of the people above 60 years old while Minnesota exhibit a higher percentage of middle age (30-40 years old) people. In this paper, both the state-wise overall uniqueness and age group uniqueness of the four states are computed based on (date of

Table 3.1: National level uniquely identifiable population percentage, 2000 and 2010 (2000 result is excerpted from [1])

Given Attributes	2000 census data	2010 census data
Gender, age, zip code	0.20%	0.14%
Gender, date of birth, zip code	63.30%	62.48%

birth, gender, zip code) using Eq. 3.2.

Scale differences

Scale has been a central topic in almost every facet of geographical research and held diverse meanings for geographers working in different domains [36]. As mentioned before, in this paper I adopt a human geographical concept of scale, meaning that it refers to the scope of the study area. (That is, a large scale refers to a large area.) For each of the four selected states, I select one metropolitan statistical area (MSA) within that state and one county within that MSA. Similarly, both the overall uniqueness and the age specific uniqueness are computed using Eq. 3.2.

3.3 Result and analysis

3.3.1 2010 national level privacy

First, I compute the national level percentage of the uniquely identifiable population using (gender, date of birth, zip code) and (gender, age, zip code), based on 2010 census data. The results are compared with those of 2000 census data. (See Table 3.1)

Second, following Golles paper [1], I provide a fine-grained characterization of the national level uniqueness of different age group population. I compute the population percentages that are 1) under 1-anonymity (uniqueness), 2) under 2-anonymity and 3) under 5-anonymity and compared with those of 2000. (See Fig. 3.2)

Table 3.1 and Fig. 3.2 shed light on the changes of national level uniqueness in the past ten years. In general, the results confirm the the conclusion from [1] that significant amount of people can be uniquely identified by the simple and frequently used demographic attributes. It additionally shows that the level of anonymity has not improved between 2000 and 2010. However, when breaking down to anonymity of

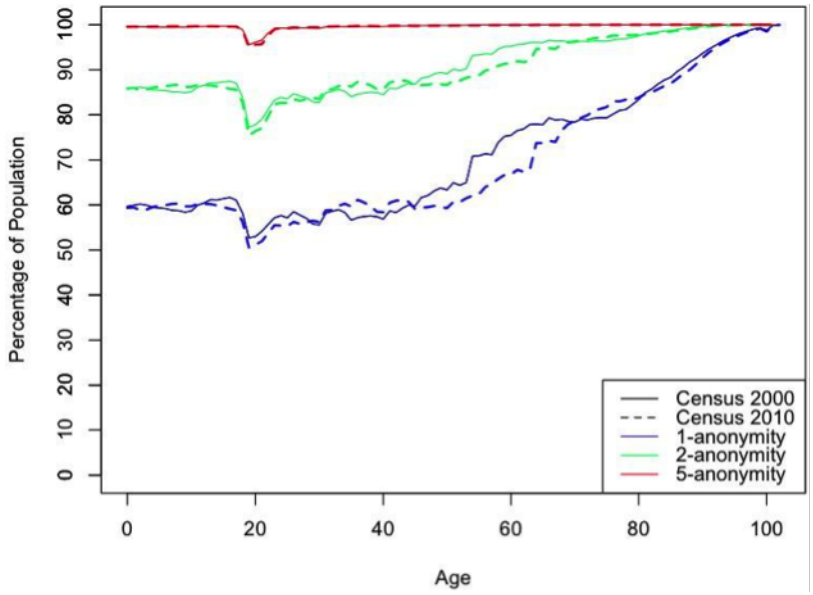

Figure 3.2: Age-specific national level anonymity of U.S population, 2000 and 2010

different age-group population, there is remarkably fewer people in middle age (from age 45 to age 70) to be uniquely identified from their peers, thus showing an improvement of individual anonymity of that particular group of people. This change is of particular interest since it cannot be completely explained by matching the uniqueness of this group to that of the corresponding group (age 35 to age 60) in 2000 result. Therefore, it indicates more complex social-economic changes of this group of people during these ten years, which is beyond the scope of this paper, but deserves further investigation.

3.3.2 Spatial variation of uniqueness

The impacts of regional differences and scale differences to uniqueness are analyzed using the selected states and scales based on the explanation in previous section I compute only the population percentage of 1-anonymity (i.e. individual uniqueness) given gender, date of birth and zip code for the easiness of comparison.

Table 3.2 and Fig. 3.3 allow us to consider how regional differences impact individual uniqueness. Table 3.2 presents the total population and the overall uniquely identifiable

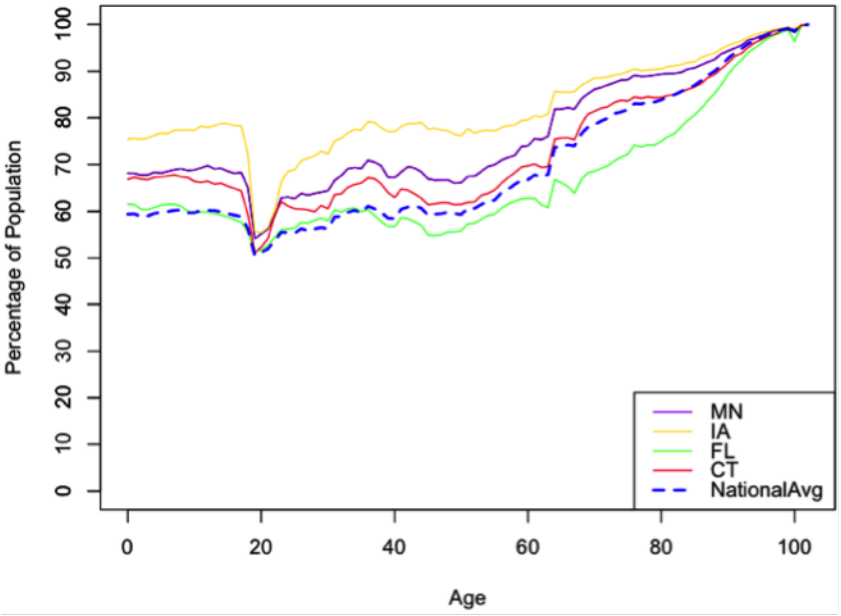

Figure 3.3: Regional differences of age-group uniqueness, using MN, FL, IA, CT

population percentage of the four states. Fig. 3.3 offers a more fine-grained comparison of the age-specific uniqueness in these states and the national average. Although the overall uniqueness differs markedly between states, some similarities show in the age-group uniqueness comparison. First, despite the different youth population percentages, the uniquely identifiable percentage of the college-age population in the four states drop to approximately the same level. Golle attributed this plunge to the concentration of campus student housing, which makes the people in these areas more homogeneous. However, the results of this paper further imply that this could be a spatially invariant feature in most part of U.S. Second, the same extent (about 10%) increase from age 63 to age 69 is observed from all four states. Unfortunately, I find no clear explanation to this trend. These two similarities are of special interests, considering the disparate state-wise demographics (see Fig. 3.1) and might imply some common spatially distinct characteristics to explore further.

