TECHNISCHE
UNIVERSITÄT
DRESDEN

Faculty of Environmental Sciences Department of Geosciences Institute of Cartography

Master Thesis

Privacy Aware Analysis of Spatial Social Media Data

submitted by **Jack Stephan**
born on 08.03.1994 in Grand Rapids, MI, U.S.A

submitted for the academic degree of
Master of Science (M.Sc.)

Submission on 06.10.2020

Supervisors

Dr.-Ing. Alexander Dunkel
Technische Universität Dresden
Marc Löchner M.Sc
Technische Universität Dresden

Original Task Description

The research objective of this thesis is a proof of sufficiency of privacy-enhanced data structures for cartographic visualization. The scope of tasks includes the comparison of a raw set of spatial data (derived from the social media service Flickr) with an equivalent data set processed using the Location Based Social Network (LBSN) structure based on the HyperLogLog algorithm by Löchner et al (2019) and a suitable selection of exemplary cartographic visualizations derived from that data.

Statement of Authorship

Herewith I declare that I am the sole author of the thesis named

„Privacy Aware Analysis of Spatial Social Media Data“

which has been submitted to the thesis assessment board today.

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Dresden, 06.10.2020 Signature

Cartography M.Sc.

Master thesis

Privacy Aware Analysis of Spatial Social Media Data

Jack Stephan

Technical University of Munich

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

TECHNISCHE UNIVERSITÄT DRESDEN

ITC UNIVERSITY OF TWENTE.

TECHNISCHE
UNIVERSITÄT
DRESDEN

Privacy Aware Analysis of Spatial Social Media Data

submitted for the academic degree of Master of Science (M.Sc.)
conducted at the Institute of Cartography, TU Dresden

Author:	Jack Stephan
Study course:	Cartography M.Sc.
Supervisor:	Dr.-Ing. Alexander Dunkel (TUD)
	Marc Löchner M.Sc (TUD)
Reviewer:	Wang Wangshu M.Sc (TUW)
Chair of the Thesis Assessment Board:	Prof. Dr.-Ing. habil. Dirk Burghardt
Date of submission:	06.10.2020

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master’s thesis entitled:

“Privacy Aware Analysis of Spatial Social Media Data”

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Munich, 06.10.2020 Jack Stephan

Acknowledgements

The completion of this thesis would not have been possible without the support of a lot of people. First and foremost, I have to extend my gratitude to my advisors: Dr.-Ing. Alexander Dunkel for his endless patience and help with understanding the topic throughout the whole semester. Marc Löchner M.Sc for his constructive ideas and much needed encouragement. And Prof. Dipl.-Phys. Dr.-Ing. habil. Dirk Burghardt for his guidance and support. Despite the pandemic and our inability to meet in person, I never felt I lacked support, guidance, or mentorship throughout the entire period. and I learned more than I could have imagined. And to all of those involved in the Cartography program who have mentored and taught me throughout the last two years, thank you. I have to give a special thanks to M.Sc. Juliane Cron for everything she has done for the program as a whole and me in particular. I cannot imagine navigating the complexities of my new life in Europe without her enduring support.

Thank you to Dr. Jeanne Gazel for inspiring me in all of my international endeavors and Dr. Annick Anctil for my first experience as a researcher, without their support and mentorship from my time at Michigan State University none of this would have been possible. Thank you to all of my friends and colleagues at the Institute of Chemical Technology in Mumbai for hosting me during my Fulbright journey that brought me to study cartography.

Thank you to my classmates who have been a part of these last two, unforgettable years. And thank you to all my friends I made in Munich, Vienna, and Dresden and my Climate-KIC summer school journey. I would list you all but there are simply too many people to thank. Thank you to my friends from back home your enduring friendship and support means more than you could know.

And finally, absolutely none of what I have accomplished in life would have happened without my Mom, Dad, brother Brad, sister Maureen, and all of my family members. Thank you for everything you have done for me and everything you continue to do.

Abstract

Spatial social media data is generated in massive quantities for the first time in history. Taking advantage of this vast new source of data is advantageous for researchers analysing spatial social media, but the potential privacy conflicts of users losing control of their personal geographic information is concerning.

This thesis aimed to investigate the use of a privacy aware data structure which uses an algorithm called HyperLogLog to process and store this spatial social media data. The data itself comes from a publicly available database called the Yahoo Flickr Creative Commons 100 Million database. This privacy aware data structure was then used to create visualizations in the form of a case study investigating the ability of this privacy aware data structure to both create adequate visualizations for social media analytics as well as preserve the privacy of the users who had contributed the data in the first place. This case study is presented in a linear demonstration of the key features and drawbacks of the privacy aware data structure and a final example demonstrating how these features can be combined.

The results of the case study showed that the use of the privacy aware data structure can in fact preserve privacy in certain use cases. The results also showed that the privacy aware data structure can create the same quality of results as a non-privacy aware counterpart for mapping problems that deal with cardinality, or distinct counts in a set e.g. users, posts, days, etc.

The results indicate that this privacy aware data structure is worth developing further into a comprehensive mapping tool for social media researchers. This mapping tool could leverage the strengths and manage the draw backs of this data structure in order become one tool in data protection.

Contents

Figures

Tables

Formulas

1 Introduction

1.1 Motivation

With the widespread adoption of the smartphone, it became possible to track the location of many individuals for the first time in human history. Currently, there are over 3 billion smartphone users worldwide and that number is trending upward ("Newzoo's Global Mobile Market Report," n.d.). Location Based Social Networks (LBSN) and Location Based Services (LBS) are just two of the ways in which smart phone users utilize their smartphones. But LBS and LBSN are also unique in that they inherently require users to share location data. In the United States, 90% of users keep their location sensor perpetually activated (*Overwhelming Number Of Smartphone Users Keep Location Services Open* |, 2016). The reasons for this can be as benign as convenience, but as Keßler & McKenzie point out, "Users often share their current location unknowingly" (Keßler & McKenzie, 2018). Users can also be enticed to share their location data ostensibly because the net benefits are perceived as positive e.g. Google maps location indicator or increased social media visibility (Kounadi et al., 2018).

This geographic information (GI) is a unique type of personally identifiable information (PII) and warrants individual examination in the greater context of privacy (Keßler & McKenzie, 2018). This GI also contrasts with volunteered geographic information (VGI) in that VGI is produced voluntarily by an untrained group of individuals and can be thought of as a type of crowdsourcing (Gómez-Barrón et al., 2016). GI can subsequently be exploited since it is publicly available for anyone to use, whether legally or illegally (Malhotra et al., n.d.). One example of such a data set is the Yahoo Flickr Creative Commons 100m data set in which one can find 100 million photos with fine grained location accuracy as well as a variety of other personally identifying pieces of information (Mao, 2015). These data sets are utilized in a variety of ways such as targeted advertising, public and private research, geomarketing, and to train algorithms.

Discussion concerning the utility, necessity, and benefit of these large data sets fits into the broader discussion around Big Data and its trends. Proponents of Big Data herald the dawn of this data driven age as a solution to problems or a tool of convenience such as personalized medical care and even personalized daily advices; the implications of sharing such a large amount of PII with commercial enterprises remains to be seen (Harari, 2017) Harari points out that algorithms will use this data to potentially know people better than themselves and thus open the door for more malicious manipulations (Harari, 2018). These algorithms can even serve to support and perpetuate flawed systems such as biased policing (*Predictive Policing Algorithms Are Racist. They Need to Be Dismantled.*, n.d.).

But these data sets can be harmful to individuals on their own, outside of the context of Big Data. The risk of an individual being identified in a dataset or reidentified in a partially anonymized dataset remains (Narayanan & Shmatikov, 2008). The question is then how can society continue benefit from the spatial analysis of social media data while ensuring the privacy of those individuals who are contributing the data in the first place?

1.2 Objectives

This thesis will focus on the issues surrounding reidentification from an individual standpoint. The objective is to perform a case study on the use of spatial social media data that has been processed into a privacy aware data structure. The structure in this case study uses HyperLogLog (HLL), a cardinality estimation algorithm (Flajolet et al., n.d.). The use of HLL will be compared with other privacy approaches and also examined as a privacy technique in conjunction with others in order to achieve what is termed “Privacy by Design” (Cavoukian, n.d.). This thesis will attempt to answer the following questions:

What is meant by privacy regarding geographic information?

Can treating geographic information with this privacy aware data structure increase the level of user privacy?

Does treating this data with the privacy aware structure allow for the same quality of subsequent visualizations for social media research?

Can the difference in privacy level be measured or quantified?

Which set operations can be carried out on the HLL shards and what are the effects on the resulting data?

What are the benefits and disadvantages to this database structure?

What are the limitations of the HLL structure and its applications?

1.3 Structure

This thesis is divided into five sections. The first section introduces the topic by giving the motivations and objectives for the research as well as outlining the structure that the thesis will manifest. The second section provides the background context and related research topics. The intersection of these three topics: privacy, social media analytics, and HyperLogLog, is what is being investigated. The third section provides the method in which the research is to be carried out by describing the data and workflow that will be used to perform a case study as well as the criteria against which the results of the case study will be evaluated. The fourth section is the case study itself which seeks demonstrate the privacy aware analysis of spatial social media data with series of visualizations. The fifth section describes the use-case of this type of spatial analysis as well as evaluates the results of the case study. This section also details the advantages and disadvantages of this method of privacy aware analysis of spatial social media data as well as the ease with which it can be performed. The sixth and final section provides a summary and conclusion about the results of the thesis as well as an outlook for future research.

2 Background

This thesis exists at the intersection of three fields of research. The first field is the spatial analysis of social media data, often in the form of cartographic visualizations. The second is the concept of privacy, geoprivacy, and privacy aware technologies. This field deals with the sometimes-murky definition of privacy, the quantification and qualification of levels of privacy, and the uniqueness of geoprivacy. The third field is research around the HyperLogLog algorithm. HyperLogLog has yet to be extensively researched for its potential to enhance privacy in cartographic visualizations of spatial social media data.

2.1 Social Media

Social Media is defined by Merriam-Webster as “Forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (such as videos)” (*Definition of SOCIAL MEDIA*, n.d.). Social media is a recent trend that coincides with the digital age and allows an unprecedented level of interconnectivity and data generation. This relatively recent increase in volume and velocity of social media data has been accompanied by a corresponding increase in interest in harvesting this data for purposes ranging from research to marketing (Chen et al., 2017). This personal data can be easily duplicated and shared and therefore can threaten the privacy of users (Malhotra et al., n.d.)

2.1.1 Location Based Services

Location based services (LBS) are services that leverage the location context of the user in order to provide increased value to the service, or in some cases, the service can only be provided with this location information (Hauser & Kabatnik, 2001). Some examples of LBS are ride sharing applications, weather information applications, and mobile service providers providing a user’s position to emergency services in the case of a crisis, among many others. It is noted that since a mobile device now function as a tool that many people have on their person at all times and that the diversity and volume of the data created and transmitted has increased as well, that a unique impression individual users would be quite simple to generate in a surveillance scenario (Hauser & Kabatnik, 2001).

2.1.2 Location Based Social Networks

Location based social networks (LBSN), also called geosocial networks, are similar to LBS in that location is leveraged to increase the value of the service, but unique in that user's voluntarily reveal their location data to increase their enjoyment of the service and to create a better online image of themselves (Kounadi et al., 2018). Some examples of LBSN are Twitter, Flickr, and Instagram. Similar to LBS, LBSN continue to generate a massive quantity of PII, some of which also contains sensitive location information. This data, although on some level provided voluntarily by the users themselves, cannot be called VGI. VGI must come from some sort of concerted data gathering effort in which the data contributors themselves are aware of the ultimate purpose of their data (Gómez-Barrón et al., 2016). It is noted that in a LBSN setting the onus of data sharing should not be unjustly foisted upon the LBSN users by hiding their agreement in the terms and conditions (Kounadi et al., 2018).

2.1.3 Analysis of Spatial Social Media Data

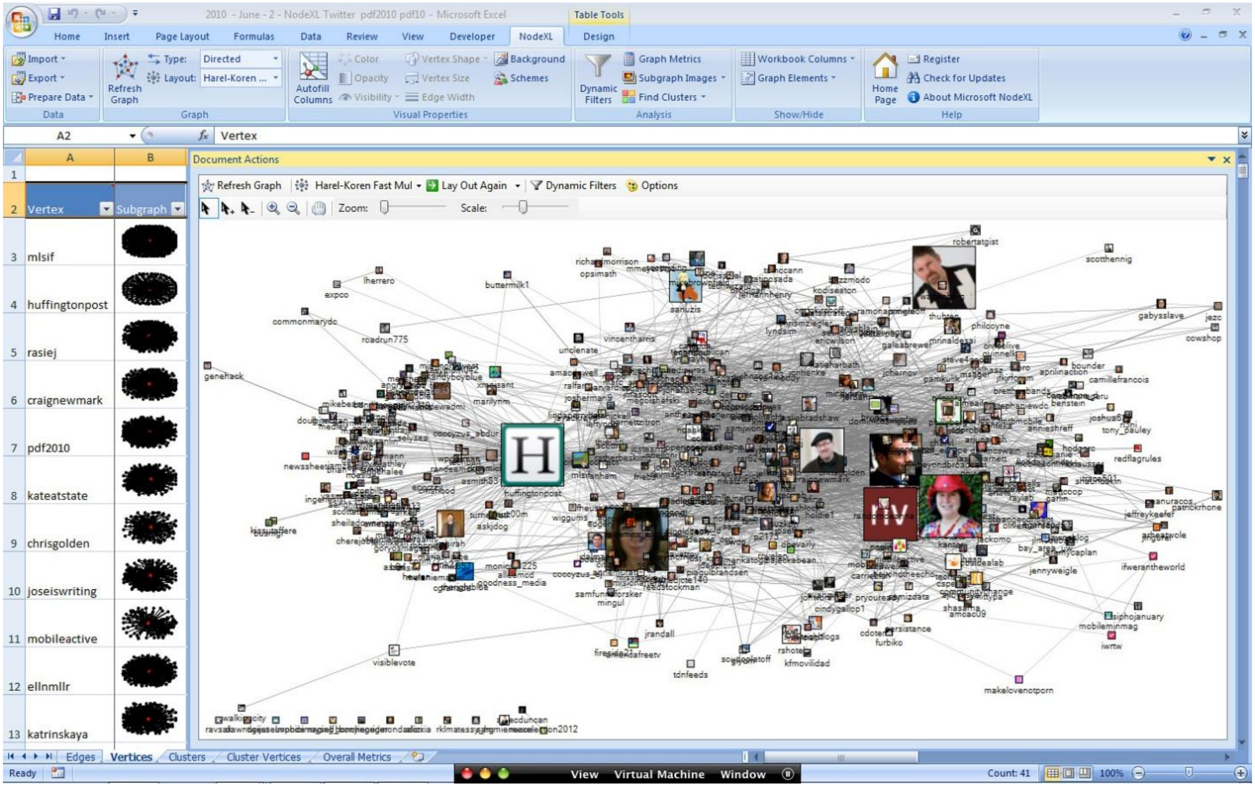
Social media companies log and store much of or all the data generated on their platforms and much of this social media data is available or accessible. Therefore, there is an interest in analyzing these massive data sets to discover underlying patterns. Spatial social media data is a specific type of social media data that includes a geographical reference. The granularity of the geographical information can vary from fine granularities like precise coordinates or addresses, to courser granularities like city, state, and country identifiers. Chen et al. (2017) classify social media data into three sub-entities, network information, geographic information, and text and content (Figure 1). Another classification uses four facets: Social, Topical, Spatial, Temporal (Dunkel et al., 2019).

Taxonomy of social media data		
Network	Geographic Information	Text and Content
• People's follower network • Messages' diffusion network • People's reposting network	• GI diffusion network • Spatial temporal event distribution • Movement trajectory	• Keywords • Topic • Sentiment

Figure 1 Taxonomy of Social Media Data

Note. Adapted from (Chen et al., 2017)

Shown in Figure 2 is an example of a network visualization of twitter users who posted the term “PDF2010” (Personal Democracy Forum 2010) on Twitter (Figure 2). The users are scaled according to the number of followers they have and connected together based on who follows whom (Marc Smith, 2010). These kinds of visualizations leverage social media networks to understand the connections between users.



Type: Directed

Layout: Harel-Koren ...

A2 Vertex

	A	B
1		
2	Vertex	Subgraph
3	mlsif	
4	huffingtonpost	
5	rasiej	
6	craignewmark	
7	pdf2010	
8	kateatstate	
9	chrisgolden	
10	joseiswriting	
11	mobileactive	
12	ellnmllr	
13	katrinskaya	

Ready

Count: 41 100%

Figure 2 Network Visualization of Twitter Users Who Mentioned PDF (Marc Smith, 2010)

Shown in Figure 3 is a geographic visualization showing which cities in the USA received the most visitors in a period of time. This visualization employs data collected from Flickr to assess which users have travelled a certain distance from their homes (Mao, 2015). These types of visualizations leverage the geographical information of the posts and their users to understand spatial dynamics.

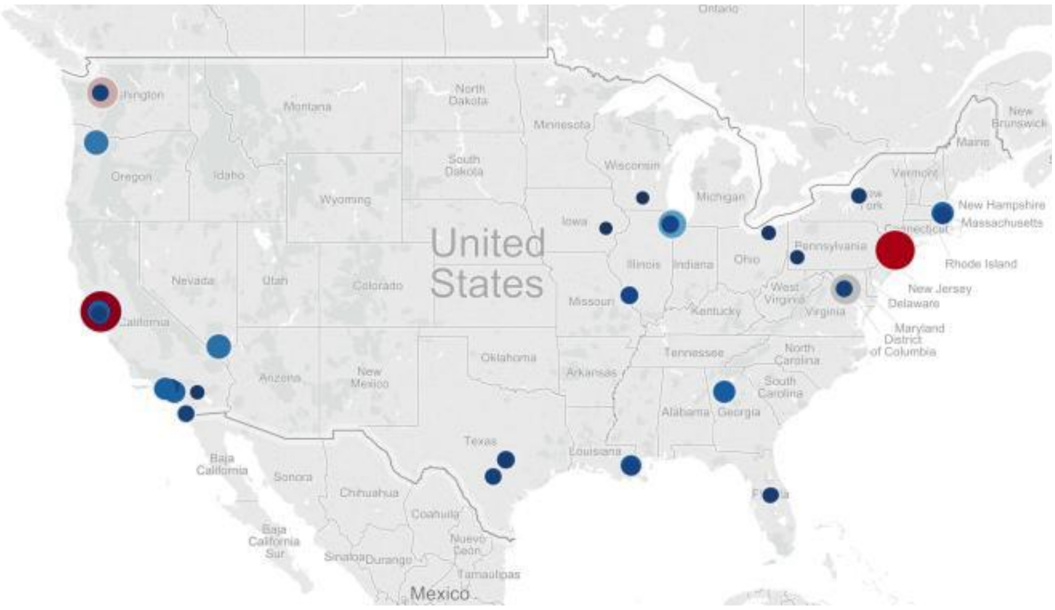


Figure 3 Filtered weighted map of the year 2012 and 2013 most visited places in US derived from the YFCC100M dataset

Note. Reprinted from (Mao, 2015)

And shown in Figure 4 is a visualization of sentimental language of twitter users on the their feelings towards the word infidel and their feelings towards the terrorist organization ISIS (Figure 4) (Regian & Noever, 2017). These types of visualizations leverage the fact that users willingly share their thoughts and feelings on social networks, albeit in an unstructured format, but nevertheless, these posts can be mined to understand the thoughts, feelings, and words of social network users.

These different sub-entities of social media can be combined as well to enhance the resulting analysis. The domain of social media data analytics covers a broad range of disciplines including data journalism, social sciences, and marketing, just to name a few.

