Privacy Implications of New York City’s Stop-and-Frisk Data

Nick Manfredi

Submitted to the Department of Computer Science in Partial Fulfillment of the Prerequisite for Honors in Computer Science

WELLESLEY COLLEGE

2015
Privacy Implications of New York City’s Stop-and-Frisk Data

by

Nick Manfredi

Submitted to the Department of Computer Science
in Partial Fulfillment of the Prerequisite for Honors in Computer Science

Abstract

A wider release of police datasets could enable social scientists, community activists, and civil libertarians to more effectively challenge discriminatory policing practices. However, the privacy implications of such sharing must be carefully considered. It is known that “de-identifying” data is not sufficient to protect the privacy of individuals.

New York City’s stop-and-frisk data is an example of one such police dataset. The stop-and-frisk data contains information (including demographic information) about all people stopped by the program from 2003 to 2012. This paper examines the identifiability of this data. It examines the uniqueness of the data to investigate the privacy implications of its release for the individuals targeted by stop-and-frisk. It also suggests ways to re-identify this data.

Thesis Supervisor: Darakhshan J. Mir
Title: Norma Wilentz Hess Fellow in Computer Science
Acknowledgments

I would like to thank my advisor, Darakhshan Mir, for introducing me to the field of data privacy, encouraging me to write a thesis, and providing invaluable advice and assistance along the way.

I would also like to thank Lyn Turbak, Ben Wood, and Dr. Ismar Volic for sitting on my thesis committee and providing feedback on this paper.

Thank you to Dr. Latanya Sweeney and Sean Hooley at Harvard University for talking with me about re-identification approaches, and to “Team Frank” (Kenyon Vafa, Christian Haigh, Samuel Peinado, Daniel Kang, and Ben Misch) for sharing their ideas.

Thank you to Dr. Jennifer Musto for talking with me about the social and historical context of the stop-and-frisk program.

Thank you to T., for his support and help proofreading.

I would also like to thank the New York Civil Liberties Union for releasing this data and challenging the stop-and-frisk program, as well as everyone involved in the struggle against police violence.
# Contents

1 Introduction 10

1.1 Stop-and-Frisk Background ................................. 10
1.2 The Stop-and-Frisk Dataset ................................. 11
   1.2.1 Stop-And-Frisk Data Characteristics ................. 13
   1.2.2 Sensitive Information in the Stop-and-Frisk Dataset .. 15
1.3 Outline .................................................. 17

2 Background and Related Work 19

2.1 De-Identification, Re-Identification .......................... 19
2.2 History of Data Privacy Techniques ........................... 21
2.3 Re-Identifying the Stop-and-Frisk Data ........................ 22

3 Uniqueness Methods 25

3.1 Challenges .............................................. 26
3.2 Quasi-Identifiers ......................................... 28
3.3 Sampling .................................................. 31
   3.3.1 Sampling Example ................................. 31
3.4 Binning .................................................... 31
   3.4.1 Binning Example .................................. 35
3.5 Uniquification ........................................... 35
   3.5.1 Uniquification Example ............................. 38

4 Uniqueness Results 42
List of Figures

1.1 Average number of sensitive attributes per stop, 2012 . . . . . . . . 18

3.1 First three figures of the sampling algorithm example using sex, race
and date of birth as the quasi-identifier. . . . . . . . . . . . . . . . . . 33
3.2 Continued sampling example. . . . . . . . . . . . . . . . . . . . . . . 34
3.3 An illustration of the binning uniqueness algorithm using sex, race and
date of birth as the quasi-identifier. . . . . . . . . . . . . . . . . . . . 36
3.4 An illustration of the uniquification uniqueness algorithm using all
combinations of two quasi-identifiers. . . . . . . . . . . . . . . . . . . . 40
3.5 Continued uniquification algorithm example. . . . . . . . . . . . . . 41

4.1 Sampled uniqueness results for all people in the 2012 NYCLU dataset. 44
4.2 Sampled uniqueness results for all people in the 2003-2012 NYPD dataset. 45
4.3 Sampled uniqueness results for all people with photo identification in
the 2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . . . . . . . . 47
4.4 Sampled uniqueness results for all people with photo identification in
the 2003-2012 NYPD dataset. . . . . . . . . . . . . . . . . . . . . . . . 48
4.5 Sampled uniqueness results for African Americans with photo identifi-
cation in the 2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . . 50
4.6 Sampled uniqueness results for women with photo identification in the
2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
4.7 Binning uniqueness results for Luis Daniel’s quasi-identifier (precinct,
sex, race, date of birth and height) in the NYCLU dataset. . . . . . . 55
4.8 Binning uniqueness results for Harvard’s quasi-identifier (sex, race, date of birth, height, eye color, hair color and build) in the 2012 NYCLU dataset. .............................................. 56