Table 3.3 and the horizontal interpretations of Fig. 3.4 explain the scale differences of uniqueness. The legend of Fig. 3.4 indicates the specific MSA and counties I selected. It

Table 3.2: Regional differences of overall uniqueness, using MN, FL, IA, CT

	Minnesota	Florida	Iowa	Connecticut
Population	5,304,141	18,801,226	3,046,945	3,574,333
Pct. of uniquely identified people	70.18%	60.99%	77.43%	66.76%

Table 3.3: Scale differences of overall uniqueness, using state, MSA, county

	State level	MSA level	County level
Minnesota	70.18%	62.89%	65.12%
Iowa	77.43%	71.5%	67.68%
Florida	60.99%	53.49%	51.90%
Connecticut	66.76%	67.32%	65.15%

is worth reiterating that these MSAs and counties are chosen to ensure that the smaller scale regions reside in bigger scale regions. This nesting of scales provides a comparison between the core of the region and the region overall. Table 3.3 reveals the general trend in which the percentages of uniquely identifiable population decrease in the core area of smaller scales, meaning proportionally more people are anonymous. The horizontal reading of Fig. 3.4 verifies this trend and further suggests that the age-group differences also tend to subside in core regions at smaller scales.

In this section, I provided the first empirical analysis of the spatial variation of uniqueness in the census data given the simple demographics. I first compare the national level uniqueness between the 2010 census and the 2000 census. The findings suggest that the overall uniqueness does not change noticeably, yet there is a significant decrease to the uniquely identifiable population of the middle-aged group. This decrease reveals that people from middle-age group (age 45 to age 60) are more spatially aggregated in 2010 comparing to 2000. I then break down the data into regions of different demographic constitutions and into scales of different areas. The analysis of the regional differences indicates some spatial invariants of the uniqueness among the four states (i.e. FL, IA, MN and CT). Although both the overall uniqueness and the age-group uniqueness vary remarkably between states, the college-age group tends to reach the same level of uniqueness. Furthermore, in the group between the age of 63 and the age of 70, a similar increase (approximately 10%) of uniquely identifiable population percentage exists, despite the different average uniqueness percentage. The study of the

scale difference suggests that the uniquely identifiable population percentage decreases as the scales become smaller and more focus on the core of the region. In addition, the age-group variation also diminishes in smaller scale regions.

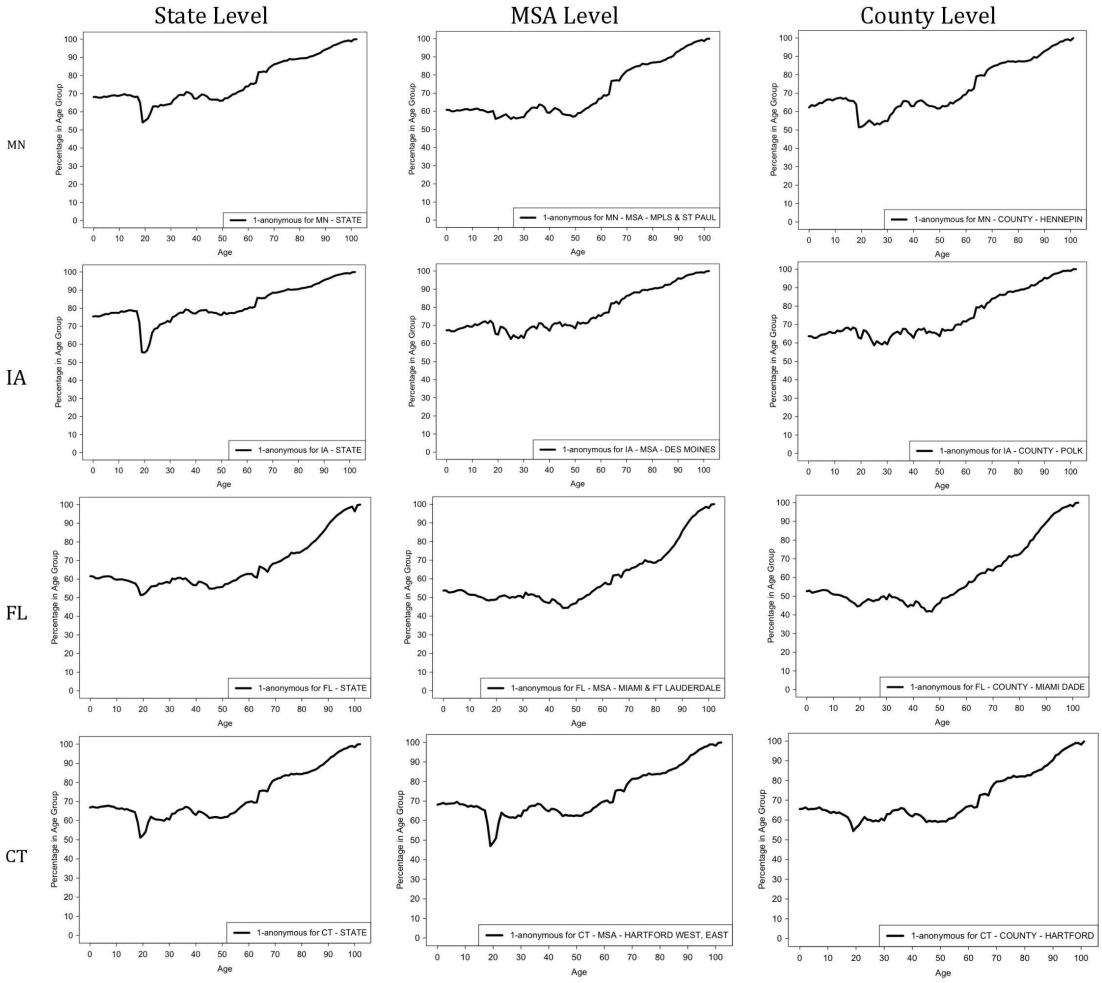

Figure 3.4: Scale differences of age-group uniqueness, using MN, FL, IA, CT

Chapter 4

Second Empirical Analysis: Finding Spatial Patterns using Global and Local Analysis

This section answers the Research Questions 2 in Section 1.2 - what are the spatial distribution patterns of the individual uniqueness in the contiguous U.S. and what characteristics do high (or low) individual uniqueness regions have? The specific types of spatial distribution patterns examined in this section are motivated by the exploratory results from Chapter 3.