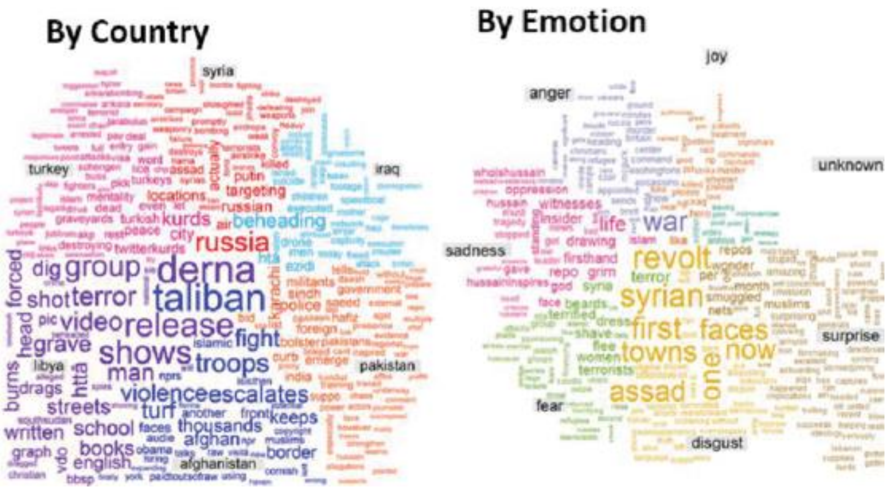


Figure 4 Sentiment Analysis of Twitter Data Concerning Terrorism Topics

Note Reprinted from (Regian & Noever, 2017)

2.2 Privacy

This massive increase in data generation and data use is a boon for those who wish to utilize the data for their own purposes, but it raises questions concerning the privacy for the individuals who are generating this data in the first place. But in order to address these questions, the very concept of privacy must first be addressed. Privacy as a concept is in fact quite nebulous and can mean anything from free thought and bodily autonomy, to the ability to live without surveillance and ownership over one's personal information (Solove, 2008). Of course, when talking about social media data, the ownership over one's personal information is explicit. But the implications are that this data could be used to harm other aspects of privacy such as personal surveillance or suppression of free speech.

One of the other problems of defining privacy comes from the fact that the concept of privacy and the idea of a right to privacy are sometimes confused (Hildebrandt, 2006). The concept of privacy is relational because privacy inherently means the relationship of an individual to others (Hildebrandt, 2006). This means that any discussion of the concept of privacy must take into consideration the context surrounding the individual whose privacy is in question. The right to privacy is something protected or ensured legally and therefore any law that enshrines this right must inherently deal with the nebulous nature of privacy itself. Because of this nebulousness, any organization that simply adheres to the regulatory framework already in place can easily find ways around these privacy protections; therefore, any organization should commit to privacy as a value and incorporate "privacy by design" into their systems (Cavoukian, n.d.).

Since this thesis deals with privacy from the perspective of analysis of social media data, a definition of privacy must be selected in order that any meaningful discussion of the privacy of the visualizations created in the case study can be discussed. The definition of privacy given in the dictionary is “the quality or state of being apart from company or observation or freedom from unauthorized intrusion” (*Definition of PRIVACY*, n.d.). This general definition would imply that any social media data with PII would violate the privacy of the individual. But there is a certain individual sense of privacy one may feel walking around a large city, a type of anonymity that comes from being in a larger group of anonymous persons (Hildebrandt, 2006). This concept is taken one step further and quantified with the property of k-anonymity. K-anonymity is a characteristic of data wherein the value k is equal to the number of other individuals in the data set who an individual cannot be differentiated from minus one (Equation 1) (Samarati & Sweeney, n.d.).

$$\text{k-anonymity} = k - 1$$

Equation 1 k-anonymity

K-anonymity can be increased through the data preparation techniques known as aggregation, generalization, and suppression (Samarati & Sweeney, n.d.). Consider the following fictional table with raw information about a street block and its residents (Table 1). Given the personal information present on for each individual and assuming the street where they lived were to be known, this table would not satisfy k-anonymity.

Table 1 Fictional Personal Information Pertaining to a Neighbourhood

Note adapted from (Samarati & Sweeney, n.d.)

Name	**Nationality**	**Sex**	**House Number**	**Age**
John	British	Male	1	21
Emma	British	Female	3	22
Brad	British	Male	3	21
Viktor	German	Male	2	35
Felix	German	Male	3	36
Franziska	German	Female	2	25
Marco	Italian	Male	1	31
Maria	Italian	Female	1	23
Sara	Italian	Female	2	23

This same information could then be suppressed to increase the k-anonymity value of the data set (Table 2)

Table 2 Suppressed Personal Information

Note adapted from (Samarati & Sweeney, n.d.)

Name	Nationality	Sex	House Number	Age
*	British	Male	1	*
*	British	Female	3	*
*	British	Male	3	*
*	German	Male	2	*
*	German	Male	3	*
*	German	Female	2	*
*	Italian	Male	1	*
*	Italian	Female	1	*
*	Italian	Female	2	*

But, if for example, some information about the age were necessary for the application of the data set, generalization could abstract the age into a class of age ranges (Table 3).

Table 3 Suppressed and Generalized Personal Information

Note adapted from (Samarati & Sweeney, n.d.)

Name	Nationality	Sex	House Number	Age
*	British	Male	1	20-30
*	British	Female	3	20-30
*	British	Male	3	20-30
*	German	Male	2	30-40
*	German	Male	3	30-40
*	German	Female	2	20-30
*	Italian	Male	1	30-40
*	Italian	Female	1	20-30
*	Italian	Female	2	20-30

And if the application data were focused only on the counts or cardinality of the information and not the specific tuples themselves, aggregation is possible to anonymize the individuals by grouping them based on the relevant data. For example, if this data's purpose was to know how many people were living in each address, the data could be aggregated just as a count (Table 4)

Table 4 Aggregate Data by Addresses

House Number	Number of Residents
1	3
2	3
3	3

Differential privacy is one example of a data anonymization concept that has been widely adopted as “the flagship data privacy definition (Desfontaines & Pejó, 2019). One way in which differential privacy is achieved is by injecting noise into the data set to protect individuals (Dwork, 2008). Google currently uses differential privacy to collect information and has published it’s libraries for others to use (“Google Wants to Help Tech Companies Know Less About You,” n.d.) A simple example of injecting noise would be in a private survey where user’s reveal information such as drug use with a simple true or false. In this example, 50% of the answers would be recorded accurately, and the other 50% of these answers would be subject to another 50% chance of being recorded as true or false. The figure used to illustrate this example uses a coin toss as stand in for a simple noise function (Figure 5)

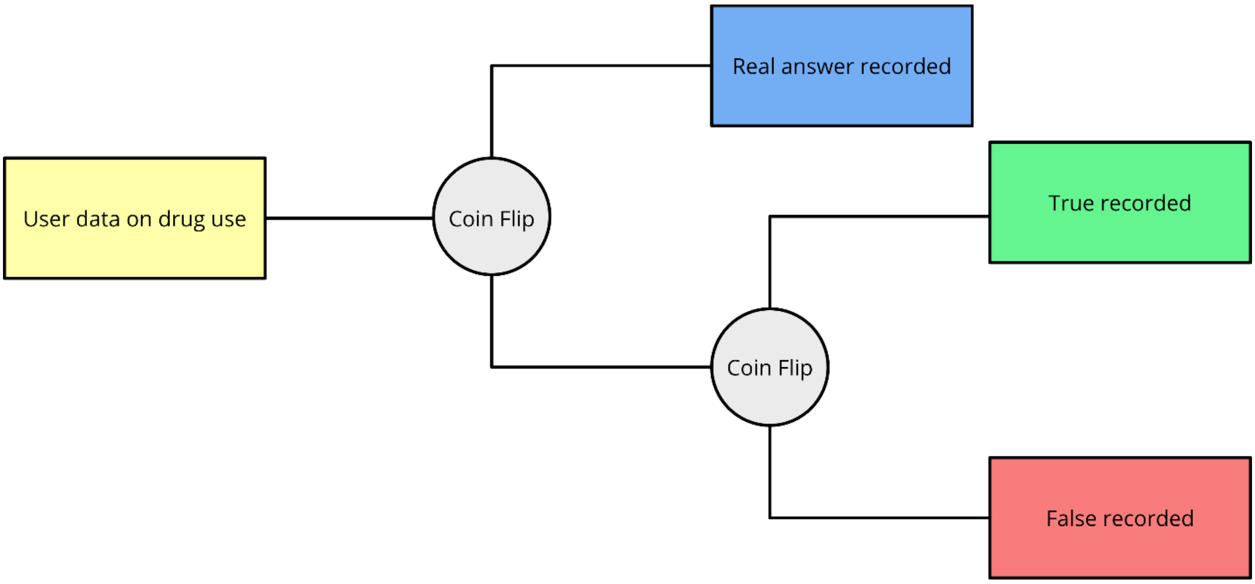


Figure 5 Differential Privacy

The main disadvantage to differential privacy is the necessity to build a new technique for each application which is a disadvantage when dealing with massive data sets from LBSN sources (Löchner et al., 2019). While the example in Figure 5 works well in a scenario with Boolean values, injecting noise into a string value makes no practical sense and therefore a specific technique would be required.

2.2.1 Big Data and Privacy

Big data is a topic of growing interest in many fields such as physics, sociology, and political science (Boyd & Crawford, 2012). Big data refers to the massive amount of data generated in today's world from the use of smart phones, social media, and other technology products. Considering enormous scope of big data and that this thesis is in the field of cartography, a definition of big data suitable for cartographers that focuses on the desire to map human patterns is "Big Data is a large dynamic data set created by or derived from human activities, communications, movements, and behaviours" (Tsou, 2015).

Although big data has the possibility to advance many disciplines, the promise of big data does not come without trepidation. Boyd and Crawford pose the question of whether big data will simply increase the quality of goods, services, and life, or whether it could be turned on people to invade their privacy and subsequently oppress them (Boyd & Crawford, 2012). It is also a possibility that both predictions can come to fruition. For example, as more people begin to trust big data algorithms with their data to help make important decisions or to understand patterns, there may be a growing incentive, or simple apathy, with the sharing of this data (Harari, 2018). This potential puts the onus on the scientific community to examine the ethical considerations surrounding big data and personal privacy. For example, a social media company that publishes its data must not only be sure to properly anonymize its data set, but also that the risk of reidentification of individual user's is understood and proper steps are taken to mitigate it (Boyd & Crawford, 2012). The types of reidentification risk have been classified as a journalism risk, where no specific individual is targeted, and a prosecutor risk where a specific individual is targeted (Wan et al., n.d.). For example, Netflix released a large, anonymized data set containing users' film ratings. This data set was successfully deanonymized by using the Internet Movie Data Base as a secondary data set, and potentially sensitive information such as political viewpoints was associated with the deanonymized individuals (Narayanan & Shmatikov, 2008).

Data privacy will continue to be a topic of growing concern. Both the European union (*General Data Protection Regulation (GDPR) – Official Legal Text*, n.d.) and the state of California (*California Consumer Privacy Act (CCPA)*, 2018) have passed comprehensive data privacy laws recently. This is a strong indication that as technologies continue to evolve more sophisticated techniques of data capture, so too will the privacy protections need to evolve.

2.2.2 Geoprivacy

Geoprivacy is a specific subset that concerns a user's intimate location information. The idea of geoprivacy, or location privacy, contrasts with normal privacy in that it is a relatively modern concept (Beresford & Stajano, 2003). It is only with the advent of information technologies that allow the capture of fine-grained location information about users for large spans of time that geoprivacy has become a topic of concern (Beresford & Stajano, 2003). It is pointed out that "Spatial is special" because the location data of an individual can reveal a wealth of sensitive information including religious practices, place of work, and other people in their real-world social network; this is known as inferred PII (Keßler & McKenzie, 2018). Not only is personal location data a particularly sensitive type of PII, but it is also one of the easiest forms of PII to collect due to the ubiquity of devices able to capture an individual's geolocation. Geoprivacy also suffers from the problem of competing goals in that maintaining strict location privacy i.e. not sharing your location is incompatible with wanting to benefit from LBS technology (Beresford & Stajano, 2003). For example, in order to benefit from the use of a weather prediction application, a user must either allow their location to be automatically known or else, at the very minimum, provide the name of a city (Keßler & McKenzie, 2018).

To protect geoprivacy, a number of techniques have been researched or implemented. Mix zones are a technique in which user's pseudonymized identities are swapped while they are in the spatial mix zone (Beresford & Stajano, 2003). Theoretically this would make any number of individuals in the mix zone indistinguishable from each other upon their departure from the zone (Figure 6).

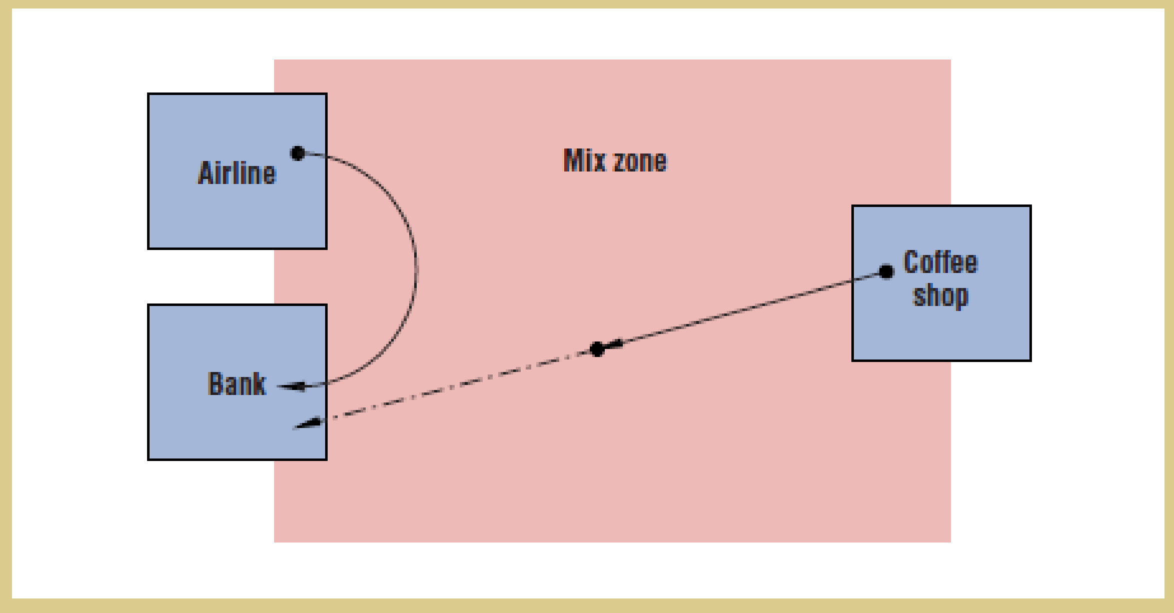


Figure 6 Mix Zones Example

Note Reprinted from (Beresford & Stajano, 2003)

Another technique known as obfuscation seeks to degrade the information quality in such a way that the individual privacy can be protected (Duckham & Kulik, 2005). Obfuscation tries to balance the LBS service provider's need to use location data to provide their service with that of an individual's desire to maintain privacy (Duckham & Kulik, 2005). An example of this technique would be increasing the granularity of a user's location so that it could not be pinpointed but an LBS such as restaurant recommendations could still deliver relevant suggestions at the scale of a city.

This also relates to one of the ideas presented by Keßler & McKenzie in their Geoprivacy manifesto that an LBS user should have the ability to control which details of their personal information are shared with the service provider to a fine level of detail (Keßler & McKenzie, 2018).

Similarly, Hauser and Kabatnik propose a privacy architecture in which a user can define which entities can receive information at which granularity and furthermore if that information can be linked to the individual's identity; this architecture is meant to function despite a lack of trust that the user feels towards the provider (Hauser & Kabatnik, 2001).

Despite the aforementioned techniques their main flaws are that whichever entity is analyzing the original data set still requires access to the raw data and that the privacy aware processed data sets cannot be updated with new data without processing the entirety of the old data once again, a severe limitation when working with LBSN data sets (Löchner et al., 2019). Furthermore, any privacy enhancing technique that shifts the burden to the user's themselves is likely to be shrouded in hard-to-understand language for the lay-users (Kounadi et al., 2018).

2.3 HyperLogLog Algorithm

HyperLogLog (HLL) is a type of probabilistic algorithm that estimates cardinality (Flajolet et al., n.d.). Cardinality is defined as the number of distinct elements in a set. Equation 2 shows the cardinality of a set containing 7 distinct elements to be equal to 7.

$$|\{a, b, c, d, e, f, g\}| = 7$$

Equation 2 Cardinality Example

HLL is an example of a solution to the count-distinct problem. The count-distinct problem is the problem of retrieving the count of distinct elements in a stream of data with duplicate elements.

2.3.1 HyperLogLog Applications

The main advantages of HLL are that it has a standard error of 2% meaning that it is incredibly accurate at estimating cardinalities, unions in HLL are lossless, and the amount of data storage is reduced significantly (Flajolet et al., n.d.). The implications of this are that HLL facilitates the streaming of large data sets that must be frequently updated. This is indeed the primary application of HLL to social media data analytics and its ability to use preserve privacy is a secondary effect that is being explored in this thesis.

The two basic set operations permitted with HLL are union and cardinality. Union allows two sets to be combined in a set union and cardinality takes any HLL shard and returns the set cardinality, or distinct count. In addition to unions, the inclusion-exclusion principle allows for intersections of HLL sets to be calculated (Equation 3). The inclusion exclusion principle is a basic concept in set logic, it is not a property of HLL.

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Equation 3 Inclusion–exclusion principle

Intersections are only meaningful in the case that the sets are of a similar size. For example, if one set has 1 billion elements and another only 10 million, the 2% error inherent to HLL would be 20 million itself, thus rendering the calculated value meaningless. In Figure 7 it is apparent that the error in the overlapping area is so large that a calculated value would be all but meaningless.

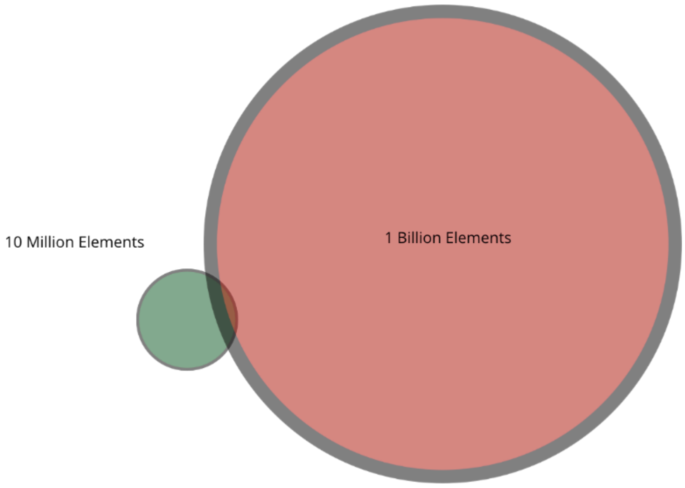


Figure 7 HLL Intersection

3 Method

This thesis will seek to answer the questions posed in section 1.2 by performing a case study. In this case study a series of cartographic visualizations will be created. These visualizations will be created from either one or both data sets being examined. In this case study a selection of “stories” will demonstrate the use of raw data and privacy aware data to create maps. These stories will then be evaluated in order to discuss the benefits and disadvantages concerning privacy and social media data analytics. The goal is to understand the specific use-cases where this privacy aware data structure could be implemented and what are its limitations.

3.1 Data

This research is carried out with the use of two parallel databases. One database can be considered raw or unprocessed. The other database is one in which the data has been processed into a privacy aware data format. Both databases are organized into a standard LBSN data model (*Summary - LBSN Structure*, n.d.).

3.1.1 LBSN Structure Data Model

The aforementioned LBSN data model is composed of **objects, bases, facets, relationships** and **overlays** (*Structure Definition - LBSN Structure*, n.d.)**. Objects** are entities from LBSN data e.g posts, users, places, and events. Each of these objects is composed of **bases** which are characteristics of the object itself e.g. title, hashtags, post creation date, etc. **Relationships** describe the interrelation of objects e.g. posts and users. All of the LBSN data is are principally categorized into four **facets**: topical, social, spatial, and temporal (Table 5) (*Structure (Bases) - LBSN Structure*, n.d.)**.** And finally, **overlays**, also called metrics, are the bases used in the context of analysis e.g. post count, user count, user days (*Metrics (Overlays) - LBSN Structure*, n.d.).