4.9 Binning uniqueness results for sex, race, date of birth, height and eye color in the 2012 NYCLU dataset. .............................................. 57

4.10 Binning uniqueness results for sex, race, date of birth in the 2012 NYCLU dataset. .............................................. 59

4.11 Binning uniqueness results for the 2003-2012 NYPD dataset. .............. 60

A.1 Sampled uniqueness results for African Americans in the 2012 NYCLU dataset. .............................................. 73

A.2 Sampled uniqueness results for people aged 25 and older in the 2012 NYCLU dataset. .............................................. 74

A.3 Sampled uniqueness results for people aged 25 and older with photo identification in the 2012 NYCLU dataset. ......................... 75

A.4 Sampled uniqueness results for Asians in the 2012 NYCLU dataset. ... 76

A.5 Sampled uniqueness results for Asians with photo identification in the 2012 NYCLU dataset. .............................................. 77

A.6 Sampled uniqueness results for Black Hispanics in the 2012 NYCLU dataset. .............................................. 78

A.7 Sampled uniqueness results for Black Hispanics with photo identification in the 2012 NYCLU dataset. .............................................. 79

A.8 Sampled uniqueness results for Native Americans in the 2012 NYCLU dataset. .............................................. 81

A.9 Sampled uniqueness results for Native Americans with photo identification in the 2012 NYCLU dataset. .............................................. 82

A.10 Sampled uniqueness results for White Hispanics in the 2012 NYCLU dataset. .............................................. 83

A.11 Sampled uniqueness results for White Hispanics with photo identification in the 2012 NYCLU dataset. .............................................. 84
A.12 Sampled uniqueness results for whites in the 2012 NYCLU dataset. . 85
A.13 Sampled uniqueness results for whites with photo identification in the
2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . 86
A.14 Sampled uniqueness results for women in the 2012 NYCLU dataset. . 87
A.15 Binning uniqueness results for sex, race, date of birth, height and
weight in the 2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . 88
A.16 Binning uniqueness results for sex, race, date of birth and height in
the 2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . . . . . . . 89
A.17 Binning uniqueness results for sex, race, date of birth and weight in
the 2012 NYCLU dataset. . . . . . . . . . . . . . . . . . . . . . . . . 90
List of Tables

1.1 Percent of stops by gender. .................................................. 13
1.2 Percent of stops by race. ..................................................... 14
1.3 For each gender stopped, percent of stops by race. .................... 14
1.4 Percent of stops in each of the five boroughs of New York. ........... 14

4.1 Racial makeup of all people vs. only those with photo ID in the 2012 NYCLU data. .................................................. 52
4.2 Gender breakdown of all people vs. only those with photo ID in the 2012 NYCLU data. .................................................. 52
4.3 Average age of all people vs. only those with photo ID in the 2012 NYCLU data. .................................................. 52
4.4 Most common birthdays of all people vs. only those with photo ID in the 2012 NYCLU data and the number of people sharing those birthdays. 53
4.5 Percent unique with different numbers of attribute combinations in the 2012 NYCLU dataset. .................................................. 58
4.6 Number of possible values for some of the examined quasi-identifiers. 61

5.1 Number of matches for quasi-identifier instances that match Tyquan Brehon’s characteristics during the time period (2003-2008) in which we know he was stopped. .................................................. 64
5.2 Results of the re-identification experiment with the top 100 most frequently stopped individuals with photo ID in the 2012 NYCLU data, using Harvard’s quasi-identifier [30]. .................................................. 66
Chapter 1

Introduction

Widespread data-sharing enables researchers, community planners and activists to make informed decisions about topics of interest to their communities. However, datasets about individuals can reveal private and personally identifiable information. To prevent unnecessary privacy breaches, data holders often omit or redact information that is seen to be uniquely identifying, like name, social security number, and address. Yet de-identifying data in this way is not sufficient to protect privacy: research has shown that 87% of individuals can be uniquely identified by their gender, date of birth, and zip code [24]. Such attributes, referred to in the literature as “quasi-identifiers,” are not directly identifying by themselves, but can be combined with outside, personally identifying information (like voter records) to re-identify people. This paper will examine the potential for unintentionally privacy breaches in New York City’s stop-and-frisk dataset.

1.1 Stop-and-Frisk Background

New York City’s stop-and-frisk program was drafted in 2001, following the 1999 murder of unarmed African immigrant Amadou Diallo, who was shot forty-one times by four white police officers. The 2001 legislation required police officers in New York City to report detailed information about the hundreds of thousands of pedestrians that they stop, question, and frisk every year, including information about the
victims’ race [19]. The data was first released in 2007, as a result of legal action by the New York Civil Liberties Union (NYCLU), and it showed a large growth in stop-and-frisk stops from the period of 2002 to 2006 [19].