Both global analysis and local analysis are carried out in this section:

- In the global analysis, I extend the analysis in Chapter 3 to consider all states in the contiguous U.S. I first provide an overall assessment of the spatial association of state-wise individual uniqueness. Second, I verify some of the trends indicated by the exploratory analysis results in the complete contiguous U.S. context. Specifically, I am interested in verifying if the individual uniqueness of college-age group population is spatially homogeneous. Global spatial analysis is conducted using both state-level and ZCTA (zip code tabulation area) level spatial scales.

- In the local analysis, the focus is on identifying the statistically significant spatial clusters of high/low uniqueness. The purpose is to see at what places people are

more vulnerable to privacy attacks using socio-demographic data. I first conduct the Clusters and Outliers Analysis using Local Indicators of Spatial Associations to identify ZCTA clusters with significant low individual uniqueness. Secondly, I overlay the low uniqueness spatial clusters with the urban areas to investigate if urban-rural divides characterize the locations of high/low uniqueness clusters. To best approach the localness, the analysis is only operated on the ZCTA level spatial scale.

4.1 Global analysis

In this section, I conduct the global spatial analysis on both state level and ZCTA level spatial scales. Following the definition from the previous section, the individual uniqueness is defined as the percentage of individual in a given geographic aggregated area that can be uniquely identified by zip code, gender, and date of birth.

4.1.1 Overall assessment of global spatial association

The state-wise individual uniqueness are computed for all states in contiguous U.S. The complete computation results are included in the Appendix. Table 4.1 and Table 4.2 are the states with the top 5 highest or lowest individual uniqueness. Fig. 4.1 is the choropleth map that shows the state level individual uniqueness in the contiguous U.S. (Classification are generated using Jenks Natural Break.) The map indicates strong spatial associations and obvious spatial heterogeneity.

Table 4.1: Top 5 states with lowest individual uniqueness

State	Population	Individual Uniqueness
CA	37249542	49.86%
DC	601723	50.11%
NV	2701225	55.11%
AZ	6394519	56.18%
NY	19378077	56.65%

- States in the southwest U.S. are associated with low individual uniqueness.

Table 4.2: Top 5 states with highest individual uniqueness

State	Population	Individual Uniqueness
VT	625741	85.86%
ME	1328255	83.73%
WV	1852774	82.30%
SD	813464	78.27%
NH	1316573	77.92%

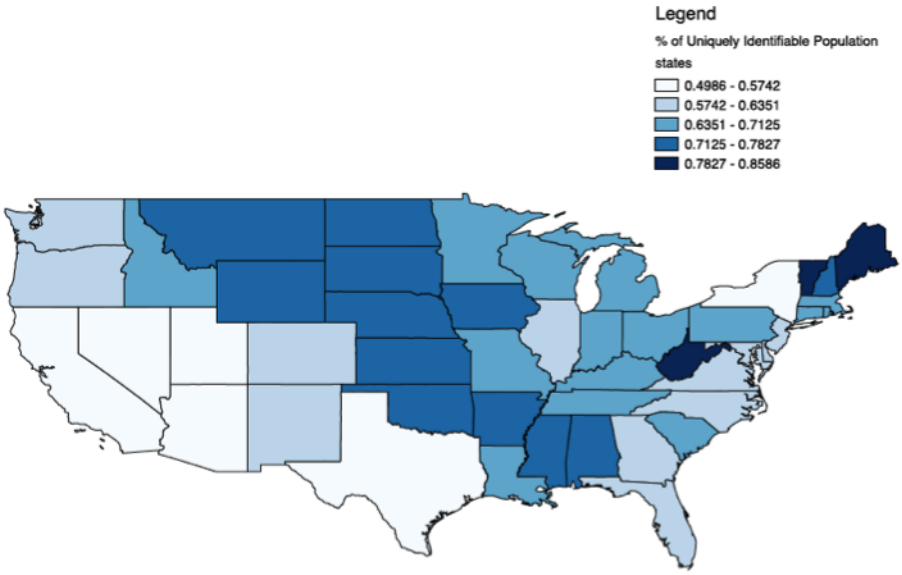

Figure 4.1: State-level individual uniqueness of the contiguous U.S.

- Midwest states are typically associated with relatively high individual uniqueness rate.
- States in the eastern U.S. are generally associated with medium uniqueness rate.
- States in the New England region are represented with highest individual uniqueness.

The relatively strong global spatial association is also presented in analysis at the ZCTA level. Compared to Fig. 4.1, Fig. 4.2 shows similar patterns of spatial association. It shows that large homogeneous ZCTAs with high individual uniqueness are observed

in the central U.S. while ZCTAs with relatively low uniqueness mostly appear in the east and west parts of the U.S.

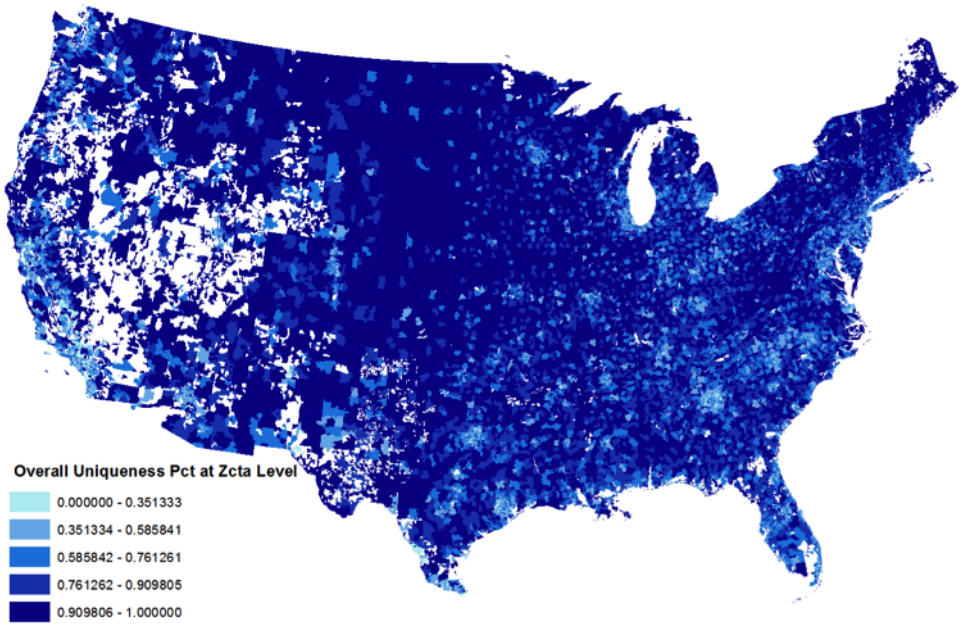

Figure 4.2: ZCTA-level individual uniqueness of contiguous U.S.