Table 5 Facets of HLL Database (*Structure Definition - LBSN Structure*, n.d.)

Facet	Description
Temporal	Related to temporal aspects of the post such as date, time, season, etc.
Spatial	Location of the post, including various granularities
Social	User identity including membership in a social group and language
Topical	User reactions including hashtags, descriptions, titles, and

In the context of this project, individual posts from the YFCC100M data set will function as the objects. And since this thesis is focusing on *spatial* social media data, the resulting case study and analysis will naturally use data from the spatial facet. That is not to say that the other facets cannot play a role in privacy aware analysis of social media data in general or that the other facets cannot be used in conjunction with each other. The other facets are simply outside the scope of the case study.

3.1.2 Yahoo Flickr Creative Commons 100 Million

The Yahoo Flickr Creative Commons 100 Million data set (YFCC100M) is a collection of 100 million Flickr posts presented in a uniform database structure (Thomee et al., 2016). The data set contains photos that were uploaded in a 10-year period between 2004 and 2014. It is also worth noting that the compiled photos all were uploaded under a creative commons license or else a non-commercial license (Thomee et al., 2016), thus allowing for the case study to be performed without any privacy conflicts (Table 6). In the LBSN data model, this raw database is categorized into the topical facet with the base representing the individual posts.

Table 6 YFCC100M Posts and Licenses

Note Adapted from (Thomee et al., 2016)

CC License	Photos	Videos
CC BY	17,210,144	137,503
CC BY-SA	9,408,154	72,116
CC BY-ND	4,910,766	37,542
CC BY-NC	12,674,885	102,288
CC BY-NC-SA	28,776,835	235,319
CC BY-NC-ND	26,225,780	208,668
Total	*99,206,564*	*793,436*

Table 7 YFFCC Record Sample

origin_id	2
post_guid	10000000
post_latlng	0101000020E6100000779D0DF967AE26401405FA449E404640
place_guid	
city_guid	12845897
country_guid	28350914
user_guid	35769202@N00
post_publish_date	08:11.0
post_body	and the number is..... TEN MILLION!! ;-)
post_geoaccuracy	city
hashtags	['clouds', 'maggiore', 'building', '10000000', 'piazza', 'bologna', 'top-v333', 'sky', 'tower', 'top-v777']
emoji	[]
post_like_count	
post_comment_count	
post_views_count	
post_title	tower in the sky (n. #10.000.000 Flickr photo)
post_create_date	14:57.0
post_thumbnail_url	http://farm1.staticflickr.com/5/10000000_106b46b078.jpg
post_url	http://www.flickr.com/photos/35769202@N00/10000000/
post_type	image
post_filter	
post_quote_count	
post_share_count	
post_language	
input_source	
user_mentions	[]
modified	25:42.3
post_content_license	2

Table 7 shows one record from the YFCC100M data set. These fields could be useful in social media data analytics. For example, the fields for hashtag, post title, date of creation and data of publication all offer valuable data. But for this data to be geographic, it must include a piece of data that offers a specific location. In this case the field titled ***post_latlng*** provides the geometry in the Well-Known Binary (WKB) data format. Worth mentioning is the ***post_geoaccuracy*** field. In this example, the post's granularity is set to city level meaning that when mapped, one cannot expect the point to be the exact location of the photo but rather within the same city where the photo was taken.

This data has been slightly processed to a standard format (*Summary - LBSN Structure*, n.d.). For example, the original YFCC100M data set has 16 levels of geoaccuracy whereas this structure only has 4 levels. These 16 levels of geoprivacy are in fact unique to Flickr. Flickr allows users to create a 'geo-fence' wherein the user places a pin and draws a radius that then corresponds to a number 1-16 (Matthew Smith et al., 2012) A summary of the unprocessed data and further explanation can be found from (Deng & Li, 2018). Important to note is that no other data set was combined to create additional PII fields in this processed data set. So, when discussing the privacy implications, these fields can be deduced or calculated simply from to the YFCC100M data set.

Besides a potential exact coordinate of where a post video or photo was captured, there remains a wide range of other types of PII in the fields of the post. For example, each post is associated with a user's globally unique identifier (GUID). In the example provided in Table 7, there is a ***post_url*** field which leads directly back to the post on the social media platform (Figure 8). Even though the geoaccuracy of this post is city level i.e. Bologna, with this extra information, namely a photograph of a specific tower, it would not be difficult to infer a finer grained location for this user. Combining this data with the date the photo was taken would allow for one to know that the user was at the Piazza Maggiore in Bologna on April 19th, 2005 at exactly 6:14 pm. The hashtags do indeed confirm this information (Figure 9)

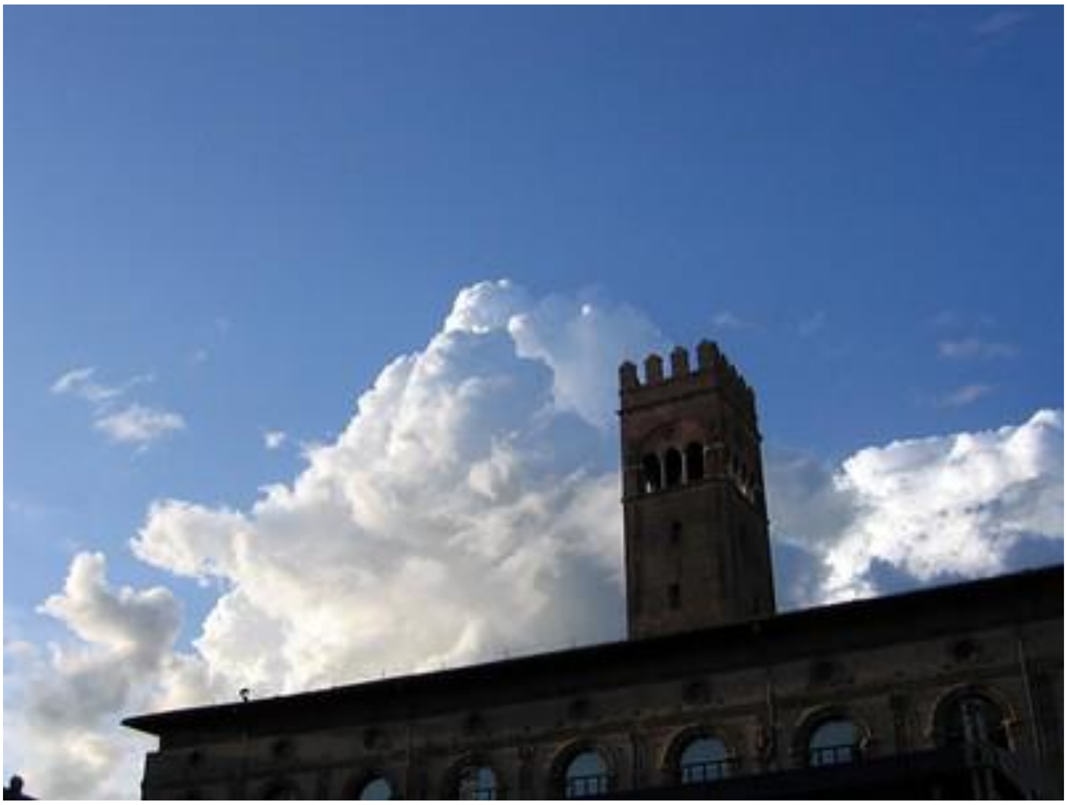


Figure 8 Photo from YFCC100M Record (Daniele, 2005)

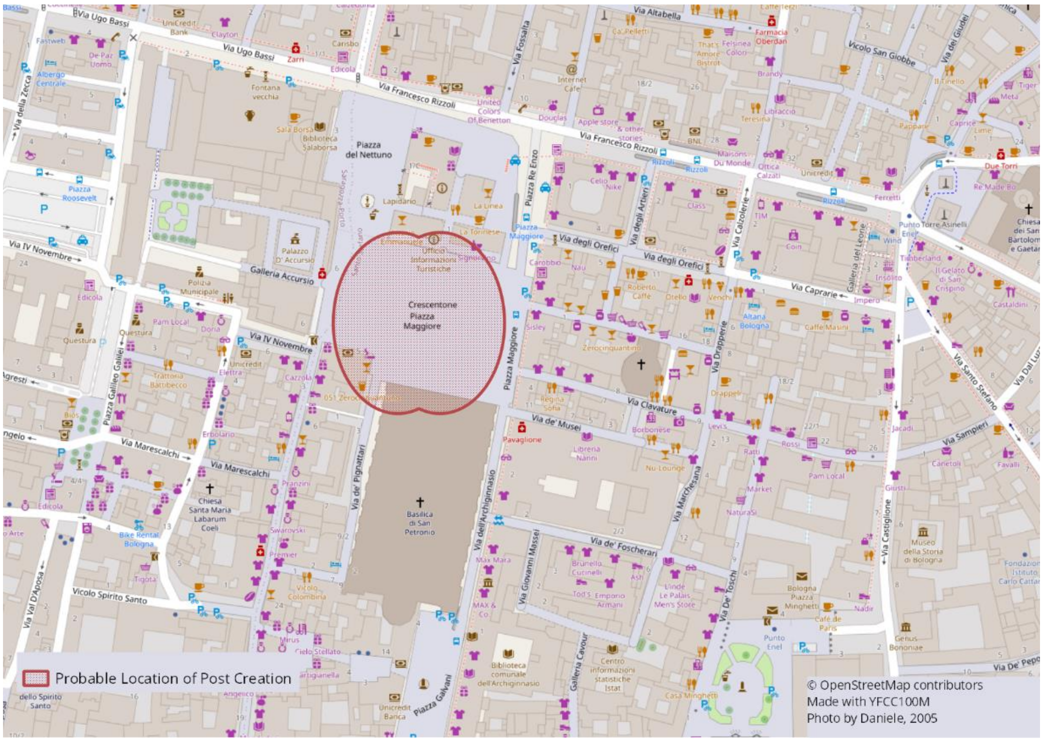


Figure 9 User Reidentified with Non-geodata

3.1.3 Privacy Aware Data Structure

The privacy aware data structure being investigated is the product of processing the YFCC100M data set. The YFCC100M data set is transformed into its privacy aware version with the use of HLL, aggregation, and a cryptographic hash function. In the context of this thesis, the processed data is categorized into the spatial facet. The bases are either user count, post count, or user days and the metrics are where the HLL algorithm is applied. Since these bases are all aggregated counts and moreover cardinalities, the HLL algorithm is suited to estimate those values. In practice, these counts are provided with a location. These locations could range from the granularity of an exact latitude and longitude to a city level or country level aggregation depending on the method by which the data is processed and the application.

It is worth emphasizing this structure does not provide an alternative to the raw data structure. That would require a similar amount of PII to be stored in the privacy aware structure. In order for one to make use of this structure, one must know what they intend to analyze before processing the data. An example of a return query from the privacy aware database with cardinalities can be found in Table 8. This data has been processed from the YFCC100M with no filters applied in regard to reactions, dates, etc. The original data is simply aggregated to location with 3 metric overlays.

Table 8 Privacy Aware Data Example

Latitude	Longitude	Post count	User days	User count
-86.0532	39.79121	6	1	1
-84.0068	164.405	7	1	1
-80.3625	20.563	91	18	2
-79.8809	-111.287	6	3	1
-78.9568	-44.2047	33	19	1

3.2 Workflow

At its core, this case study is investigating a workflow wherein privacy is increased at each step (Figure 10). The first step is the YFCC100M data set being organized into the LBSN data structure. Maps created from this data structure are represent a method of spatial social media analysis wherein no attention is paid to the protection of data privacy. This raw database can then be processed into the privacy aware structure described in Table 8. During the creation of the privacy aware database, a number of filters can be applied to the data. For example, one can filter for keywords in the text i.e. reactions, location of the original post, or a date range in which the post was taken. In fact, it is vital that one filters the data properly because if the data is simply processed into the HLL database without any context as to what is being analyzed, there is no way to retrieve that information to update the HLL database. After data is processed into the HLL database, it can always be further aggregated depending on necessity for visualization as well as a way to further increase privacy. The opposite is not true. Once the data has been processed, it can no longer be disaggregated.

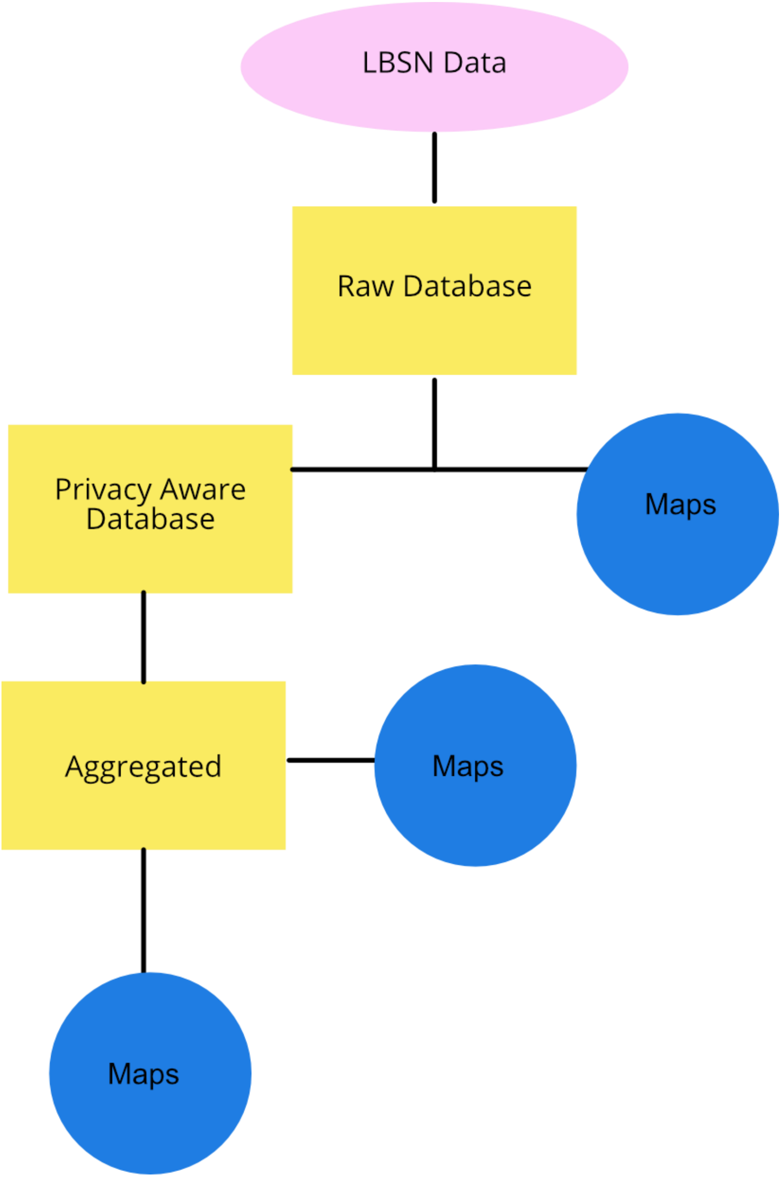


Figure 10 Workflow Diagram

3.2.1 Raw Data

The raw database can be accessed with a simple SQL query and mapped. Shown below is the SQL to query every post created on March 8th, 2012 (Code 1).

```
SELECT *, ST_X(t.post_latlng) as Long, ST_Y(t.post_latlng) as Lat
FROM topical.post as t
WHERE  t.post_geoaccuracy = 'latlng'
   AND t.post_create_date >= '2012-03-08 00:00:00.000'
   AND t.post_create_date < '2012-03-09 00:00:00.000'
```

Code 1 SQL Query for Raw Database

This data can of course be mapped to simply visualize where posts where created on this date. Figure 11 demonstrates this data in North America and also provides a heat map visualization.

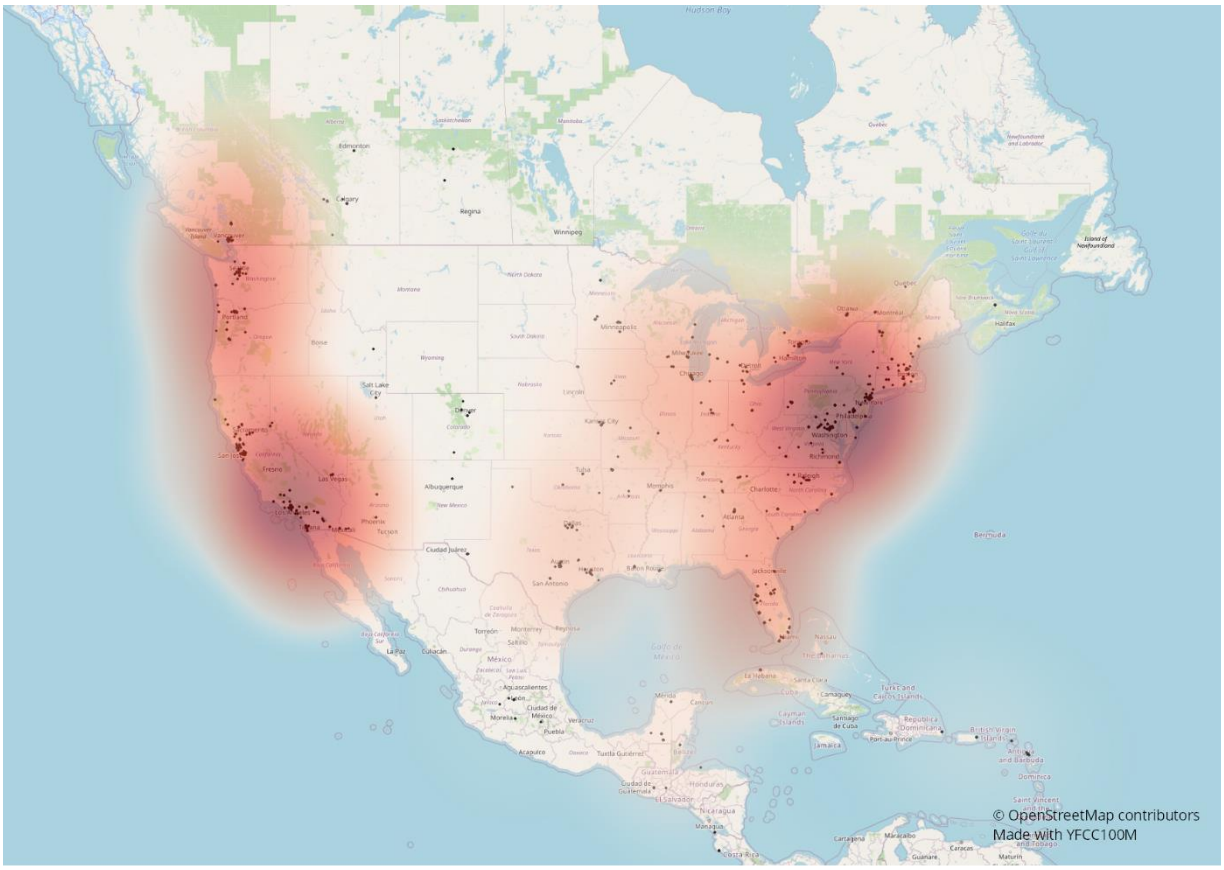


Figure 11 Raw Database Query Mapped

3.2.2 Transformation of Data

The method to transform the data from the raw database into the HLL database follows these steps: 1) one must create the common schema discussed in section 3.1.1. 2) one must use a cryptographic hash function. Below is an example of a cryptographic hash function (Code 2). Without the cryptographic hash function, the privacy of the individuals cannot actually be preserved (Desfontaines & Pejó, 2019). 3) one must select which data is to be processed. The example provided queries the same data i.e. all post created on March 8th 2012. Code 3 demonstrates filtering the data but not yet transforming it. It is important that both a unique user id as well as the post creation date are selected in this step. These two pieces of data are vital to the creation of the HLL metrics user days and user count.

```
/* Produce pseudonymized hash of input id with skey
 * - using skey as seed value
 * - sha256 cryptographic hash function
 * - encode in base64 to reduce length of hash
 * - remove trailing '=' from base64 string
 * - return as text
 */
CREATE OR REPLACE FUNCTION
extensions.crypt_hash (id text, skey text)
RETURNS text
AS $$
    SELECT
        RTRIM(
            ENCODE(
                HMAC(
                    id::bytea,
                    skey::bytea,
                    'sha256'),
                'base64'),
            '=')
$$
LANGUAGE SQL
STRICT;
```

Code 2 Cryptographic Hash Function

Note Copied from (*A Basic Visual Analytics Example - LBSN Structure*, n.d.)

```
CREATE MATERIALIZED VIEW mviews.spatiallatlng_raw_marcheighth AS
    SELECT  extensions.crypt_hash(t1.post_guid, 'samplekey') as
"post_guid",
            ST_Y(t1.post_latlng) As "latitude",
            ST_X(t1.post_latlng) As "longitude",
            extensions.crypt_hash(t1.user_guid, 'samplekey') as
"user_guid",
            to_char(t1.post_create_date, 'yyyy-MM-dd') as "post_cre-
ate_date"
    FROM   topical.post t1
    WHERE
    t1.post_geoaccuracy = 'latlng'
AND t1.post_create_date >= '2012-03-08 00:00:00.000'
AND t1.post_create_date < '2012-03-09 00:00:00.000'
```

Code 3 Filtering Data

4) The raw database must be connected to the HLL database and destination table must be prepared for the pre-filtered raw data to be processed into with the same aforementioned schema. 5) The pre-filtered raw data must then be processed into the prepared destination table (Code 4). In this step functions not inherent to the database management software are utilized. They are part of an extension called Citus that allows for HLL to be utilized in PostgreSQL specifically (*Citusdata/Postgresql-Hll*, 2013/2020). After this step, the data is able to be mapped or further aggregated as demonstrated in the workflow diagram (Figure 10).