Community activists and civil liberties groups have viewed the growth of the program with contempt, arguing that stop-and-frisk disproportionately targets poor Black and Latino New Yorkers [18]. The data played an important role in challenging the program, since it showed that stop-and-frisk disproportionately targeted People of Color and was ineffective at achieving its stated aim of reducing the number of guns on the streets [18]. For instance, the NYCLU report [19] showed that:

1. Young Black and Latino men make up just 4.7% of the city’s population, yet they make up 41% of all stops.

2. 90% of young Black and Latino men stopped were innocent (not issued a summons or arrested).

3. Black and Latino New Yorkers were more likely to be frisked, but less likely to have a weapon than frisked whites.

4. Guns were recovered in only one-sixth of one percent of all the stops.

Consequently, the program is being phased out (at least in name) following court rulings and mayor Bill de Blasio was elected in part due to his opposition to the program [21].

1.2 The Stop-and-Frisk Dataset

The stop-and-frisk data contains 101 different categories of information, including fields specific to the individual being stopped: date of birth, race, sex, height, build, eye color, and hair color [29] [8]. It is possible that combining such fields with outside information (like social media profiles, voter records, criminal databases, or personal knowledge of an individual fitting the description) might be sufficient to uniquely identify individuals in the dataset. This presents researchers and activists with a
double bind situation: the data release enabled researchers and community groups to challenge the harm caused by the stop-and-frisk program, yet such data could be abused to cause additional harm the individuals in the dataset. For instance, it is possible that a nosy neighbor or employer who knows that someone has been stopped by police, along with general physical characteristics of their target, could cross reference this information against the stop-and-frisk data to see exactly why their neighbor or potential employee was singled out for police harassment and would learn what occurred during the stop. This private information could then be used discriminate against them in some way. It is therefore necessary to examine the risks and benefits of such data disclosure.

We cannot predict what future datasets will be released, and such datasets could make it easier to compromise the privacy of people in this dataset. However, any such privacy breach would require that individual’s data be unique in the stop-and-frisk dataset, since this would allow an adversary to cross reference the stop-and-frisk data with an additional data source to definitively re-identify the individual stopped: if more than one individual shares the same characteristics, an adversary would not be able to link a name to a particular stop with certainty. To examine the potential for harm posed by the stop-and-frisk data, this thesis examines both: the extent to which the personal attributes contained in the dataset are uniquely identifying; and ways to re-identify this data.

In all parts, this paper focuses on the data released for the year 2012 by the NYCLU [29], though data from 2003 to 2012 released by the New York Police Department (NYPD) [8] is occasionally examined as well. This is motivated by the fact that we are studying the uniqueness of the data in order to get at individuals’ representation in the data, and we consider factors like height and weight which are likely to change dramatically over longer periods of time. This is particularly important considering that 49.9% of all stops over all years are of people aged 14-24, whose height and weight are likely to fluctuate more than those of older people [19]. Additionally, 2012 is the most recent year that the NYPD reported over half a million stop-and-frisk stops (there were 532,911), down from a peak of 685,724 in 2011 [19]. Due to political
pressure and challenges to the program, the number of stops decreased dramatically to 191,851 in 2013 [19]. Additionally, the 2013 data does not contain dates of birth, which is a major factor in the uniqueness of the data and, consequently, the privacy risks associated with its release. Finally, the 2012 data is sanitized by the NYCLU in a way that preserved the maximum amount of information from the raw data: incomplete and improperly coded records were fully reconstructed, without deleting any entries. For example, in some years of the NYPD dataset all the stops from Staten Island are improperly coded and no stops from this borough appear in the data. Additionally, in some years of the NYPD data, the number of stops involving individuals of unknown sex meets or exceeds the number of stops involving women, whereas in the NYCLU data all stops are coded as either male or female.

### 1.2.1 Stop-And-Frisk Data Characteristics

It is necessary to understand the basic characteristics of the NYCLU data, since it will inform the experiments that follow. Of the 532,911 stops in 2012, the majority involved men: men were stopped 487,065 times, accounting for 91.4% of the individuals stopped, as seen in Table 1.1. Women were stopped 38,062 and are 7.1% of all stops. Additionally, the majority of stops involved African Americans: African Americans were stopped 284,229 times (53.3% of all stops), white Hispanics were stopped 129,368 times (24.3%), whites were stopped 50,366 times (9.5%), Black Hispanics were stopped 35,772 times (6.7%), Asian and Pacific Islanders were stopped 17,058 times (3.2%), and American Indians and Native Alaskans were stopped 2,257 times (0.4%) (Table 1.2). The race of people stopped was similar for each gender, except that white women were more likely to be stopped when compared to all women than white men compared to all men (Table 1.3).