Finally, the strong overall global spatial association can be confirmed by the Moran's I value. Table 4.3 shows the Global Moran's I computed at state level and ZCTA level, using various spatial weight functions. P-values are computed using the permutation test [55]. The results show that the spatial associations of individual uniqueness are highly significant at both the large scale and the small scale, regardless of the spatial weight function.

The computation of Moran's I entails the construction of a spatial weight matrix. It is worth noting that the ZCTA defined by Census Bureau does not cover the complete contiguous U.S. This is because remote non-residential places such as national parks and water are not included. This is reasonable in our scenario since the individual uniqueness is meaningless in places with no people. As a result, the holes are treated as the natural phenomenon rather than the missing value in this thesis and can be excluded in the Moran's I computation.

Table 4.3: Global Moran's I (p-value) of individual uniqueness in contiguous U.S. at State level and ZCTA level.

Spatial Weight Function	ZCTA Level Scale	State Level Scale
Queen 1st order Contiguity (row standardized)	0.532 (0.001)	0.3648 (0.002)
Rook 1st order Contiguity (row standardized)	0.530 (0.001)	0.3640 (0.001)
Nearest Neighbor (4 for State, 8 for ZCTA)	0.537 (0.001)	0. 1877 (0.006)

4.1.2 Spatially homogeneously low uniqueness for college-age population

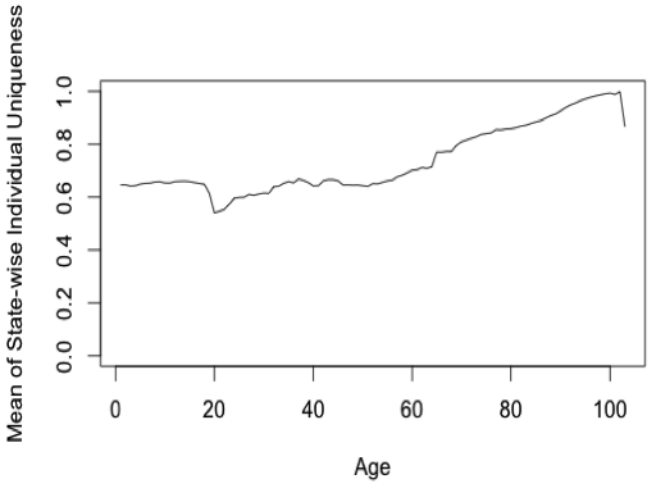

Figure 4.3: Mean of age group individual uniqueness of states in contiguous U.S.

This analysis tests whether the low individual uniqueness of the college age group, observed in the exploratory analysis, is indeed a global spatial phenomenon. In Chapter 3, in all of MN, IA, FL and CT, the individual uniqueness indicator drops significantly to a similarly low value for the age group 18-22. As I suggested, the decrease and low individual uniqueness values can be explained by college students clustered residency around campus. I am interested to see if the individual uniqueness is spatially homogeneously low in every part of the U.S.

I verify this by computing the descriptive statistics of age-wise individual uniqueness

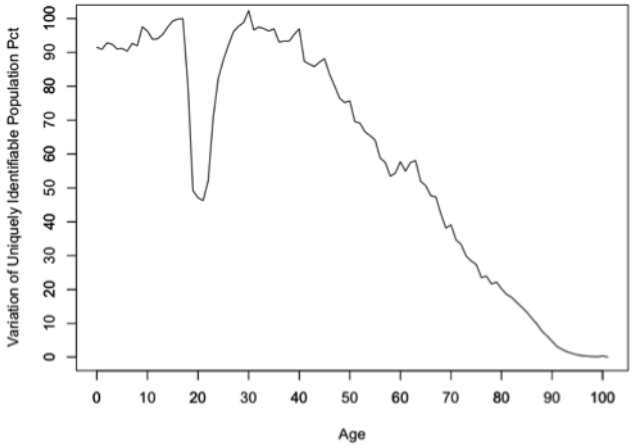

Figure 4.4: Variation of age group individual uniqueness of states in contiguous U.S.

and their Moran's I for all states in the contiguous U.S.. Fig. 4.3 shows that the mean of individual uniqueness is the smallest around the age group of 20. Fig. 4.4 indicates that on the state level, the variation of the individual uniqueness drops significantly at the age 20. These two evidences reveal that the individual uniqueness of all the states varies little for the college-age age group and remains at a low value. Fig. 4.5 demonstrates the spatial homogeneity, which shows that the spatial association is also lowest for the individual uniqueness of the college age group. The weak spatial association suggests weak spatial heterogeneity in spatial demographic studies according to our discussion in Section 2.3.2. These three figures jointly suggest that the low individual uniqueness for the college age is a spatially homogeneous phenomenon.

4.2 Local analysis

In this section, the local spatial analysis is carried out on ZCTA level. The focus is to identify the significant spatial clusters of individual uniqueness. This analysis focuses on locating clusters and outliers, which are defined as follows:

- Clusters:
  - High/high (HH) cluster: ZCTA with high individual uniqueness that are

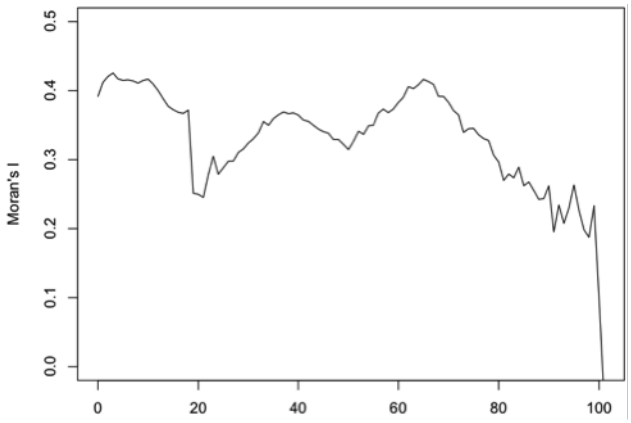

Figure 4.5: Moran's I of age group individual uniqueness of states in contiguous U.S.

surrounded by ZCTAs with high individual uniqueness

  - Low/low (LL) cluster: ZCTA with low individual uniqueness that are surrounded by ZCTAs with low individual uniqueness

- Outliers:

  - High/low (HL) outlier: ZCTA with high individual uniqueness that are surrounded by ZCTAs with low individual uniqueness
  - Low/high (LH) outlier: ZCTA with low individual uniqueness that are surrounded by ZCTAs with high individual uniqueness

In addition to the Cluster/Outlier Analysis, further analysis is conducted to investigate the socio-demographic characteristics of these clusters and outliers. Specifically, I investigate the effect of urban-rural divide.