3.2.3 HLL Database

The HLL database can then be accessed with an SQL query (Code 5) after the process outlined in section 3.2.2 is successfully completed. This SQL query also utilizes functions from Citus. This is because the metric values are stored in what are known as HLL shards (Figure 12). These shards are what allow for the data to undergo lossless unions. The Citus functions that will be used are union, union aggregate (for more than two values), and cardinality. The query returns a record with coordinates and a metric overlay for each coordinate: post count, user count, and user days. Figure 13 shows the results of this query mapped. Figure 13 and Figure 11 comparable but one major difference is that in order to generate the heat map with the HLL data, one must use one of the metrics as a weight for each point, whereas the raw data is not yet aggregated. This is a central difference when comparing the data sets in the coming case study.

$$\backslash\backslash x128b7fa71702334490600f57769b7f0c8ee4f6$$

Figure 12 HLL Shard in Binary String Representation

```
INSERT INTO spatial.latlng_marcheighth(
        latitude,
        longitude,
        user_hll,
        post_hll,
        date_hll,
        latlng_geom)
    SELECT  latitude,
            longitude,
            hll_add_agg(hll_hash_text(user_guid)) as user_hll,
            hll_add_agg(hll_hash_text(post_guid)) as post_hll,
            hll_add_agg(
                hll_hash_text(user_guid || post_create_date)
                ) as date_hll,
            ST_SetSRID(
                ST_MakePoint(longitude, latitude), 4326) as latlng_geom
    FROM extensions.spatiallatlng_raw_marcheighth
    GROUP BY latitude, longitude;
```

Code 4 Inserting Pre-Filtered Data into HLL Table

```
SELECT latitude,
           longitude,
           hll_cardinality(post_hll)::int as postcount,
           hll_cardinality(date_hll)::int as userdays,
           hll_cardinality(user_hll)::int as usercount

    FROM spatial.latlng_marcheighth
```

Code 5 SQL Query for HLL Database

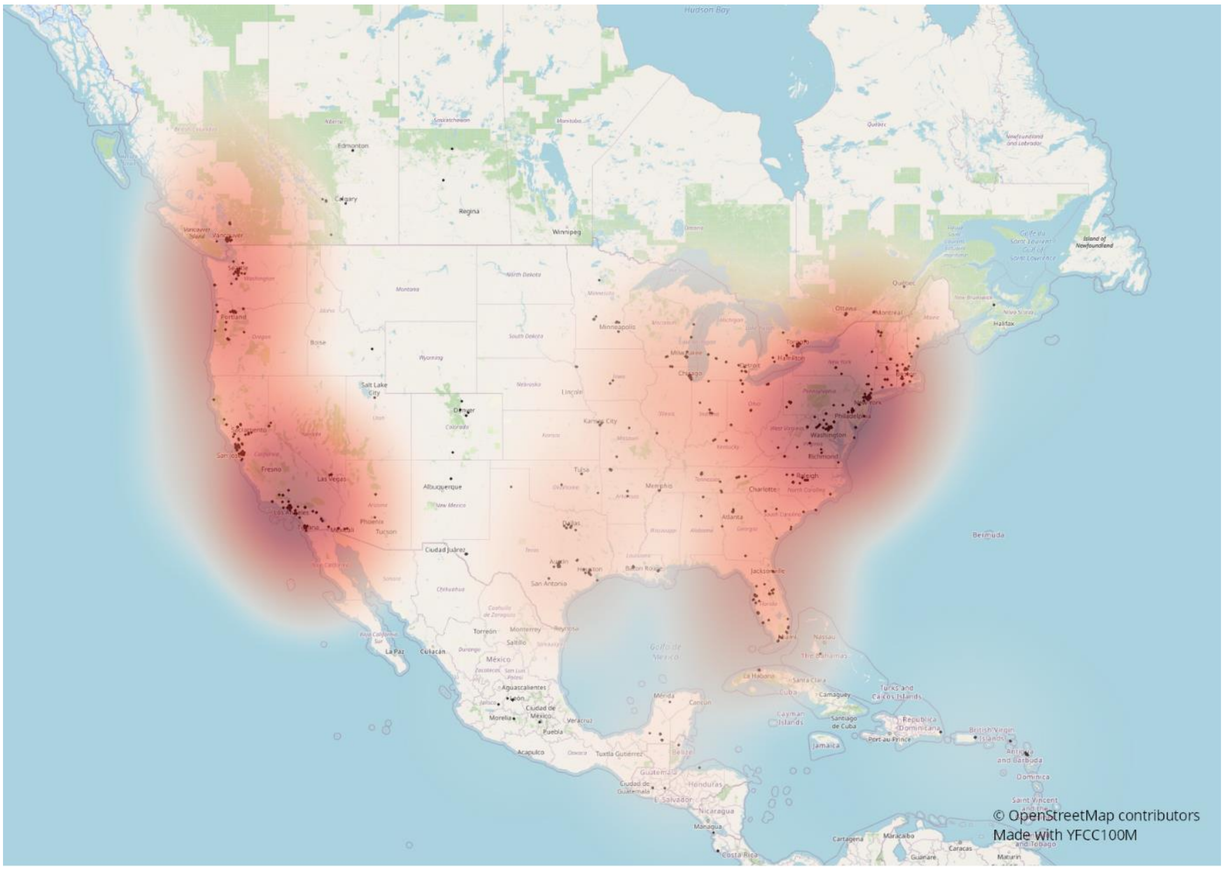


Figure 13 HLL Database Query Mapped

3.3 Evaluation

The following case study seeks to demonstrate the use of the privacy aware data structure for spatial social media analytics. Using techniques outlined in section 3.2 (workflow) maps will be created and subsequently evaluated. These evaluations will remain qualitative. The main criteria against which these maps will be judged is whether or not an individual's geoprivacy is maintained or increased by the use of the privacy aware data structure and whether or not the visualization itself is comparable in quality and use case to a comparable visualization made with the raw data set.

3.3.1 Definition of Geoprivacy

A precise definition for geoprivacy must be provided in order for the aforementioned evaluations to be carried out. Since privacy and geoprivacy are multifaceted topics, this definition is not meant to be a discipline spanning standard but rather a useful definition to apply in this particular case study. For example, the concept of geoprivacy being an individual's ability to determine which personal geographic information is shared and to what granularity (Keßler & McKenzie, 2018) does not provide a useful criteria with which to evaluate the case study visualizations. All of the data in this case study was permitted for use under creative commons licenses. Moreover, in the case of LBSN company analyzing their customers' geographic data, the terms, and conditions they provide are often convoluted, and therefore it is not in the spirit of geoprivacy by which they obtain consent of the users. This definition shifts the perspective and burden of privacy to the user themselves whereas the case studies are evaluating a scenario in which the users' data is already compiled into a large data set e.g YFFC100M.

From the perspective of social media analytics, the most useful definition against which these maps can be evaluated refers to whether or not a single user can be identified from the data set. These kinds of privacy risks, already discussed in section 2.2.1, are known as prosecutor and journalism risks (Wan et al., n.d.). Therefore, the definition of geoprivacy that will be used for evaluation is a single user's ability to remain anonymous within the publicly available dataset.

4 Case Study

This case study is composed of four sections that demonstrate how HLL can be used to make cartographic visualizations and a final section that uses the techniques described in the previous four sections.

4.1 Cardinality

As mentioned previously, cardinality refers to the unique number of items in a set. In the case of the HLL database, there are three pre-processed fields for cardinality. Those fields are post count, user days, and user count. Post count refers the number of unique posts at a given set of coordinates. User days refers the cumulative number of unique days and unique users posted at a given set of coordinates. For example, if one user posts 3 times on 3 separate days, that would be 3 user days. If another user has posted 3 times but on 2 separate days, that is 2 user days. These values are summed at each coordinate. The final metric is unique users that have posted at a given coordinate. These three metrics do not preclude other meaningful metrics from being created, these three metrics are simply the typical metrics described in the privacy aware database structure (*Metrics (Overlays) - LBSN Structure*, n.d.)

It is important to note that cardinality can still be measured with the raw data set. Code 6 shows an SQL query to select only those posts that are within an envelope that encompasses the city of Grand Rapids, Michigan in the United States. The result of this query is mapped in Figure 14 with symbols sized proportionally to the number of posts at each location.

```
SELECT count(t1.post_latlng),
    ST_X(t1.post_latlng), ST_Y(t1.post_latlng)
FROM topical.post AS t1
WHERE t1.post_latlng
        && -- intersects
        ST_MakeEnvelope (
         -85.550477, 43.064883, -- bounding
         -85.860819, 42.849689, -- box limits (Grand Rapids)
         4326)
 GROUP BY t1.post_latlng
```

Code 6 Cardinality from Raw Database, (Grand Rapids, Michigan)

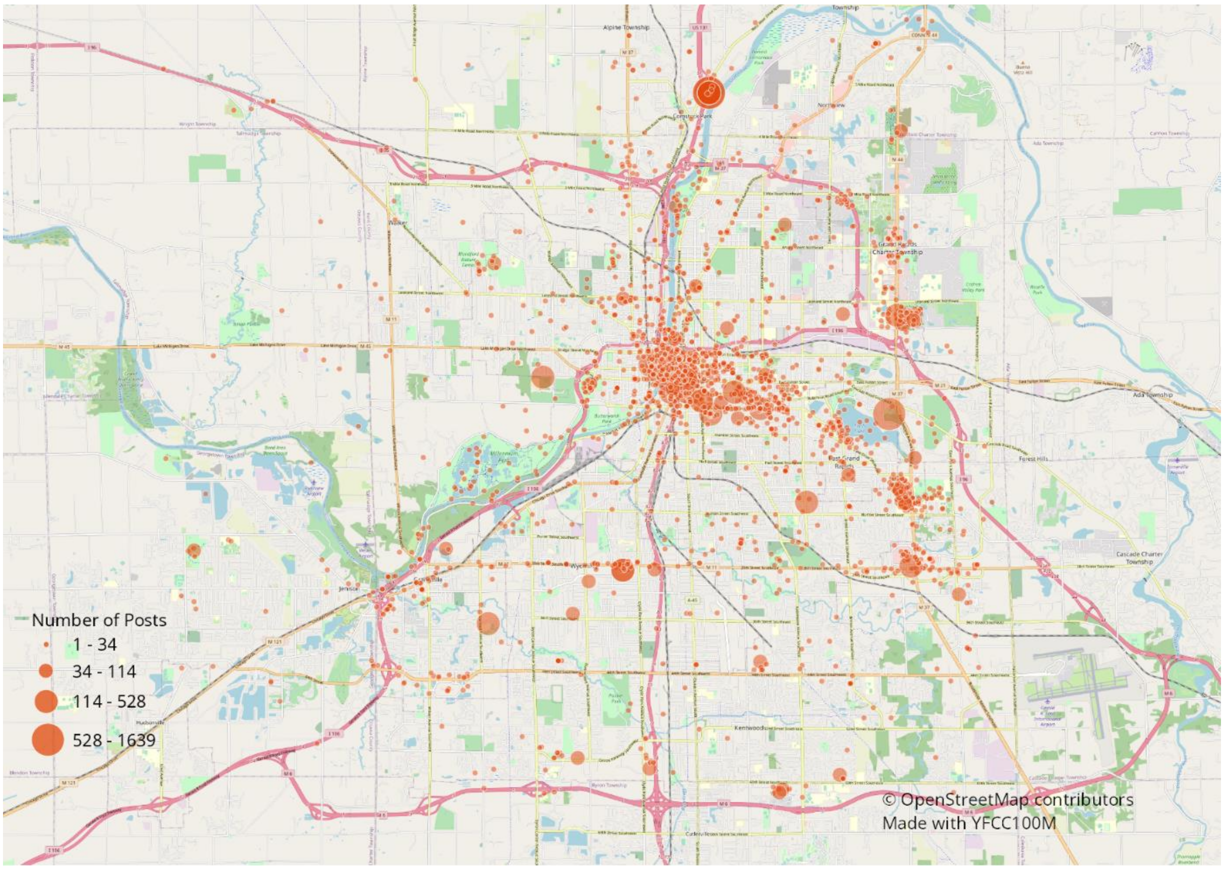
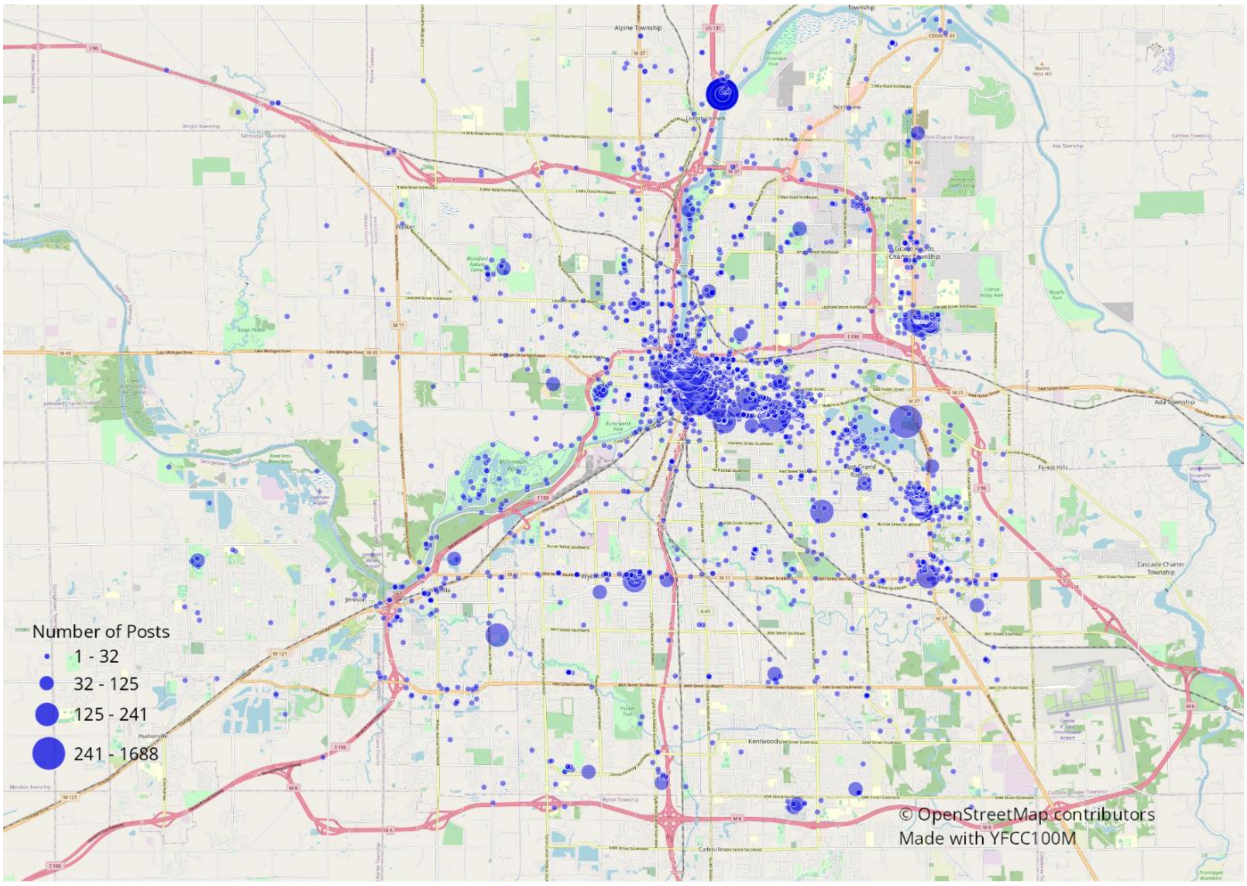


Figure 14 Cardinality from Raw Data, (Grand Rapids, Michigan)

Code 7 shows queries the exact same data but from the HLL database. The main difference in this query is that it returns the post count in the HLL format that must be converted using the cardinality function. The result of this query is mapped in Figure 15 with symbols sized proportionally to the number of posts at each coordinate pair.

```
SELECT latitude, longitude, hll_cardinality(post_hll)::int
FROM spatial.latlng t1
WHERE t1.latlng_geom
    && -- intersects
    ST_MakeEnvelope (
     -85.550477, 43.064883, -- bounding
     -85.860819, 42.849689, -- box limits (Grand Rapids)
     4326)
```

Code 7 Cardinality from HLL Database, (Grand Rapids, Michigan)



© OpenStreetMap contributors
Made with YFCC100M

Figure 15 Cardinality from HLL Data, (Grand Rapids, Michigan)

4.1.1 Comparison of the Two Data Sets

From simple visual comparison, not much difference can be detected between the results in both Figure 14 and Figure 15. A quantitative comparison of some of the statistics can be seen in Table 9. It is clear that for most applications this marginal difference would not affect the quality of the subsequent visualization.

Table 9 Comparison of Data Sets

	Raw	HLL	Percent Difference
Total Records	4518	4513	0.1%
Total Unique Posts	24110	23716	1.6%
Maximum Value	1639	1688	3.0%

4.1.2 Reidentification

One key difference between the two data sets is not only the ability to extract PII from the raw data set, but the ease with which one can extract this information. In Figure 16 one user was random selected from the data set and all of their posts within the envelope encompassing Grand Rapids, Michigan were mapped. The geographic information itself already provides a sensitive piece of information. The group of posts within the blue circle are located at a religious university. A reasonable inference would be that this user is a member of this particular religious denomination. This is potentially sensitive information that, in an LBSN context where the user had not provided explicit permission, could be compromising to the individual. Combing the geographic information with other information from the post, the red circles denote two posts that include the word wedding as well as a name of the people being married. Furthermore, one can view the photos themselves from the URL provided in the database (see Table 7). These pieces of information could easily be used to reidentify the poster and potentially others who are in the photographs.

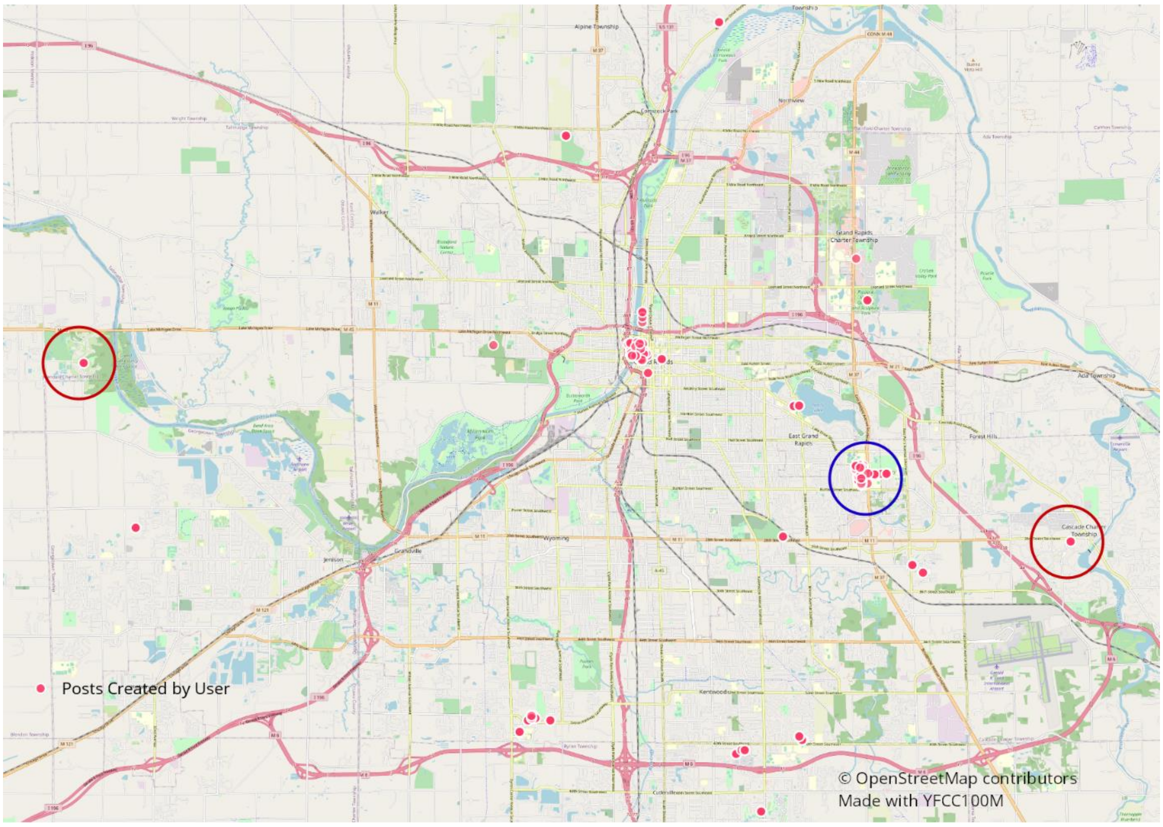


Figure 16 Reidentification of a Single User

4.2 Union & Aggregation

4.2.1 Geographic Aggregation

One of the key features of HLL is the ability to perform lossless unions. A lossless union allows for two separate HLL sets to be combined so that the cardinalities are maintained. This is in contrast to simple addition of the values. For example, if the number of users for a location is calculated on March 8th to be 6 and on June 8th to be 10, without the ability to perform a union, one can only add the two values together to equal 16. A union allows the sets to be unified without counting the same user twice.

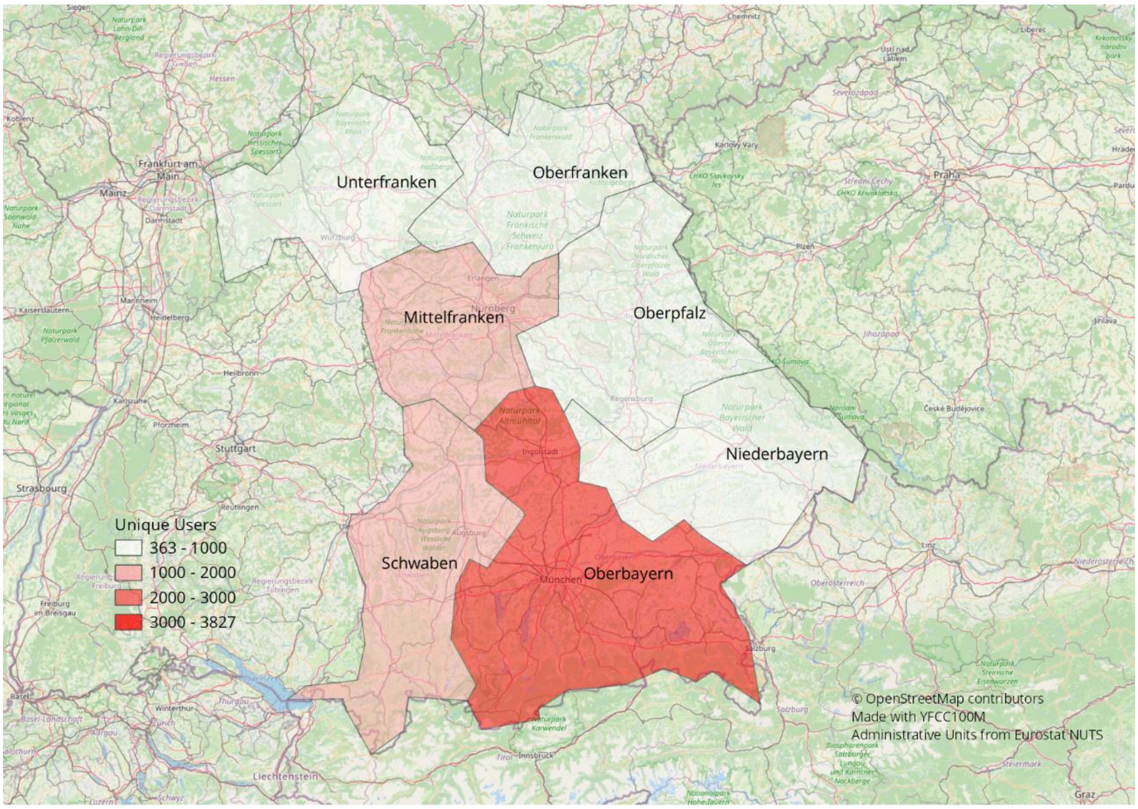
This ability aggregate through unions allows for HLL to be aggregated into courser and courser granularities. Code 8 demonstrates the aggregation of the individual point coordinates to the individual administrative regions in Bavaria using a union aggregation function. It is worth noting that this technique employs a geographic relation representing the administrative regions in the database and PostGIS to calculate the intersection. The result of this aggregation is mapped in Figure 17. Once this query is performed, one can access the data with a simple query to retrieve the metrics for each Bavarian administrative region. The results are shown in Table 10 and mapped using unique user count in Figure 17.

```
SELECT
    by.nuts_name,
    by.wkb_geometry,
    hll_union_agg(s.user_hll) AS "user_hll",
    hll_union_agg(s.post_hll) AS "post_hll",
    hll_union_agg(s.date_hll) AS "date_hll"
FROM public.bayern AS by
JOIN spatial.latlng AS s
ON ST_Intersects(by.wkb_geometry, s.latlng_geom)
GROUP BY by.nuts_name, by.wkb_geometry;
```

Code 8 Aggregating Counts to Geographic Areas, i.e. Bavarian Regions

Table 10 Metrics Aggregated on Bavarian Administrative Regions

	nuts_name	usercount	postcount	userdays
0	Unterfranken	483	25165	2772
1	Mittelfranken	1011	25305	3504
2	Schwaben	1312	29436	3961
3	Oberpfalz	430	8374	1432
4	Oberbayern	3827	140245	18274
5	Niederbayern	363	8075	1098
6	Oberfranken	456	12789	2414



© OpenStreetMap contributors
Made with YFCC100M
Administrative Units from Eurostat NUTS

Figure 17 Unique Users in Bavaria by Administrative Region

As long as the values remain in their HLL shard form (see Figure 12), HLL union operations can continue to be performed. This means that the entirety of Bavaria can be aggregated into a single unit and combined with other aggregations as demonstrated in Figure 18. The same values are returned whether or not the aggregation occurs from the point data or the already aggregated administrative region data.

```
SELECT
    bl.nuts_name,
    hll_cardinality(hll_union_agg(s.user_hll))::int AS "user_hll",
    hll_cardinality(hll_union_agg(s.post_hll))::int AS "post_hll",
    hll_cardinality(hll_union_agg(s.date_hll))::int AS "date_hll"
FROM public.yfcc_latlng_deutsch_bl AS bl
JOIN public.yfcc_latlng_bayern AS s
ON ST_Intersects(bl.wkb_geometry, s.wkb_geometry)
   WHERE bl.nuts_name = 'BAYERN'
GROUP BY bl.nuts_name, bl.wkb_geometry;
```

Code 9 Secondary Aggregation of Data

Code 9 demonstrates how the previously aggregated data can be aggregated further to the area encompassed by all of Bavaria. Figure 19 demonstrates the same concept applied to aggregating all of the federal states of Germany into one unit. A better illustration of this ability can be seen in Figures 20-25. The points are aggregated to administrative regions in Figures 20 & 21. The administrative regions are aggregated to the federal state level in Figures 22 & 23. And finally, the federal states are aggregated to country level in Figures 22 & 23.