4.2.1 Clusters and outliers analysis

ZCTA level individual uniqueness dataset are used to identify the clusters and outliers. The identification procedure first calculates a local Moran's I (see Section 2), associated z-score and p-value for each ZCTA. It then classifies the ZCTAs into clusters and outliers based on the local Moran's I value and its statistical significance (p¡0.05). Specifically,

the procedure selects ZCTAs with significant positive local Moran's I value to be clusters and significant negative local Moran's I value to be outliers. Finally, the individual uniqueness of that ZCTA and surrounding ZCTA are used to determine the specific type of the clusters/outliers.

It is worth noting that since the distribution of individual uniqueness on ZCTA level is heavily skewed (See Fig. 4.6), each feature should have at least eight neighbors [56].

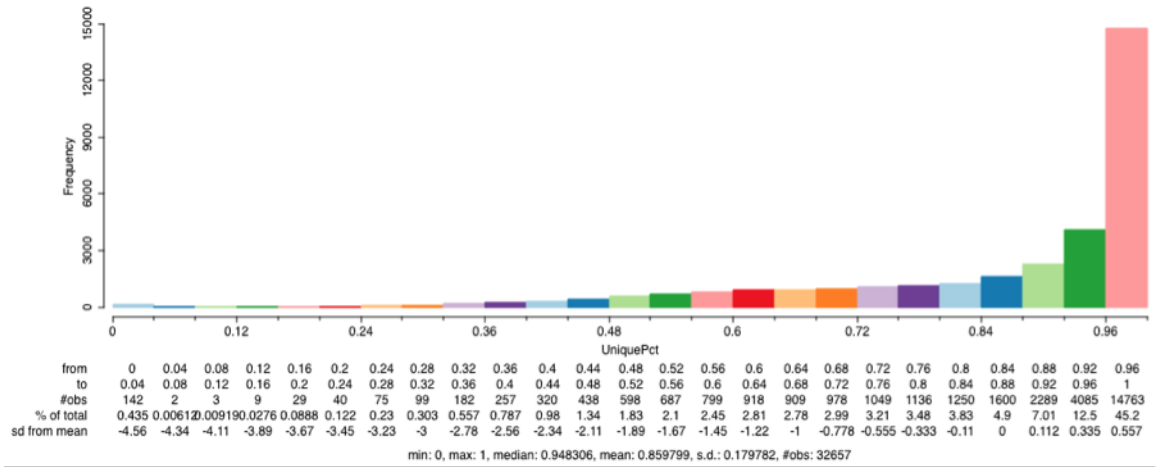

Figure 4.6: Distribution of ZCTA level individual uniqueness in the contiguous U.S.

Table 4.4: Summaries of the Clusters and Outliers Analysis on ZCTA level

Type		Stands for	Counts
Outlier	LH	Low uniqueness surrounded by high uniqueness	187
	HL	High uniqueness surrounded by low uniquneess	250
Cluster	LL	Low uniqueness surrounded by low uniqueness	4922
	HH	High uniqueness surrounded by high uniquness	0

The interpretation focuses on the LL cluster, because they dominate the significant cluster and outliers analysis, according to Table 4.4. Fig. 4.7 shows the overall distribution of the LL clusters on ZCTA level. For the convenience of interpretation, the result is further post-processed by local aggregation, which aggregates neighboring LL clusters to form Local LL Regions. These Local LL Region are essentially collections of neighboring significant LL ZCTAs. The purpose of this post-processing is to allow us

to interpret the location of the LL ZCTAs in a more sensible way.

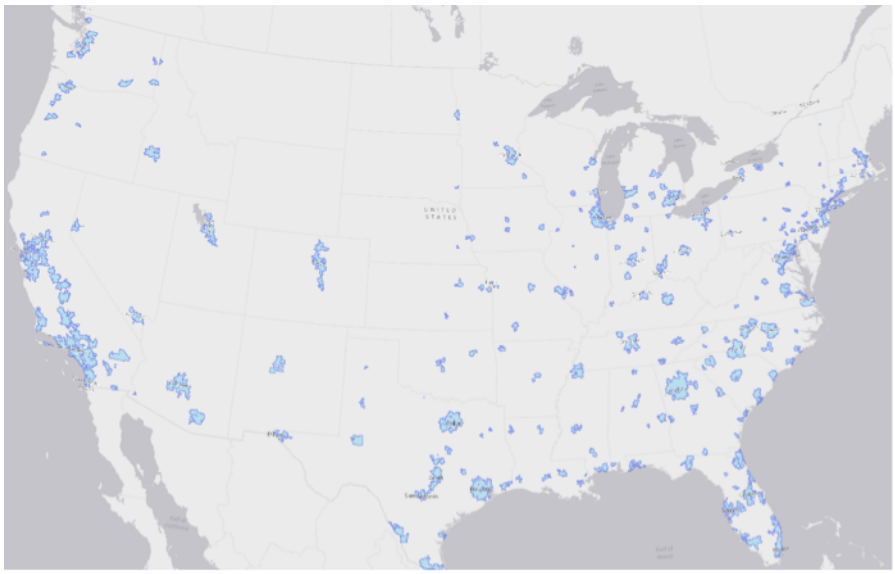

Figure 4.7: Low/Low individual uniqueness clusters on ZCTA level

The post-processing produced 1467 Local LL Regions. These regions are ranked by their area size from the biggest to the smallest to show where the LL clusters are more aggregated. As Fig. 4.8 presents, Southern California, San Francisco, Atlanta, Houston, Dallas, and Chicago are the top 6 Local LL Regions where LL clusters aggregate the most. These are considered the areas where people are less unique and enjoy better privacy. One commonality of these top 6 Local LL Region is they are all centered on very urbanized areas.