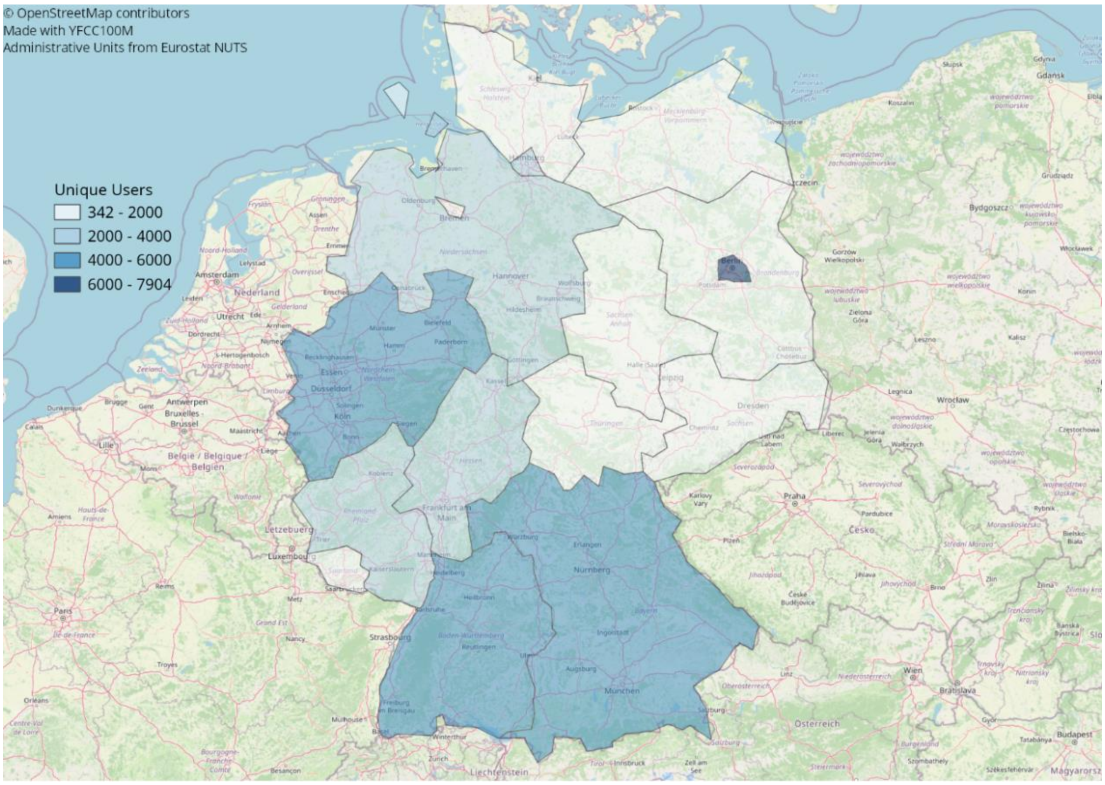


Figure 18 Unique Users in Germany by Federal State

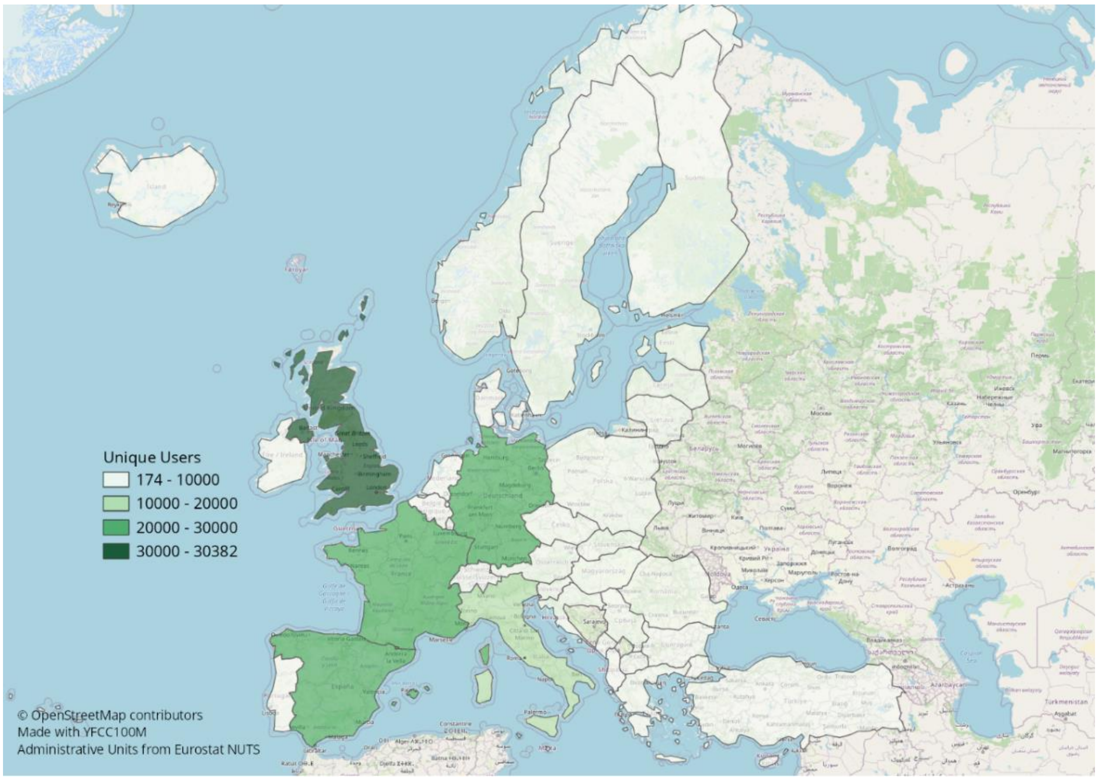


Figure 19 Unique Users in Europe by Country

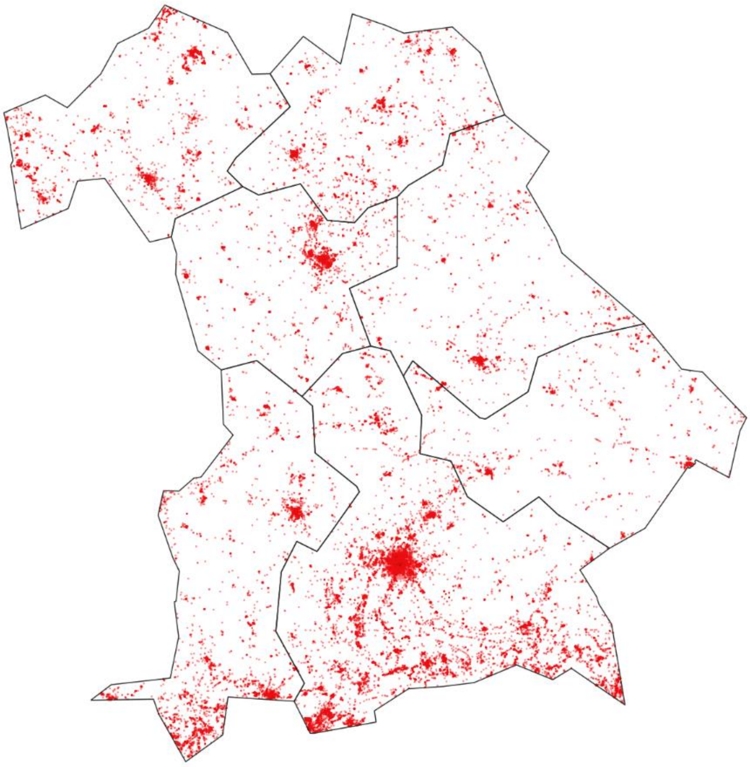


Figure 20 Example of Aggregation in Bavarian Administrative Regions Before

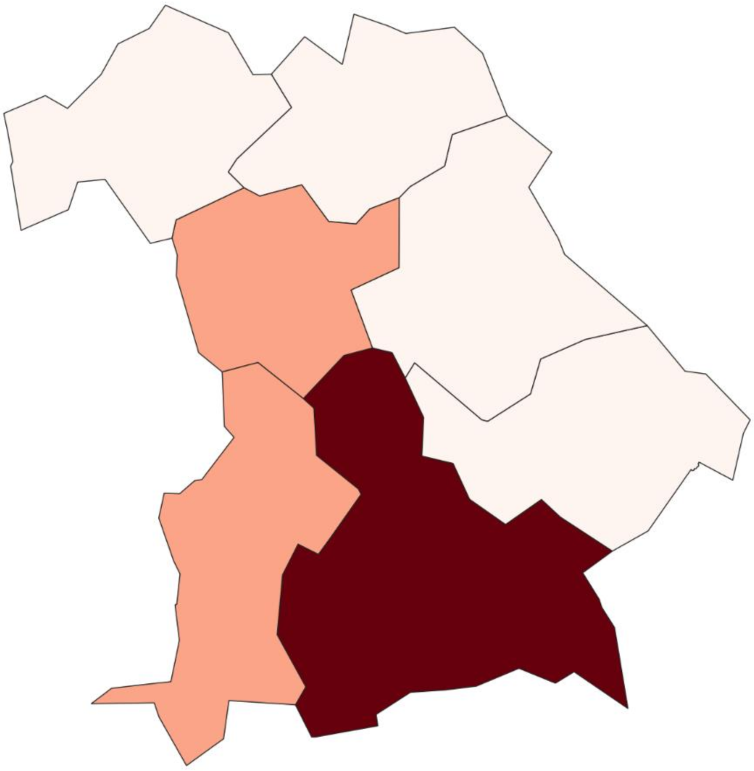


Figure 21 Example of Aggregation in Bavarian Administrative Regions After

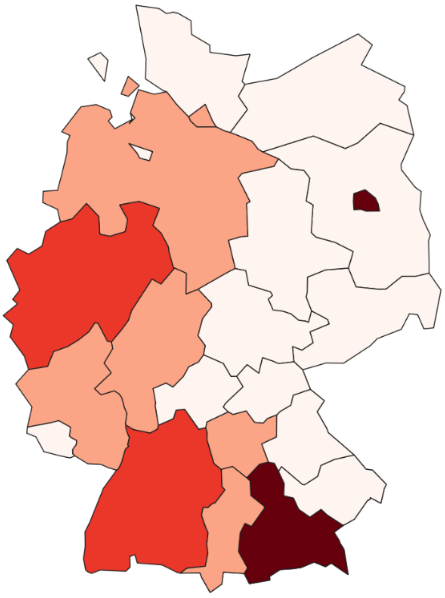


Figure 22 Example of Aggregation in Bavaria Before

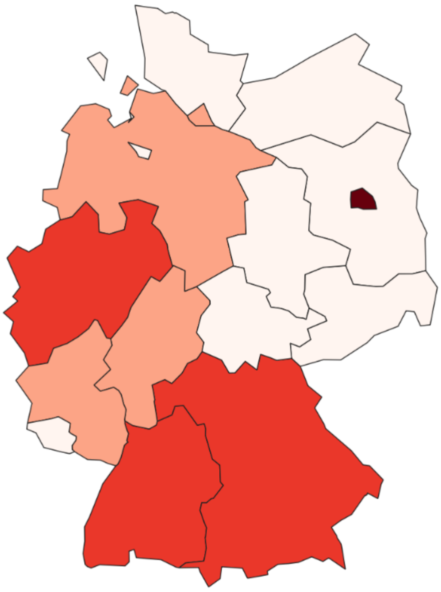


Figure 23 Example of Aggregation in Bavaria After

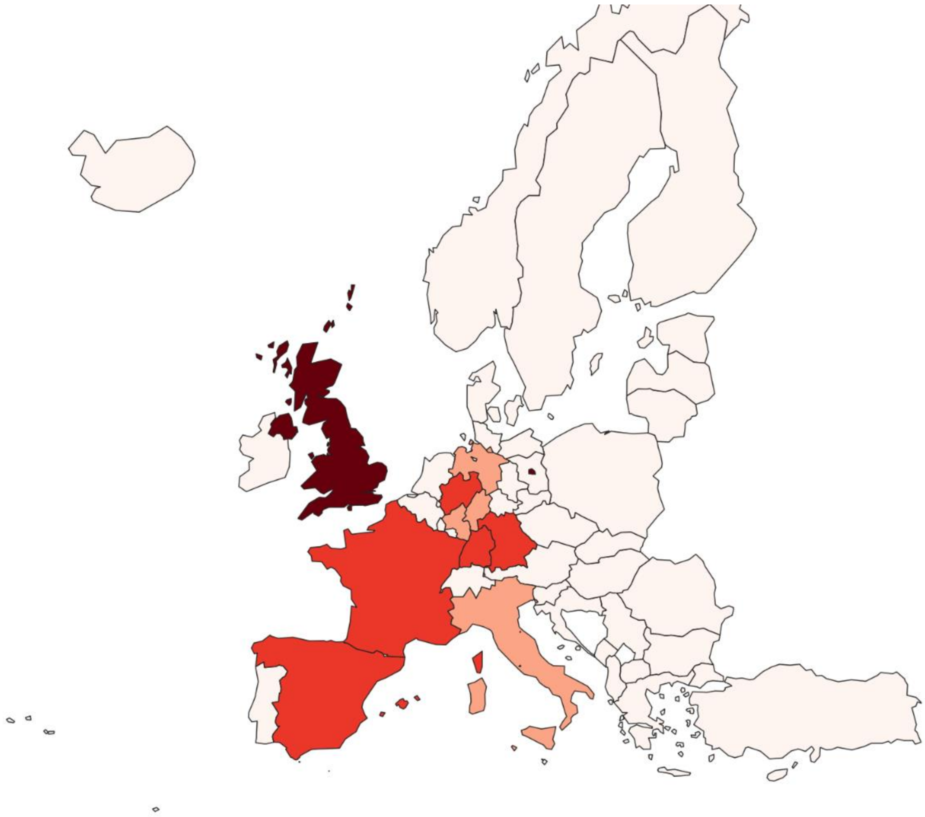


Figure 24 Example of Aggregation in Germany Before

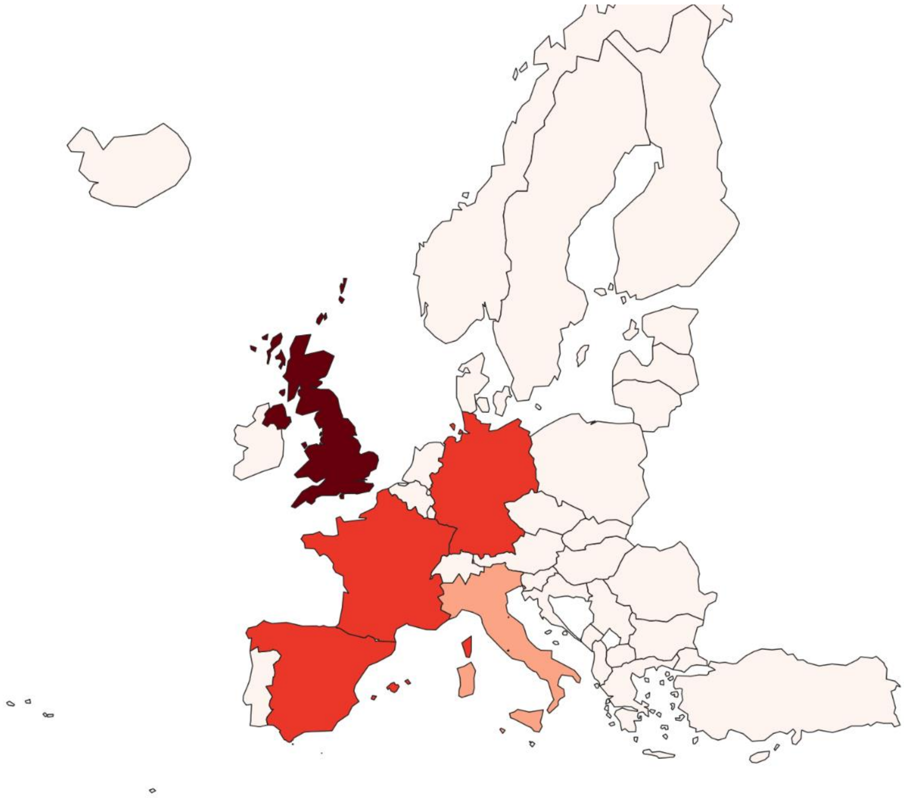


Figure 25 Example of Aggregation in Germany After

4.2.2 Lossless Unions

Lossless unions also have the practical application of allowing an already filtered data set to be updated. For example, the data set containing all posts from March 8th, 2012 (see section 3.2) could be updated with all posts from June 8th, 2012 using a union between the sets. In this example it assumed that the data for March 8th is already filtered and processed into its privacy aware format. Therefore, the first step is to filter the new data from the raw database according the steps outlined in section 3.2.3. Once the data has been processed, a union between the two data sets can be performed. This union requires 3 queries. One query to perform a union on the points that appear in both data sets (Code 10). One query to collect the points that only have posts on March 8th (Code 11). And one query to collect the points that only appear in June 8th (Code 12) These records can simple be concatenated into one relation then mapped (Figure 26). The practical application for these unions is data streaming.

```
SELECT  m.latitude, m.longitude,
hll_cardinality(hll_union(m.user_hll, j.user_hll))::int as usercount,
hll_cardinality(hll_union(m.post_hll, j.post_hll))::int as postcount,
hll_cardinality(hll_union(m.date_hll, j.date_hll))::int as "userdays"
FROM spatial.latlng_marcheighth as m, spatial.latlng_juneeight as j
WHERE m.latlng_geom = j.latlng_geom
```

Code 10 Union Query for YFCC Data from March 8th and June 8th, 2012

```
SELECT  m.latitude, m.longitude,
              hll_cardinality(m.user_hll)::int as usercount,
              hll_cardinality(m.post_hll)::int as postcount,
              hll_cardinality(m.date_hll)::int as userdays
FROM spatial.latlng_marcheighth as m
WHERE m.latlng_geom NOT IN ( SELECT j.latlng_geom
FROM spatial.latlng_juneeight as j)
```

Code 11 Query for YFCC Data from March 8th

```
SELECT  j.latitude, j.longitude,
              hll_cardinality(j.user_hll)::int as usercount,
              hll_cardinality(j.post_hll)::int as postcount,
              hll_cardinality(j.date_hll)::int as userdays
FROM spatial.latlng_juneeight as j
WHERE j.latlng_geom NOT IN ( SELECT m.latlng_geom
FROM spatial.latlng_marcheighth as m)
```

Code 12 Query for YFCC Data from June 8th

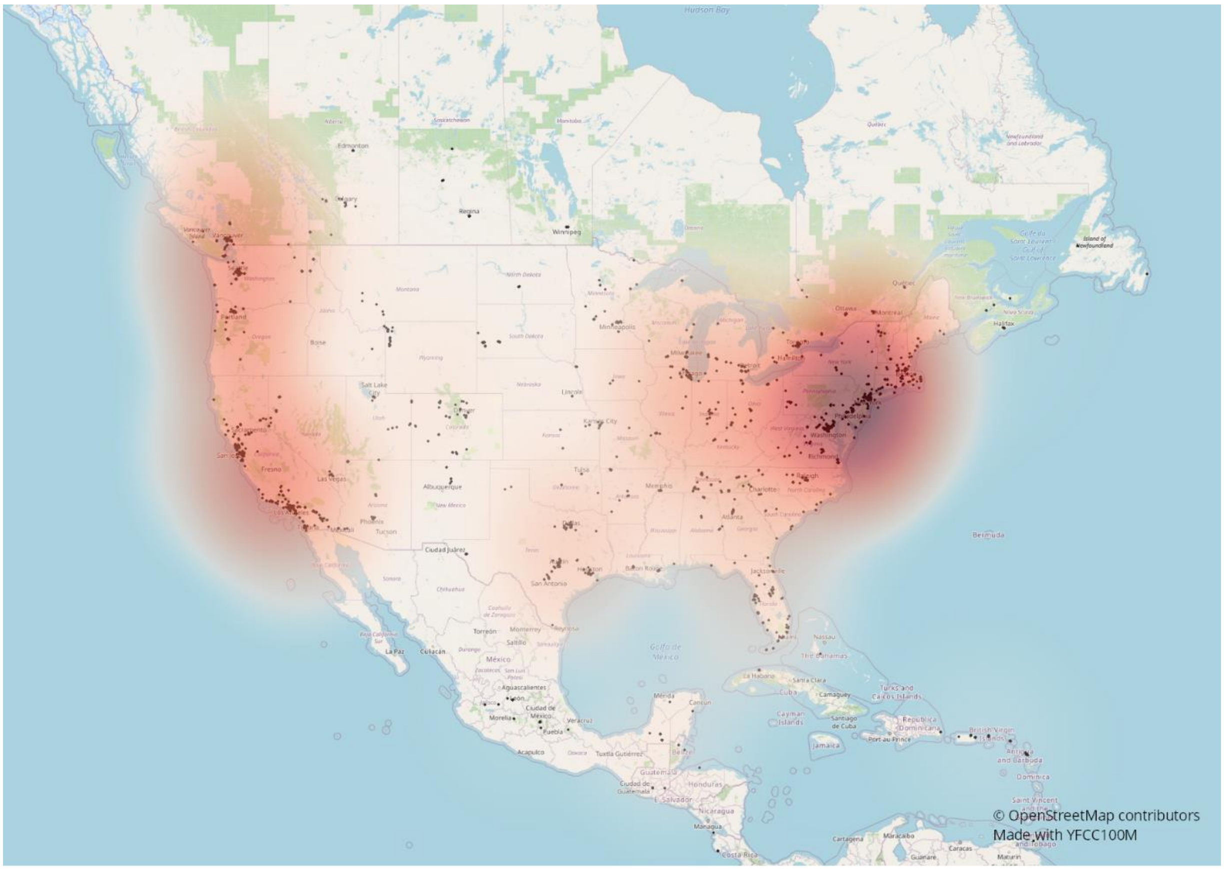


Figure 26 Posts from June 8th Unioned with Posts from March 8th.

4.3 Intersection

Intersections are another possibility with HLL sets. The Inclusion-Exclusion principle allows intersections to be calculated using the cardinalities and unions of two or more sets (Equation 4, Equation 5). In these equations $||$ denotes cardinality, $\cup$ denotes union, and $\cap$ denotes intersection. This equation can also be depicted graphically (Figure 27).

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Equation 4 Inclusion-Exclusion Principle for 2 sets

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

Equation 5 Inclusion-Exclusion Principle for 3 sets

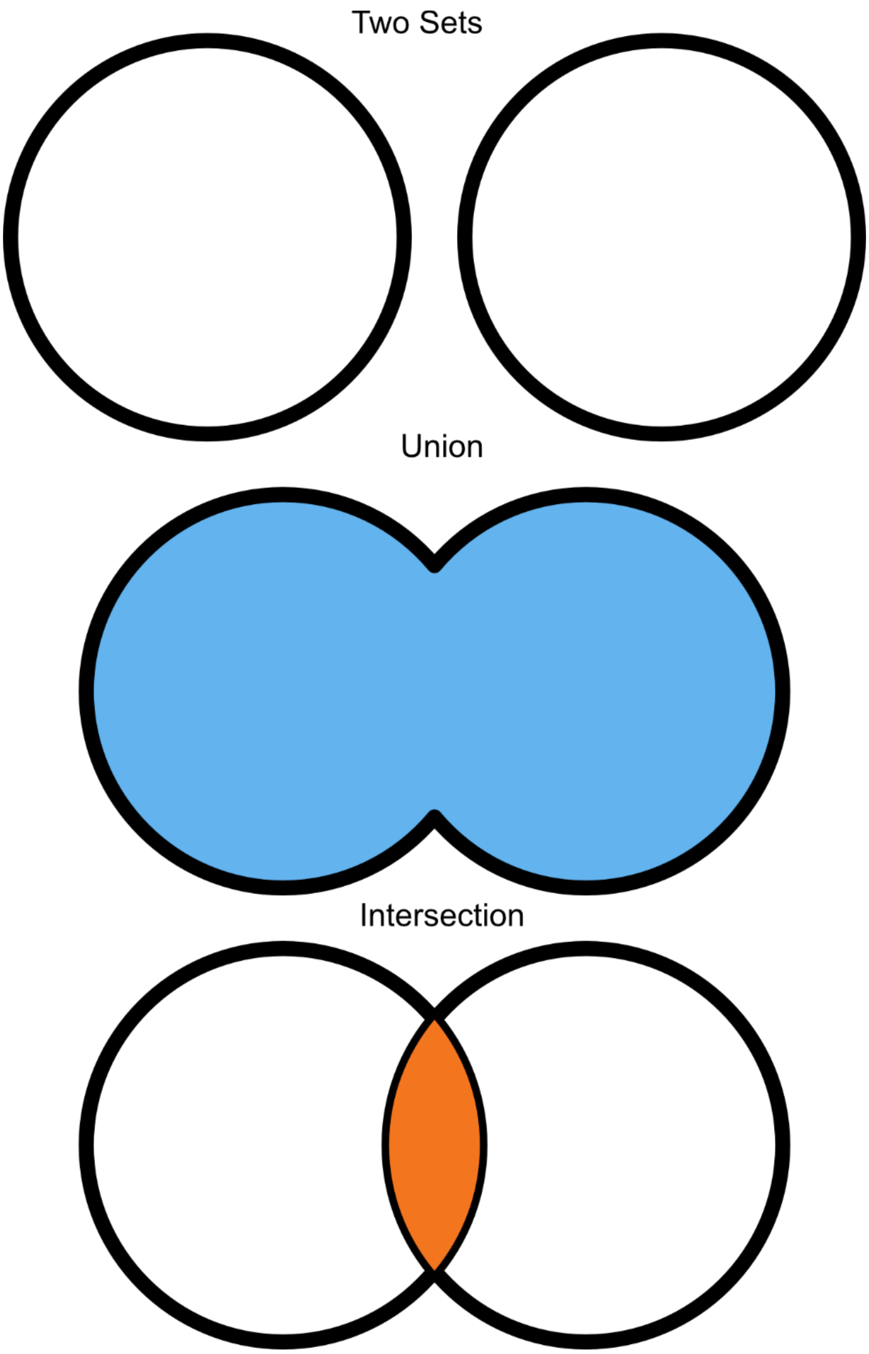


Figure 27 Graphical Depiction of Union and Intersection

Intersections can be applied toward social media analytics in answering questions such as how many unique users posted in Michigan, Colorado, and New Mexico in the USA. Calculating the intersection of three sets follows Equation 5. First the individual cardinalities must be calculated (green fields), the cardinalities of each union (blue fields), and the individual intersections between each state (orange fields)(Table 11). Then the total union of all three sets must be calculated which in this case is 9724 unique users. The calculated result is 243 unique users who had created posts in all three states (Equation 6).

Table 11 Cardinality, Union, and Intersection of Unique Users in Michigan, Colorado, and New Mexico

	Michigan	Colorado	New Mexico
Michigan	4127	890	279
Colorado	8275	5038	785
New Mexico	6118	6523	2270
Color Coding	**Union**	**Intersection**	**Cardinality**

$$|A \cap B \cap C| = 9724 - 4127 - 5038 - 2270 + 890 + 279 + 785 = 243$$

Equation 6 Calculating the Intersection of Unique Users in Michigan, Colorado, and New Mexico

Unlike unions, intersections are not lossless. As mentioned previously, when two sets of unequal size are intersected, the error can render the results greatly skewed or meaningless (Figure 7). The values in Table 11 were compared with the same values calculated from the raw database (Figure 28). The percent difference between the actual number and the HLL number is 72.5% for Michigan and Colorado. Although the percent error is lower for the other two intersections, this error propagates in the final result with an error of 71.2%