4.2.2 Urban-rural divides on individual uniqueness

One observation from previous section is that the top Local LL Regions are all in highly urbanized areas. In fact, the boundaries of the Local LL Regions overlap well with the city boundary. This observation encourages me to explore 1) if urban areas distribution characterized the low individual uniqueness clusters and, 2) if there is an urban-rural divide on individual uniqueness. I first overlay the urban area with the LL individual uniqueness cluster to show their relatedness. Second, I computed the

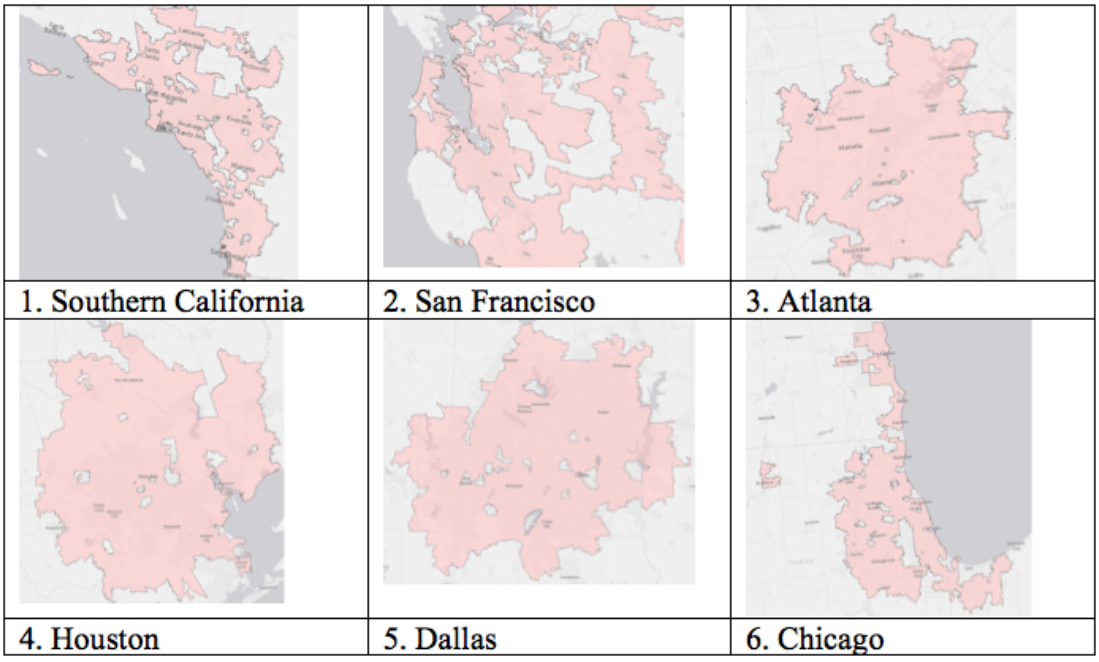

Figure 4.8: Top 10 Low/Low regions by area

individual uniqueness associated with urbanized areas and rural areas respectively to quantitatively evaluate the urban-rural divides.

The urban-rural classification is from the U.S. Census Bureau [57]. The classification set a series of qualifying criteria for a region to be considered urban area. Any territory that does not satisfy the standard is classified as rural area. The qualifying criteria are population-based, which essentially select census tracts that meet minimum population with its adjacent non-residential areas. The urban areas are further classified into Urbanized Area (over 50,000 population) and Urban Clusters (between 2,500 populations to 50,000 populations). The Census Bureau also provides a Relationship File to relate ZCTAs to urban areas [57].

Table 4.5 shows that more than 99% of LL clusters locate in the urban area, which indicates a strong relationship between being urbanized and having low individual uniqueness. For outliers, it is worth noting that large proportions of LH (low uniqueness ZCTAs surrounded by high uniqueness ZCTAs) outliers are also in the urban areas, but only

Table 4.5: Overlay between significant Clusters/Outliers and Urban Area

	Total	In Urbanized Area		In Urban Area	
		Count	Percentage	Count	Percentage
LL clusters	4922	4883	99.21	4919	99.93
LH outliers	187	43	22.99	181	96.79
HL outliers	250	180	72.00	191	76.40
LH outliers	0	0	NA	0	NA

small proportions are in Urbanized Area. I explain this by the fact that LH outliers are on the urban side of the edge of the urban-rural area. Although they have low uniqueness by themselves, they are adjacent to the high uniqueness rural area. Since they are away from the urban center, their population might not be high (thus not classified as Urbanized Area).

The fact that urban areas characterize most of the LL clusters suggests an urban-rural divide on the distribution of individual uniqueness. It suggests the hypothesis that individual uniqueness in urban areas is lower than that in rural areas. I examined this hypothesis by computing the overall individual uniqueness rate for urban areas and rural areas respectively.

Table 4.6: Comparison of overall uniqueness between urban and rural area

	Urban Area	Rural Area
Population	266,641,301	45,821,696
Percentage uniquely identified people	58.02%	88.43%

Table 4.6 compares the overall individual uniqueness between the urban area and the rural area. The comparison reveals a clear urban and rural divide. The uniqueness rate of individuals in rural areas is significantly higher than the uniqueness rate of individuals in urban areas. In other words, if a person lives in an urban area, he will be 52% less likely to be uniquely identified than if he lives in a rural area.

Extending the analysis from Chapter 3, Chapter 4 uses the complete U.S. census dataset to explore the specific spatial patterns both on global and local scales. In the global analysis, I analyzed the global spatial association patterns for both overall individual uniqueness and the age-group individual uniqueness. The results indicate

obvious spatial heterogeneity on overall uniqueness and spatial homogeneity on uniqueness for a specific group. The uniqueness rates are low in the western regions of the U.S., medium-high in the Midwest and medium-low in the eastern of the U.S. and high in the New England region. Since low uniqueness represents less identifiability thus high privacy, people who live in the New England and the Midwest are more vulnerable to privacy attacks from a pure census demographic data perspective. In addition, I find that college-age groups low uniqueness rate is globally invariant and spatially homogeneous. In other words, although the overall uniqueness spatially varies, college-age people enjoy the same high privacy in every state in contiguous U.S.

The local analysis focuses on identifying the location of high/low uniqueness clusters and exploring their characteristics, using ZCTA spatial scale. LISA is used as an indicator to show if a region is a significant local cluster. The results show that the LL clusters dominate the statistical significant clusters and outliers. The LL clusters are merged with their adjacent LL cluster to form a Local LL Region. A ranked list based on the area of the Local LL Regions is provided. The ranked list shows that the top Local LL Regions are all highly urbanized area that are densely populated. Following this strong correlation between the urbanized area and the low uniqueness clusters, I further compare the urban and rural individual uniqueness. The comparison showed a clear urban and rural divide, which suggests that if a person lives in urban area, he is 52% less likely to be uniquely identified than if he lives in rural area from a pure census demographic perspective.

Chapter 5

Conclusion and Discussion

To the best of my knowledge, this thesis provides the first empirical spatial analysis of the individual uniqueness based on simple demographic information in U.S. at different scales and comparing different regions. In the context of this thesis, individual uniqueness is calculated as the proportion of the uniquely identifiable population, given (zip code, gender, date of birth) in a geographic aggregated area and is used as a proxy to personal privacy. The combination of (zip code, gender, date of birth) represents the most basic and commonly used demographic attributes and is easily exposed in our daily encounters including online registration, financial inquiries even casual conversations. This research extends previous studies, which failed to consider the spatial nature of the question. I first conduct an exploratory analysis using exemplary regions to qualitatively verify the existence of spatial variation and to infer interesting spatial distribution on individual uniqueness. This is followed by the study using a complete US dataset to quantitatively measure the magnitude of spatial association and identify the significant high/low individual uniqueness clusters. My analyses reveal both global and local patterns of the individual uniqueness at multiple spatial scales.

5.1 Conclusion

5.1.1 Global patterns

- **On national level, individual uniqueness remains the same between 2000 and 2010, but middle-age population enjoy better privacy in 2010.**

  I compare the national level individual uniqueness computed using 2000 census and 2010 census. Results show that national level individual uniqueness remains similar (62.48% in 2010, compared to 63.30% in 2000). However, middle-age population (age of 45-60) enjoys lower individual uniqueness in 2010. This means that on national average, fewer middle-aged people can be uniquely identified using this three demographic attributes thus having better privacy.

- **State-wise individual uniqueness presents significant spatial heterogeneity. However, there is little spatial variation in the individual uniqueness of college-age group.**

  The significant spatial association confirms the spatial heterogeneity of individual uniqueness in state-level analysis. A clear spatial regime is shown. In general, states in Midwest and in New England are associated with high individual uniqueness while the states in the southeastern and the eastern U.S. are assocaited with low individual uniqueness.

  The above two findings weaken the Census Bureau's claim that "no one can be uniquely identified." In fact, there is a high possibility that an individual could be uniquely identified and this possiblity is unevenly distributed in U.S. It suggests that the Census Bureau should consider the spatial heterogeneity when devising census data aggregation units and propose spatially varying aggregation criteria to ensure the privacy is ensured equally across space.

  In contrast to the significant spatial heterogeneity on overall individual uniqueness, age-wise uniqueness presents different patterns. Most notably, there is little spatial variation for the individual uniqueness of the college-age population. Considering also the low mean and variation of college-ages individual uniqueness, it means that these people enjoy a low identification rate in every part of U.S.

5.1.2 Local patterns

- **Low/Low Clusters dominates the significant clusters and outliers on ZCTA level and they aggregate most in big urban area.**

  HH, LL, HL and LH 4 types of statistically significant cluster and outliers are identified from the Cluster and Outliers Analysis using LISA on ZCTA level. Results show that the LL clusters dominate the statistically significant clusters. This indicates that the low uniqueness ZCTA surrounded by other low uniqueness ZCTAs is the most prominent type of spatial pattern.

Table 5.1: Summaries of the Clusters and Outliers Analysis on ZCTA level

Type		Stands for	Counts
Outlier	LH	Low uniqueness surrounded by high uniqueness	187
	HL	High uniqueness surrounded by low uniquneess	250
Cluster	LL	Low uniqueness surrounded by low uniqueness	4922
	HH	High uniqueness surrounded by high uniquness	0

In addition, by merging neighboring LL clusters, I found that LL clusters mostly occur in larger urban areas. The ranked list of the top 6 regions/cities by area is as follows: Southern California, San Francisco, Atlanta, Houston, Dallas and Chicago. These are the places where low uniqueness ZCTA clusters aggregate the most.

- **Urban-rural divides in individual uniqueness distribution** I also investigate if the individual uniqueness in urban areas differs greatly from that in rural areas. The classification of rural and urban areas follows the criteria from US Census Bureaus definition and mainly depends on the population density. Results show that individual uniqueness in urban areas is 58.02% compared to that of 88.43% in rural areas. This means, if a person is from urban area, given the zip code, gender and date of birth information, he/she is much less likely to be uniquely identified in comparison to if he/she lives in a rural area.

  The urban-rural divides is one of the nation's long-lasting splits, which is manifested on economics, education, opportunity and many other aspects. The finding from this thesis adds to the discussion by further indicating that privacy, mesured by individual uniqueness, also follow this urban-rural divide.

5.2 Future research

One future direction of research involves extending the analysis of urban-rural divide to investigate the specific social-demographic attributes that explain the pattern of individual uniqueness. This could involve running multivariable spatial regressions to find out the relations between specific demographic attributes and the uniqueness rate. Finding this correlation will help the improvement of privacy protection strategy in census data by applying place-specific aggregation limits [30].

Second, on national level, the marked decrease of middle-age uniqueness in 2010 indicates a more spatially aggregated residency of middle-aged population. However, the reason for the increasing spatial aggregation remains uninvestigated and deserves exploration with additional theory from population studies.

Finally, given the potential of the simple demographic data in contravening individual privacy protection, coupling census data with other publicly available personal level data, such as online social media, for combination attacks are of great interest for future research.

References

[1] Philippe Golle. Revisiting the uniqueness of simple demographics in the us population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80. ACM, 2006.

[2] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[4] U.S. Census Bureau. Population distribution in the united states and puerto rico [map], 2010.

[5] Julia Angwin and Jennifer Valentino-Devries. Apple, google collect user data, April 2011.

[6] Charles Arthur. Android phones record user-locations according to research, April 2011.

[7] Malte Spitz. Malte spitz: Your phone company is watching, June 2012.

[8] Latanya Sweeney. Uniqueness of simple demographics in the us population. Technical report, Technical report, Carnegie Mellon University, 2000.

[9] Daniel J Solove. Understanding privacy. 2008.

[10] Samuel D Warren and Louis D Brandeis. The right to privacy. *Harvard law review*, pages 193–220, 1890.

[11] Sissela Bok. *Secrets: On the ethics of concealment and revelation*. Vintage, 1989.

[12] Richard A Posner and Aspen Publishers. Economic analysis of law. 2007.

[13] Charles Fried. Privacy. *Yale Law Journal*, 77(3):475, January 1968.

[14] Daniel J Solove. Conceptualizing privacy. *California Law Review*, pages 1087–1155, 2002.

[15] Daniel J Solove. 'i've got nothing to hide'and other misunderstandings of privacy. *San Diego law review*, 44:745, 2007.