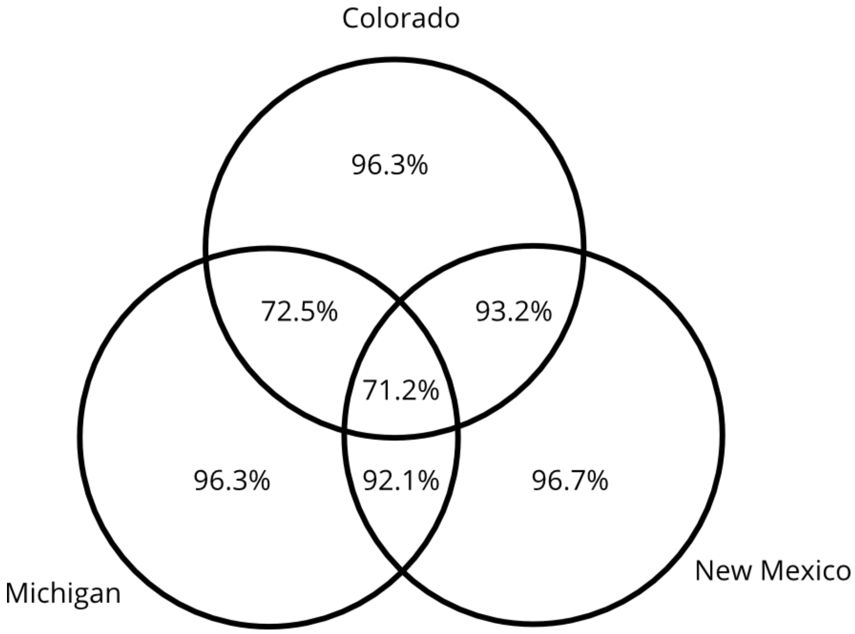


Figure 28 Percent Difference between HLL values and Raw database Values

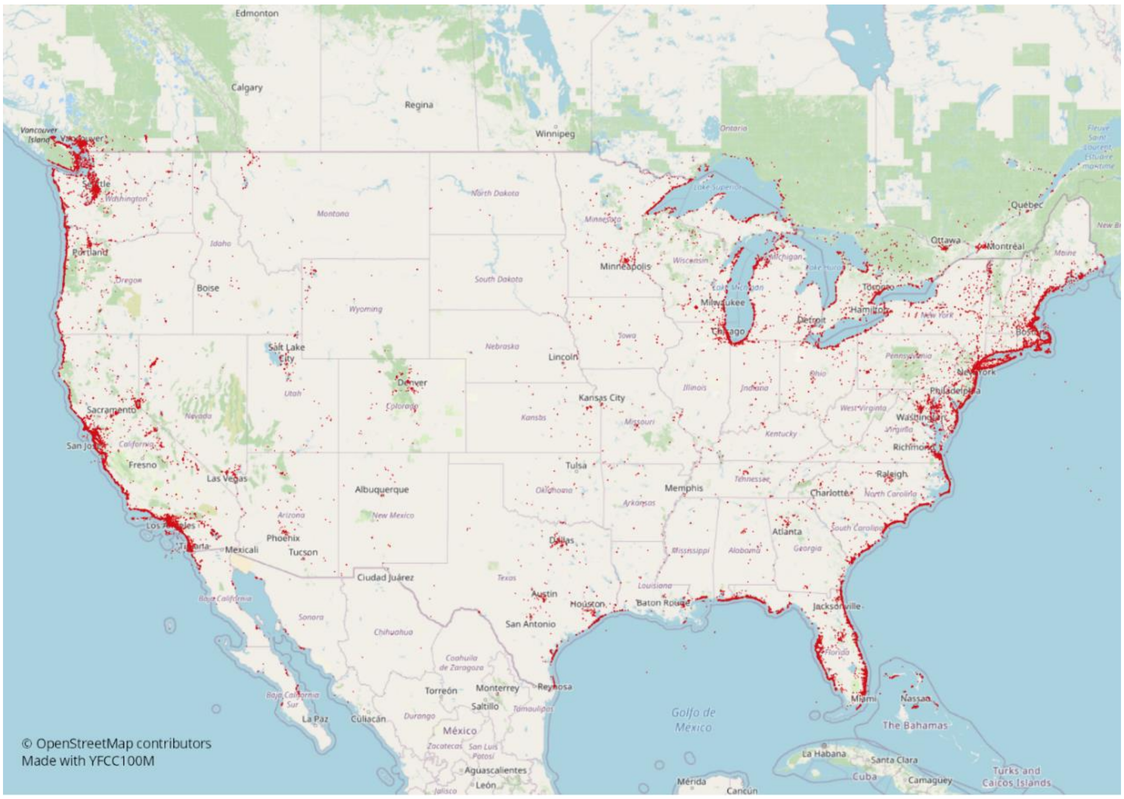
4.4 Filtering by Topic

Filtering the data from the raw database by topic is another possibility. Much of social media analytics requires not just data from a specific location, but rather data that relates to a certain theme or topic. This data is filtered in the data transformation step of the processing raw data into the HLL structure. One common filter is for reactions to a particular topic. Code 13 demonstrates filtering of the raw data for reactions that relate to the beach in a variety of languages. The filtered data is then processed into the HLL table and mapped in North America in Figure 29.

```
CREATE MATERIALIZED VIEW mviews.spatiallatlng_raw_beach AS
          SELECT  extensions.crypt_hash(t1.post_guid, 'samplekey') as "post_guid",
            ST_Y(t1.post_latlng) As "latitude",
                             ST_X(t1.post_latlng) As "longitude",
                             t1.post_latlng As "geom",
                             extensions.crypt_hash(t1.user_guid, 'samplekey') as "user_guid",
                             to_char(t1.post_create_date, 'yyyy-MM-dd') as "post_create_date",
                             t1.post_geoaccuracy
          FROM  topical.post t1
          WHERE  t1.post_latlng
                             && ST_MakeEnvelope (
                             -125.0011, 24.9493,
                              -66.9326, 49.5904,
                              4326) AND
                              t1.post_geoaccuracy IN ('place', 'latlng', 'city')
                              AND (
                             -- tags concat: check if any is contained by
                             -- citext format: case-insensitive text conversion
                             ARRAY['beach', 'strand', 'la_playa',' plage', 'playa',
                             'plaża','spiaggia','lakeshore','seashore','shore-
line','shore']::citext[] && t1.hashtags::citext[]
                             OR
                             -- tags separated: check if is both contained by
                             ARRAY['sea','ocean']::citext[] <@ t1.hashtags::citext[]
                             OR
                             -- words in title: check if any exists as a part of the text
                             -- note that underscores ( _ ) are replaced by space
                             -- character here
                             lower(t1.post_title) ilike any (
                                       array['% beach %','% strand %','% playa %',
                                       '% plage % la_playa %','% plaża %','% spiaggia %',
                                       '% lakeshore %','% seashore %','% shoreline %','% shore
%'])
                             OR
                             -- words in post body (photo description):
                             -- check if any exists as a part of the text
                             -- note that underscores ( _ ) are replaced by space
                             -- character here
                             lower(t1.post_body) ilike any (
                                       array['% beach %','% strand %','% playa %',
                                       '% plage % la_playa %','% plaża %','% spiaggia %',
                                       '% lakeshore %','% seashore %','% shoreline %','% shore
%']))
```

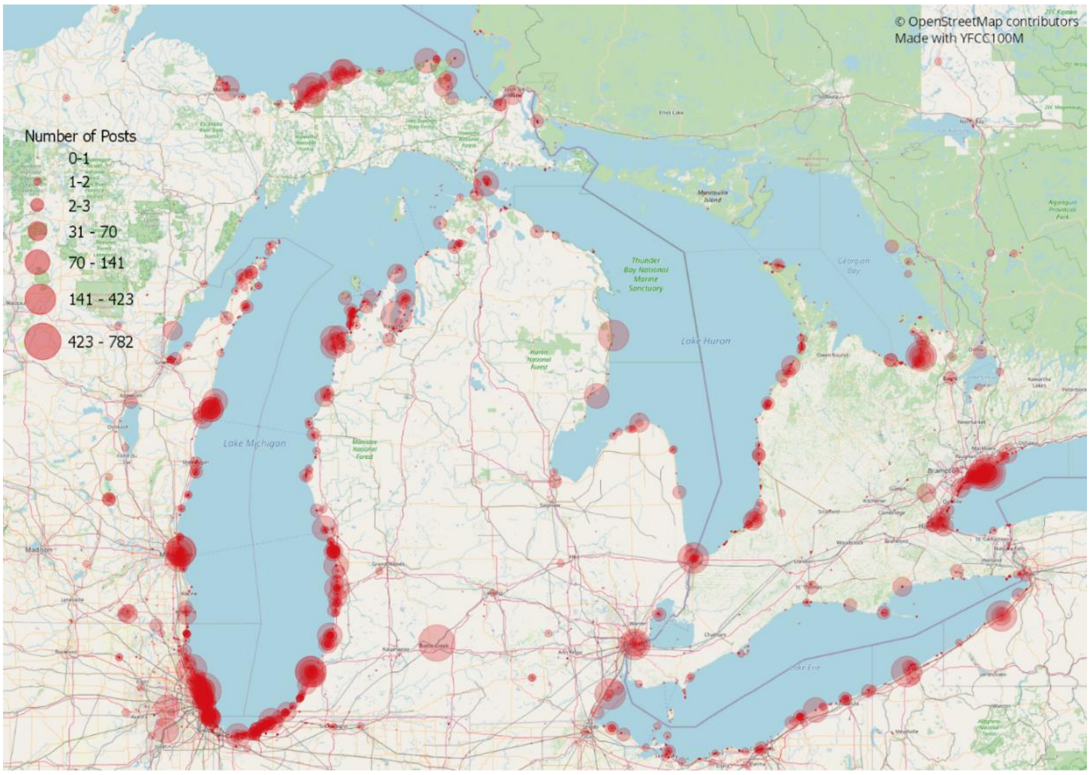
Code 13 SQL Query Filtering Reactions like Beach

Note adapted from (*A Basic Visual Analytics Example - LBSN Structure*, n.d.)



© OpenStreetMap contributors
Made with YFCC100M

Figure 29 Reactions to terms like “Beach” in the US



© OpenStreetMap contributors
Made with YFCC100M

Figure 30 Number of Posts Reacting to terms like “Beach” around Michigan

Figure 30 demonstrates the same data at a larger scale and proportionally scaled to the number of posts at each location. These proportional and overlapping symbols allow one to see hotspots around Lake Michigan, Lake Superior, and Lake Ontario.

Figure 31 uses data processed in the exact same way as the previous example, but for reactions related to the inauguration of the president of the US. With this data it is also possible to create a heat map weighted to the number of posts at a given coordinate pair. The resulting visualization shows the pattern of Flickr users' posts in Washington D.C.

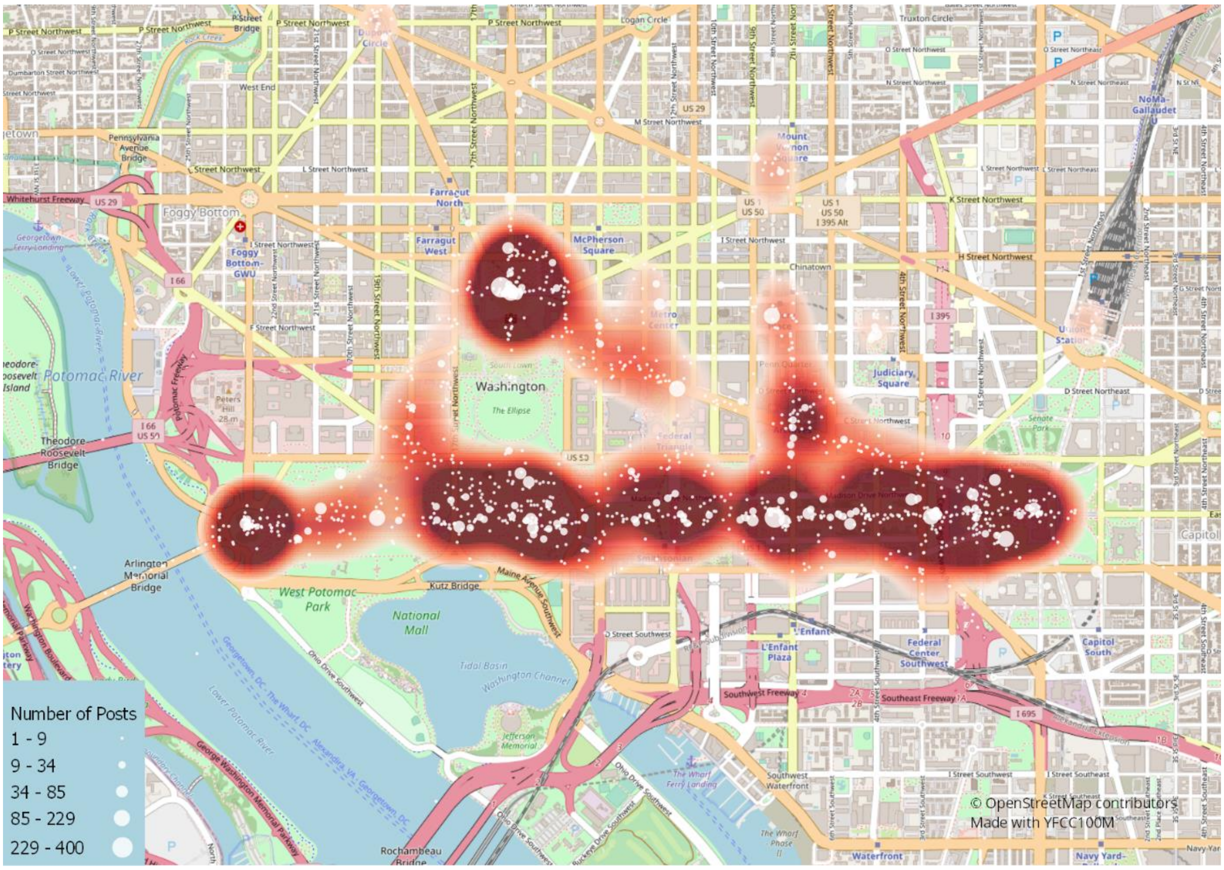
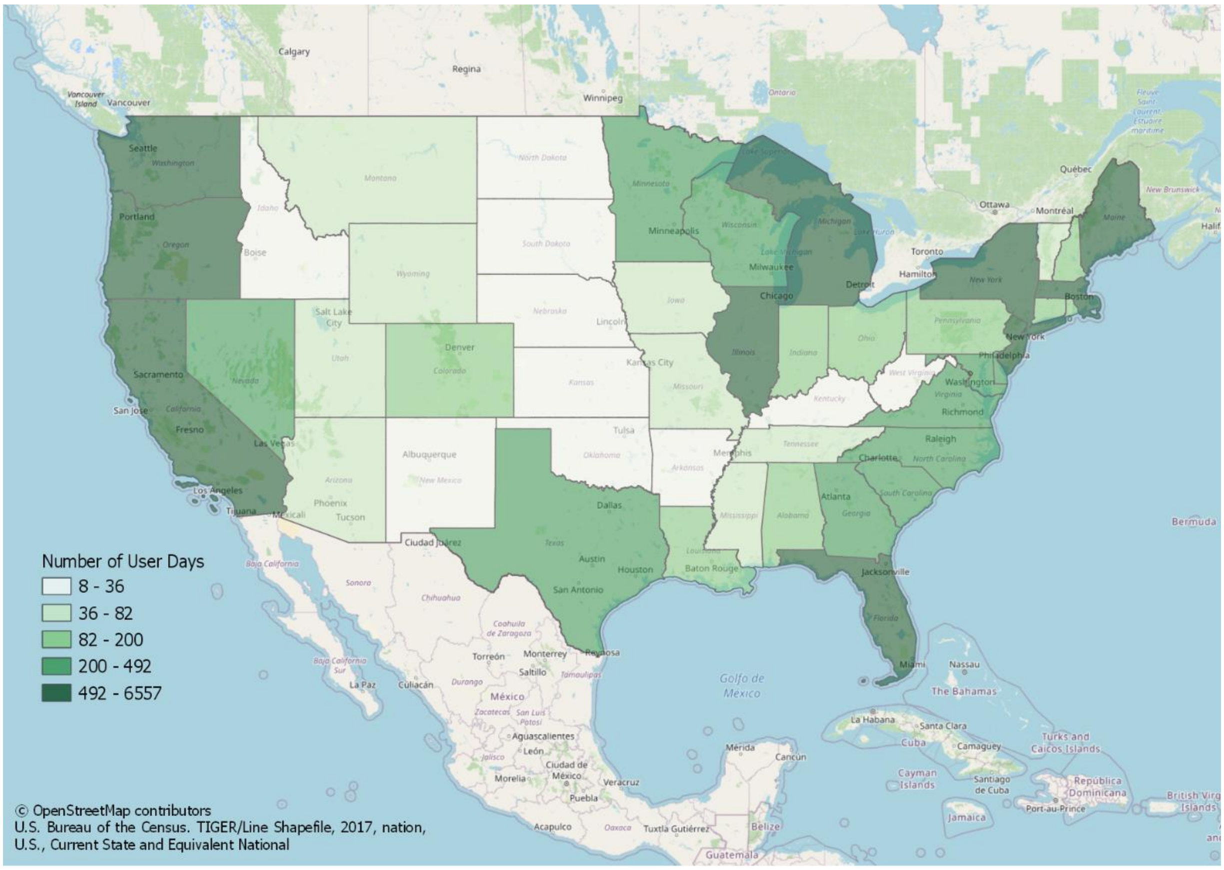


Figure 31 Number of Posts Reacting to Terms like "Inauguration" in Washington D.C.

4.5 Increased Complexity

The methods and techniques demonstrated in the previous sections of the case study are not mutually exclusive i.e. they can be combined. For example, once the data for beach reaction has been processed into the HLL format, it is possible to aggregate them to geographic areas. Figure 32 demonstrates the data related to beach reactions aggregated to the contiguous 48 states.



© OpenStreetMap contributors
U.S. Bureau of the Census. TIGER/Line Shapefile, 2017, nation, U.S., Current State and Equivalent National

Figure 32 Number of User Days Reacting to Terms like “Beach” Aggregated by State

Furthermore, this data can be used to calculate intersections. Using the inclusion exclusion principle (Equation 4) the intersection between states is easily calculated, e.g. Florida and California or Massachusetts and Oregon (Table 12,Table 13). Although a value is calculated for the intersection of post count and user days, these values are essentially meaningless and fall within the margin of error one can expect when intersecting data sets. The intersection of user days can only be meaningful when constrained to a smaller geographic area like a city or tourist region.

Table 12 Union and Intersection of HLL Values in Florida and California

	Florida	California	Union	Intersection
Post Count	61726	143667	200695	4698
User Days	9351	23995	33519	173
User Count	2730	6557	8750	537

Table 13 Union and Intersection of HLL Values in Massachusetts and Oregon

	Massachusetts	Oregon	Union	Intersection
Post Count	13917	13296	26826	387
User Days	1735	2121	3913	57
User Count	710	849	1526	33

To provide an example with a meaningful intersection that is not just unique users, an HLL field for this data set was added that counts not unique user days, but rather unique days themselves. The specific dates were then aggregated based on latitude. One group included post at a latitude higher than or equal to 37.5 N (Code 14) and the other group included posts below the latitude 37.5 N (Code 15). The division of posts is mapped in Figure 33.

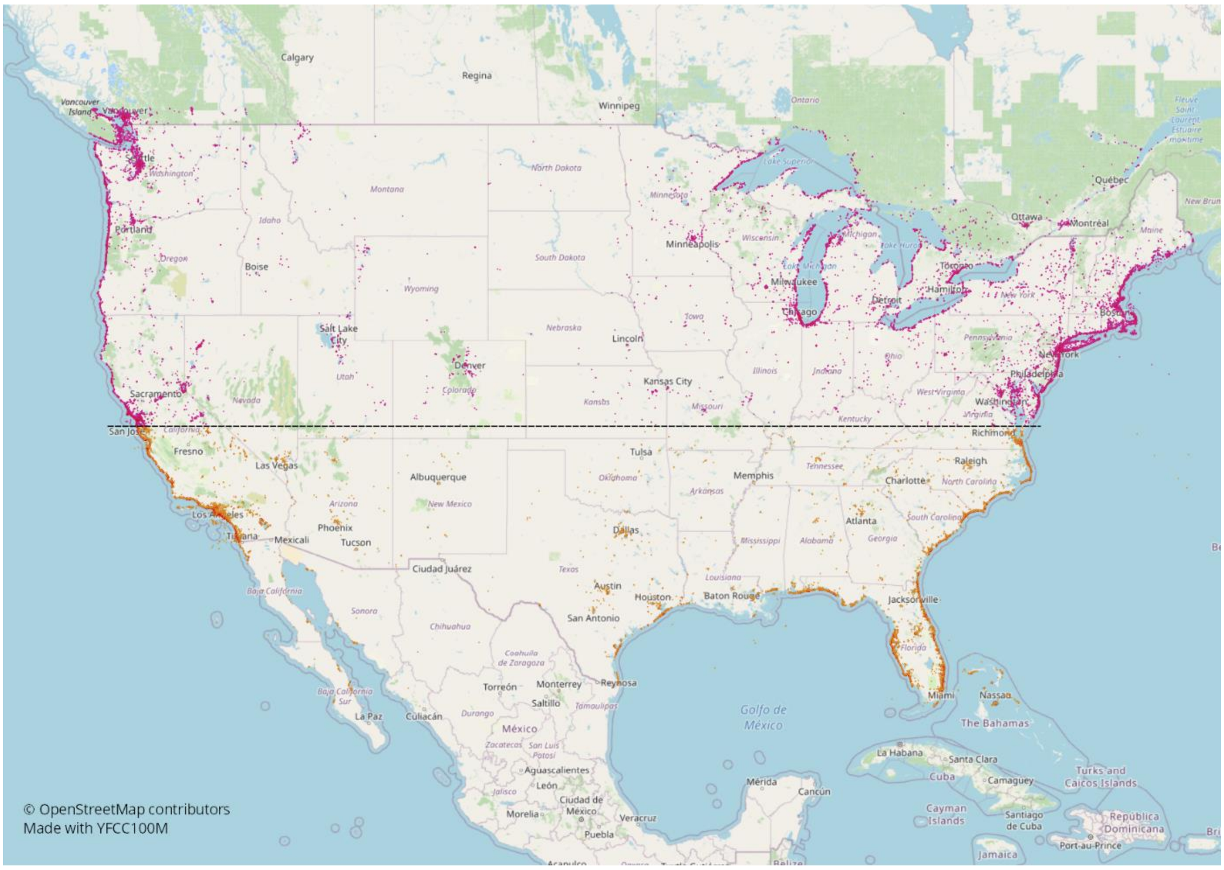


Figure 33 Posts Reacting to “Beach” divided at 37.5 N

```
SELECT
        hll_union_agg(specific_date_hll)::hll as date,
        hll_union_agg(user_hll)::hll as usercount
FROM spatial.latlng_beach
WHERE latitude <37.5
```

Code 14 Northern Beach Posts

```
SELECT
        hll_union_agg(specific_date_hll)::hll as date,
        hll_union_agg(user_hll)::hll as usercount
FROM spatial.latlng_beach
WHERE latitude >= 37.5
```

Code 15 Southern Beach Posts

The intersection of unique users and unique days is then shown in Table 14. Since the unique days are calculated based on date of post creation and not date uploaded to Flickr, there are more unique days than would be in 10 years.