[16] Ken Gormley. One hundred years of privacy. *Wis. L. Rev.*, page 1335, 1992.

[17] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[18] Helen Nissenbaum. Privacy as contextual integrity. *Washington law review*, 79(1), 2004.

[19] Seda Gürses. Can you engineer privacy? *Communications of the ACM*, 57(8):20–23, 2014.

[20] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[21] Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

[22] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005.

[23] William Winkler. Using simulated annealing for k-anonymity. Technical report, Research Report 2002-07, US Census Bureau Statistical Research Division, 2002.

[24] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[25] U.S. Census Bureau. Data protection and privacy, Oct. 2014.

[26] Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637, 2008.

[27] Michael F Goodchild. Twenty years of progress: Giscience in 2010. *Journal of spatial information science*, (1):3–20, 2015.

[28] William B Allshouse, Molly K Fitch, Kristen H Hampton, Dionne C Gesink, Irene A Doherty, Peter A Leone, Marc L Serre, and William C Miller. Geomasking sensitive health data and privacy protection: an evaluation using an e911 database. *Geocarto international*, 25(6):443–452, 2010.

[29] Marc P Armstrong, Gerard Rushton, Dale L Zimmerman, et al. Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18(5):497–525, 1999.

[30] Khaled El Emam, Ann Brown, and Philip AbdelMalik. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *Journal of the American Medical Informatics Association*, 16(2):256–266, 2009.

[31] Dale L Zimmerman and Claire Pavlik. Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical Analysis*, 40(1):52–76, 2008.

[32] Dave Stinchcomb. Procedures for geomasking to protect patient confidentiality. In *ESRI International Health GIS Conference*, pages 17–20, 2004.

[33] Andreas Pfitzmann and Marit Hansen. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management, 2009.

[34] Agusti Solanas and Antoni Martínez-Ballesté. Privacy protection in location-based services through a public-key privacy homomorphism. In *Public Key Infrastructure*, pages 362–368. Springer, 2007.

[35] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42. ACM, 2003.

[36] Eric Sheppard and Robert B McMaster. *Scale and geographic inquiry: Nature, society, and method*. John Wiley & Sons, 2008.

[37] Michael F Goodchild. Scales of cybergeography. *Scale and Geographic Inquiry*, page 154, 2008.

[38] Stan Openshaw and Peter J Taylor. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, 21:127–144, 1979.

[39] Petre Stoica and Randolph L Moses. *Introduction to spectral analysis*, volume 1. Prentice hall Upper Saddle River, 1997.

[40] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, pages 234–240, 1970.

[41] Luc Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 1988.

[42] Luc Anselin. Thirty years of spatial econometrics. *Papers in regional science*, 89(1):3–25, 2010.

[43] Luc Anselin. Spatial dependence and spatial structural instability in applied regression analysis*. *Journal of Regional Science*, 30(2):185–207, 1990.

[44] Advanced Spatial Analysis Program. An introduction to spatial heterogeneity.

[45] Luc Anselin. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical analysis*, 20(1):1–17, 1988.

[46] Michael John De Smith, Michael F Goodchild, and Paul Longley. *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador Publishing Ltd, 2007.

[47] Sándor Kabos and Ferenc Csillag. The analysis of spatial association on a regular lattice by join-count statistics without the assumption of first-order homogeneity. *Computers & geosciences*, 28(8):901–910, 2002.

[48] Paul R Voss. Demography as a spatial social science. *Population Research and Policy Review*, 26(5-6):457–476, 2007.

[49] Paul R Voss, Katherine J Curtis White, and Roger B Hammer. Explorations in spatial demography. In *Population change and rural society*, pages 407–429. Springer, 2006.

[50] J Keith Ord and Arthur Getis. Distributional issues concerning distance statistics. Technical report, Working paper, 1994.

[51] Luc Anselin. Local indicators of spatial associationlisa. *Geographical analysis*, 27(2):93–115, 1995.

[52] U.S. Census Bureau. Zip code tabulation areas (zctas), 2010.

[53] Catherine A Fitch and Steven Ruggles. Building the national historical geographic information system. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 36(1):41–51, 2003.

[54] U.S. Department of Labor. Geographic practice cost index values by zip code, 2010.

[55] David O'Sullivan and David Unwin. *Geographic information analysis*. John Wiley & Sons, 2014.

[56] Andy Mitchell. The esri guide to gis analysis: Spatial measurements and statistics. vol 2. redlands, 2005.

[57] U.S. Department of Labor. 2010 census urban and rural classification and urban area criteria, 2010.

Appendix A

Long Tables

A.1 Individual uniqueness computed for all states in the contiguous U.S.

Table A.1: Individual uniqueness computed for all states in the contiguous U.S.

Begin of Table		
State	StatePop	Individual Uniqueness
PR	3723066	0.518219034
MO	5989128	0.703833063
MA	6547612	0.663817136
HI	1360301	0.556410455
RI	1052581	0.660715527
NH	1316573	0.779180321
ME	1328255	0.837286963
VT	625741	0.858555599
CT	3574097	0.667562608
NY	19378077	0.5665054
NJ	8791894	0.630760547
PA	12702375	0.694690586

Continuation of Table A.1		
State	StatePop	Individual Uniqueness
DE	897925	0.628634245
DC	601723	0.501074327
VA	8001239	0.634605049
MD	5773561	0.601959398
WV	1852774	0.822925709
NC	9535477	0.635095828
SC	4625364	0.664595578
GA	9687711	0.594169779
FL	18801226	0.609859272
AL	4779588	0.74925967
TN	6339942	0.658816571
MS	2967323	0.735331065
KY	4345548	0.71248424
OH	11536514	0.666007591
IN	6483792	0.68471888
MI	9883612	0.674067109
IA	3046945	0.77430413
WI	5687012	0.704444345
MN	5304141	0.701792691
SD	813464	0.782654607
ND	671781	0.751554788
MT	990003	0.747940405
IL	12830581	0.591803409
KS	2853186	0.735972969
NE	1826766	0.763392966
LA	4532830	0.693105156
AR	2916042	0.7308755
OK	3751609	0.724084511
TX	25144800	0.574175275
CO	5029374	0.619638003

Continuation of Table A.1		
State	StatePop	Individual Uniqueness
WY	563839	0.75063789
ID	1566982	0.67340227
UT	2763264	0.570688305
AZ	6394519	0.561831622
NM	2056349	0.628467874
NV	2701225	0.551116811
CA	37249542	0.498600177
OR	3831165	0.633780285
WA	6724629	0.630894408
AK	709930	0.692710814
End of Table		