Table 14 Union and Intersection of HLL Values in Above and Below 37.5 N

	Above 37.5 N	Below 37.5 N	Union	Intersection
Unique Days	4120	4017	4560	3577
User Count	10189	8917	16636	2470

5 Discussion

The above case study is incomplete without addressing how the visualizations meet the criteria for high quality maps as well as their ability preserve privacy. The benefits and disadvantages to this application the HLL data structure must be addressed. It is also important to understand HLL's limitations for spatial social media data as well as how easily it can be deployed.

5.1 Use-Case

HyperLogLog is not privacy preserving unto itself; it must be combined with other methods to preserve privacy. In one study, it is demonstrated that cardinality estimators cannot preserve privacy (Desfontaines et al., 2019). Desfontaines et al. give two requirements that must be met in order to mitigate privacy risks in HLL. Since HLL uses a non-cryptographic MurMurHash function (Desfontaines et al., 2019) it is necessary to use a cryptographic hash function to secure the unique IDs. This prerequisite is met in this case study by using the pgcrypto extension to secure the unique user IDs. The second prerequisite is for the raw data to be hidden behind a restricted API. This prerequisite is not met per se in the case study for multiple reasons. Firstly, the YFCC100M data set is already available in its raw format for anyone to access. Secondly, no API has been developed that would allow for this type of case study to occur. Therefore, the HLL sets as accessed during the case study would need to be restricted in a true privacy aware scenario.

One risk described is known as an intersection attack and can be employed in HLL sets where the user cardinality is one. As a simplified example, if one were to union an HLL set with a user cardinality of one with another set, it would be known whether or not that unique user is in both sets or only one based on whether or not the cardinality of the union changed by one (Desfontaines et al., 2019). Desfontaines et al. describe intersection attacks as realistic but in practice more complicated than simply unioning HLL sets.

5.2 Evaluation of Maps

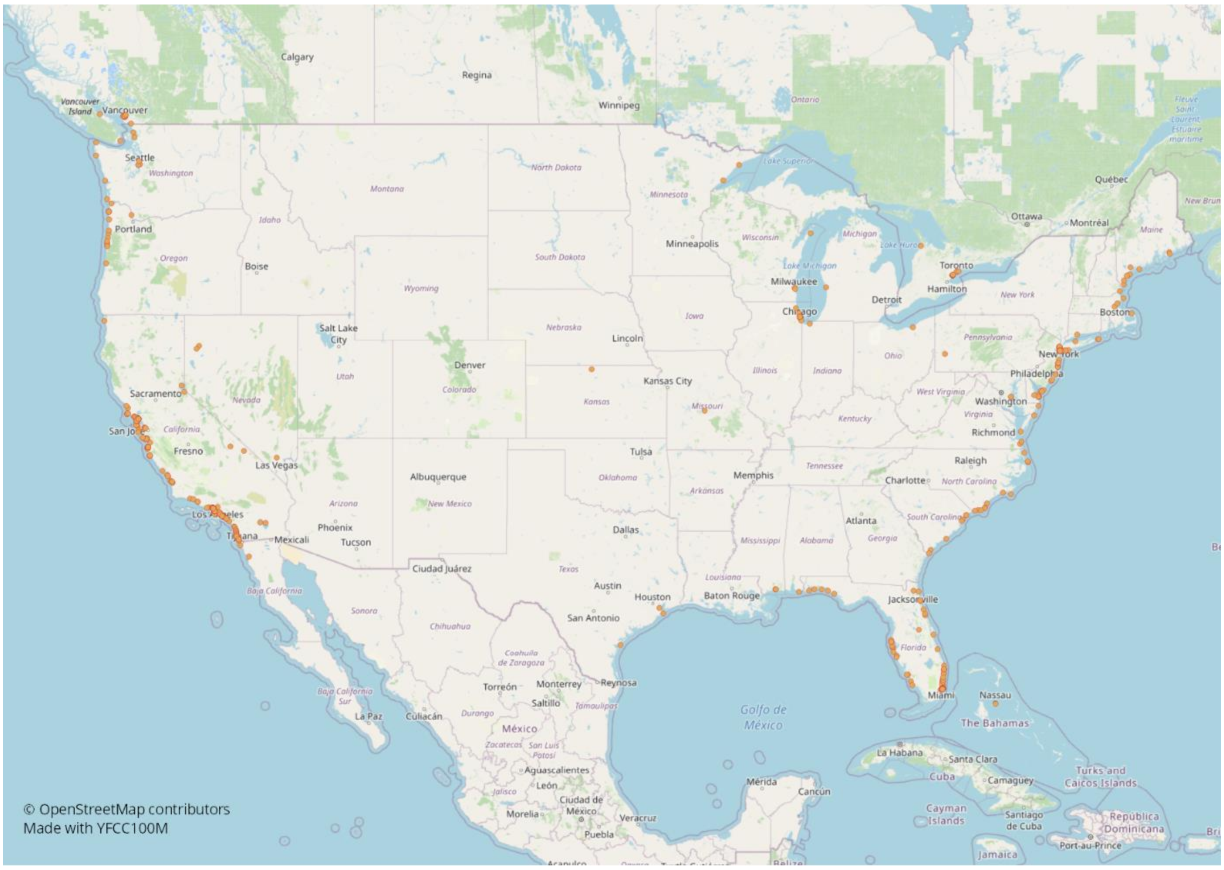
In order to evaluate the results of the case study, it is important to revisit the definition of geoprivacy given previously i.e. the ability of a single user to remain anonymous in the publicly available data set. Privacy is not a static concept and there is a tradeoff between privacy and the level of detail of the information being mapped.

5.2.1 Maps as Public Data

If it is assumed that the data is not publicly available until the creation of the maps, then the maps themselves serve as the publicly available information. In this context, it would be very difficult to identify a single user from any of the maps provided. Figure 31, which depicts the number of posts reacting to the inauguration of the president, is both the largest scale example from the case study as well as the most sensitive due to its political nature. In this visualization it would be impossible to identify a single user without the aid of additional information.

5.2.2 Database as Public Data

In the case that the publicly available information is the HLL database but not the raw database, the privacy of the users could still be preserved. This is because during the processing of the data from its raw form into the privacy aware structure, a cryptographic hash function is used on the unique user (Code 2). But there does remain the risk of intersection attacks. A potential solution would be to not allow HLL sets with a fine location granularity and a unique user cardinality of one to be shared in their HLL format. But if these values were not returned in their HLL format, no further set operations could be performed i.e. union, union aggregation. In the HLL data set that is not filtered by topic, 98.8% of all coordinate pairs have a user count of one. This would eliminate a vast number of records that one could use for further set operations. This constraint becomes even more restrictive if applied to the data set that has been filtered for reactions relating to the beach. In this data set 99.6% of all records have a user cardinality of one. The coordinate pairs with a user cardinality greater than one are mapped in Figure 34. One can see large reduction in data in contrast to Figure 29 in which the same data is mapped with no restriction. One potential solution to this restriction would be to classify the user cardinalities as ranges of value as this is often how they are mapped anyway e.g. proportional symbol maps. This issue demonstrates well the tradeoff between privacy and detail. The threat of an intersection attack on HLL sets is a potential area of further research.



© OpenStreetMap contributors
Made with YFCC100M

Figure 34 Reactions to Beach with Unique Users More than 1 Mapped

5.2.3 Quality of Maps

Finally, the overall quality of the maps is only hindered by data accuracy and type of visualization. The data can always be processed in such a way that provides coordinate pairs with numerical metrics overlays. For example, one could create a new metric overlay that is a count of how many unique camera types were used at a specific location. Furthermore, the points themselves can be aggregated to larger and larger granularities. Once the points are provided, they can be mapped and visualized in the same way as any other point data set can be mapped. This means that as long as the concept being mapped can be mapped as a cardinality and as long as the error of the cardinality estimation is not too great for the application, the quality is comparable to maps made with raw data.

5.3 Advantages of HLL

The HLL data structure has advantages that are not related to its ability to preserve privacy. For example, the HLL data structure greatly reduces the amount of the data stored. This decrease in memory storage is an asset when it comes to working with big data in social media. This also allows for an increase in processing speeds when performing computationally intensive calculations. Another advantage is the ability to perform lossless unions as demonstrated in Section 4.2.2. These lossless unions allow the data to be updated seamlessly so in the case that the data being analyzed should be in real time, there is no issue in updating the already processed data sets and maintaining the 2% margin of error inherent to the algorithm. Lastly, the HLL data structure can be combined with other techniques for social media analysis as well as preserving privacy. This thesis demonstrated that flexibility by combining HLL with a cryptographic hash and aggregations.

5.4 Disadvantages of HLL

The limitations the HLL data structure are mostly limitations to what topics can actually be analyzed using social media date. Some types of visualizations implicitly, if not explicitly, require each record have unique and identifiable values e.g. dates, user Ids, etc. For example, in one study a tourist index was created by mapping out YFFC100M posts where the users could be assumed to be a certain distance from their homes and thus tourists (Mao, 2015). It is implicit that one must be able to track the individuals as they leave a specified region in order to create this map. Therefore, it would be difficult to end up with the same result using the HLL data structure. The 2% error inherent to the estimation of cardinality can be a disadvantage depending on the required accuracy for the context, but this margin will usually fall within the acceptable tolerance when mapping social media data.

Another noteworthy disadvantage is that the social media researcher must already understand the data set they are creating the privacy aware data from. Moreover, the researcher must have an understanding of what they intend to map. With raw data, it is possible to experiment with the data sets and its various fields, but with HLL once the data has been processed, it cannot be repurposed easily. This limitation also applies to the aggregation of data. If the data has already been aggregated, there is no mechanism to disaggregate. This means that the map creator will need to know which granularity is necessary to create the intended map. The ability to aggregate into coarser granularities has been demonstrated, but there is a tradeoff in privacy when collecting data at a finer granularity than is necessary.

5.5 Ease of Deployment

A final consideration is the ease with which one can deploy HLL for spatial social media analytics. Conceptually, HLL is not so intuitive. Specifically, the ability for large errors to propagate when performing intersections could hinder the ability for a researcher to deploy this technique. Another issue with HLL is that one must use SQL to perform unions, aggregations, and calculate cardinalities. Simply receiving the data in a table is not enough. In order to use the HLL functions, one must have continual access to the database.

On the other hand, there is already a proposed LBSN database structure that allows for seamless integration of HLL (*A Basic Visual Analytics Example - LBSN Structure*, n.d.). Moreover, cardinality itself is not a complicated subject matter. Therefore, the set operations union, aggregation, and cardinality using the Citus extension are relatively straight forward and do not require a large body of prior knowledge to operate.

In order to make these tools easier to put into the hands of social media researchers, a programming language library could be developed to facilitate the use of HLL while overcoming complications and pitfalls by abstracting them in the internal classes of the library. This hypothetical library could interface with the restricted API from which data would be queried as well as incorporate the standard LBSN database structure for a more comprehensive and flexible mapping tool.

6 Conclusions

In a world where data sets are growing at an increasing rate and where mobile position devices are ubiquitous, a balance must be struck between individual users' rights to privacy and society's ability to benefit from big data. The ability to use these large social media data sets to understand the world has proven a boon to many fields of research, but the cost of potentially compromising the privacy of so many individuals should be carefully weighed. This is especially pertinent when discussing the geoprivacy and location data that is made available from the use of these platforms.

This thesis attempts to describe one technique to strike a balance between social media analytics and personal geoprivacy using the HyperLogLog algorithm and a privacy aware database for cartographic visualizations. In the process, it addressed a few overarching themes: 1) How well can this database technique preserve individual privacy? 2) Are the subsequent visualizations as good as or potentially better than their non-privacy aware counterparts? 3) What are the limitations in accuracy of the results as well as which types of visualization can be produced?

This thesis attempts to address these themes through the use of a case study demonstrating exactly how a raw data set could be transformed into a parallel, privacy aware version and subsequently used to create a variety of maps. Each of these maps demonstrates a facet of the HLL data structure and how it can be used for social media analysis. The final map demonstrates the combination of these facets. it was determined that the privacy aware data structure is in fact privacy preserving and can in fact create visualizations on par with its non-privacy aware counterpart. Furthermore, it has some distinct advantages in its application towards understand big social media data like reduced data storage, faster processing, and flexibility. On the other hand, this privacy aware data structure is not as intuitive to use and requires a background knowledge for deployment. Moreover, it is not privacy aware unto itself but rather must be combined with other privacy preserving techniques, and there still remains some privacy risks that must be addressed.

The use of HLL for privacy aware analysis of spatial social media shows great potential. Its ability to be combined with other privacy aware techniques is one of its greatest advantages. Moreover, its potential to be incorporated into a privacy aware mapping library for a major programming language makes the continued research and development of HLL as a tool in the field of geoprivacy a worthwhile endeavor.

7 References

A basic visual analytics example—LBSN Structure. (n.d.). Retrieved September 3, 2020, from https://lbsn.vgiscience.org/yfcc-geohash/

Beresford, A. R., & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive Computing*, *2*(1), 46–55. https://doi.org/10.1109/MPRV.2003.1186725

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

California Consumer Privacy Act (CCPA). (2018, October 15). State of California - Department of Justice - Office of the Attorney General. https://oag.ca.gov/privacy/ccpa

Cavoukian, A. (n.d.). *The 7 Foundational Principles*. 2.

Chen, S., Lin, L., & Yuan, X. (2017). Social Media Visual Analytics. *Computer Graphics Forum*, *36*(3), 563–587. https://doi.org/10.1111/cgf.13211

Citusdata/postgresql-hll. (2020). [C]. Citus Data. https://github.com/citusdata/postgresql-hll (Original work published 2013)

Daniele. (2005). *tower in the sky (n. #10.000.000 Flickr photo)* [Photo]. https://www.flickr.com/photos/jereye/10000000/

Definition of PRIVACY. (n.d.). Retrieved July 20, 2020, from https://www.merriam-webster.com/dictionary/privacy

Definition of SOCIAL MEDIA. (n.d.). Retrieved August 5, 2020, from https://www.merriam-webster.com/dictionary/social+media

Deng, N., & Li, X. (2018). Feeling a Destination through the “Right” Photos: A Machine Learning Model for DMOs’ Photo Selection. *Tourism Management*, *2018*. https://doi.org/10.1016/j.tourman.2017.09.010

Desfontaines, D., Lochbihler, A., & Basin, D. (2019). Cardinality Estimators do not Preserve Privacy. *Proceedings on Privacy Enhancing Technologies*, *2019*, 26–46. https://doi.org/10.2478/popets-2019-0018

Desfontaines, D., & Pejó, B. (2019). SoK: Differential Privacies. *ArXiv:1906.01337 [Cs]*. http://arxiv.org/abs/1906.01337

Duckham, M., & Kulik, L. (2005). *A Formal Model of Obfuscation and Negotiation for Location Privacy*. *3468*, 152–170. https://doi.org/10.1007/11428572_10

Dunkel, A., Andrienko, G., Andrienko, N., Burghardt, D., Hauthal, E., & Purves, R. (2019). A conceptual framework for studying collective reactions to events in location-based social media. *International Journal of Geographical Information Science*, *33*(4), 780–804. https://doi.org/10.1080/13658816.2018.1546390

Dwork, C. (2008). Differential Privacy: A Survey of Results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

Flajolet, P., Fusy, É., & Gandouet, O. (n.d.). *HyperLogLog: The analysis of a near-optimal cardinality estimation algorithm*. 20.

General Data Protection Regulation (GDPR) – Official Legal Text. (n.d.). General Data Protection Regulation (GDPR). Retrieved August 19, 2020, from https://gdpr-info.eu/

Gómez-Barrón, J.-P., Manso-Callejo, M.-Á., Alcarria, R., & Iturrioz, T. (2016). Volunteered Geographic Information System Design: Project and Participation Guidelines. *ISPRS International Journal of Geo-Information*, *5*(7), 108. https://doi.org/10.3390/ijgi5070108

Google Wants to Help Tech Companies Know Less About You. (n.d.). *Wired*. Retrieved July 28, 2020, from https://www.wired.com/story/google-differential-privacy-open-source/

Harari, Y. N. (2017). *Homo Deus: A Brief History of Tomorrow* (Reprint edition). Harper.

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

Hauser, C., & Kabatnik, M. (2001). Towards Privacy Support in a Global Location Service. *In Proc. of the IFIP Workshop on IP and ATM Traffic Management (WATM/EUNICE 2001*, 81–89.

Hildebrandt, M. (2006). *Privacy and Identity*. 15.

Keßler, C., & McKenzie, G. (2018). A geoprivacy manifesto. *Transactions in GIS*, *22*(1), 3–19. https://doi.org/10.1111/tgis.12305

Kounadi, O., Resch, B., & Petutschnig, A. (2018). Privacy Threats and Protection Recommendations for the Use of Geosocial Network Data in Research. *Social Sciences*, *7*(10), 1–17. https://doi.org/10.3390/socsci7100191

Löchner, M., Dunkel, A., & Burghardt, D. (2019). *Protecting privacy using HyperLogLog to process data from Location Based Social Networks*.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (n.d.). *Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scal…* 20.

Mao, T. (2015). Mining One Hundred Million Creative Commons Flickr Images Dataset to Flickr Tourist Index. *International Journal of Future Computer and Communication*, *4*(2), 104–107. https://doi.org/10.7763/IJFCC.2015.V4.365

Metrics (Overlays)—LBSN Structure. (n.d.). Retrieved September 1, 2020, from https://lbsn.vgiscience.org/concept-metrics/

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 111–125. https://doi.org/10.1109/SP.2008.33

Newzoo's Global Mobile Market Report: Insights into the World's 3.2 Billion Smartphone Users, the Devices They Use & the Mobile Games They Play. (n.d.). *Newzoo*. Retrieved July 10, 2020, from https://newzoo.com/insights/articles/newzoos-global-mobile-market-report-insights-into-the-worlds-3-2-billion-smartphone-users-the-devices-they-use-the-mobile-games-they-play/

Overwhelming Number Of Smartphone Users Keep Location Services Open |. (2016, April 22). https://geomarketing.com/overwhelming-number-of-smartphone-users-keep-location-services-open

Predictive policing algorithms are racist. They need to be dismantled. (n.d.). MIT Technology Review. Retrieved July 28, 2020, from https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

Regian, W., & Noever, D. (2017, November 27). *Generative Representation of Synthetic Threat Actors for Simulation and Training Generative Representation of Synthetic Threat Actors for Simulation and Training*.

Samarati, P., & Sweeney, L. (n.d.). *Protecting Privacy when Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression*. 19.

Smith, Marc. (2010). *2010—June—3—NodeXL - Twitter—PDF2010* [Photo]. https://www.flickr.com/photos/marc_smith/4667113851/

Smith, Matthew, Szongott, C., Henne, B., & von Voigt, G. (2012). Big data privacy issues in public social media. *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 1–6. https://doi.org/10.1109/DEST.2012.6227909

Solove, D. J. (2008). *Understanding privacy*. Harvard University Press.

Structure (Bases)—LBSN Structure. (n.d.). Retrieved September 1, 2020, from https://lbsn.vgiscience.org/concept-structure/

Structure Definition—LBSN Structure. (n.d.). Retrieved August 31, 2020, from https://lbsn.vgiscience.org/structure/

Summary—LBSN Structure. (n.d.). Retrieved August 28, 2020, from https://lbsn.vgiscience.org/

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, *59*(2), 64–73. https://doi.org/10.1145/2812802

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, *42*(sup1), 70–74. https://doi.org/10.1080/15230406.2015.1059251

Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M., Heatherly, R., & Malin, B. A. (n.d.). *A Game Theoretic Framework for Analyzing Re-identification Risk: Supporting Information*. 5.