<table>
<thead>
<tr>
<th>Gender of All Stops</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>91.4%</td>
</tr>
<tr>
<td>Women</td>
<td>7.1%</td>
</tr>
</tbody>
</table>

Table 1.1: Percent of stops by gender.
Race of All Stops

<table>
<thead>
<tr>
<th>Race</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>53.3%</td>
</tr>
<tr>
<td>White Hispanic</td>
<td>24.3%</td>
</tr>
<tr>
<td>White</td>
<td>9.5%</td>
</tr>
<tr>
<td>Black Hispanic</td>
<td>6.7%</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>3.3%</td>
</tr>
<tr>
<td>American Indian/Native Alaskan</td>
<td>0.4%</td>
</tr>
</tbody>
</table>

Table 1.2: Percent of stops by race.

<table>
<thead>
<tr>
<th>Race by Gender of Stops</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>54.9%</td>
<td>53.3%</td>
</tr>
<tr>
<td>White Hispanic</td>
<td>25.0%</td>
<td>24.0%</td>
</tr>
<tr>
<td>White</td>
<td>9.4%</td>
<td>13.7%</td>
</tr>
<tr>
<td>Black Hispanic</td>
<td>7.0%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>3.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td>American Indian/Native Alaskan</td>
<td>0.4%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Table 1.3: For each gender stopped, percent of stops by race.

Additionally, 34.4% of all stops occurred in Brooklyn, 20.8% in Queens, 20.7% in Manhattan, 19.2% in the Bronx, and 3.9% in Staten Island (Table 1.4). The most common recorded height was 68 inches (13.7%), and the most common weight was 160 pounds (13.0%). 40.2% of all stops had no date of birth recorded, which is significant for this paper’s later uniqueness and re-identification studies.

<table>
<thead>
<tr>
<th>Borough</th>
<th>Percent of stops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brooklyn</td>
<td>35.4%</td>
</tr>
<tr>
<td>Queens</td>
<td>20.8%</td>
</tr>
<tr>
<td>Manhattan</td>
<td>20.7%</td>
</tr>
<tr>
<td>Bronx</td>
<td>19.2%</td>
</tr>
<tr>
<td>Staten Island</td>
<td>3.9%</td>
</tr>
</tbody>
</table>

Table 1.4: Percent of stops in each of the five boroughs of New York.

In addition to information about the individuals stopped, the data contains a number of fields related to the stop-and-frisk incident. For instance, a summons or arrest was issued in only 11.1% of all stops, meaning 88.9% stopped were innocent, according to the NYCLU’s definition [19]. Furthermore, only 11.1% of stops were due to a suspicion of violent crime. At least one act of force by police was used in
17.3% of all stops (which includes an officer using their hands, putting the suspect on the ground or against a wall, drawing their weapon, pointing their weapon, using their baton, using handcuffs, pepper spray, or other acts of force). Finally, guns were recovered in only 714 stops, or 0.01% of the total stops made in 2012.

1.2.2 Sensitive Information in the Stop-and-Frisk Dataset

To get a sense of the amount of information that an attacker could gain from re-identifying the stop-and-frisk dataset, it is useful to consider the content and amount of the information coded in each stop. In addition to the demographic information, the 101 fields of the dataset include information about why stops were made and what occurred during the stop. Linking this data back to an individual may harm their reputation or livelihood, even if they were completely innocent of any crime (as 88.9% of people in the data were). While any information could potentially constitute sensitive information (for instance, knowing where and when someone was stopped might be sensitive if that person did not want their location to be known), we consider the following fields to be particularly sensitive, since they tend to cast the individual in a distinctly unfavorable light:

1. The crime code describing the crime involved.
2. The offense for which the individual was arrested.
3. The offense for which a summons was issued.
4. The crime that the individual was suspected of.
5. The reason the officer used force against the individual in the data.
6. That the individual was frisked.
7. That the individual was searched.
8. That an arrest was made.
9. That a summons was issued.
10. That contraband was found.

11. That a weapon was found on the suspect, including: a pistol, rifle, assault weapon, knife or cutting instrument, machine gun, assault weapon, or other weapon.

12. That the officer used force against the suspect in the following ways: using his hands, placing the suspect on the ground, placing the suspect against the wall, drawing a weapon, pointing a weapon, using a baton, using handcuffs, using pepper spray, or using other force.

13. The reason for the stop, including: carrying a suspicious object, fitting a relevant description, casing a victim or location, acting as a lookout, wearing clothes commonly used in a crime, actions indicative of a drug transaction, furtive movements, actions of engaging in a violent crime, or a suspicious bulge.

14. The reasons for a frisk, including: violent crime suspected, other suspicion of weapons, inappropriate attire for season, actions of engaging in a violent crime, refusal to comply with the officer’s directions, verbal threats by the individual, knowledge of the individual’s prior criminal behavior, furtive movements, having a suspicious bulge, having a hard object, having the outline of a weapon, or an admission by the individual.

15. The basis for a search, including: having a hard object, having the outline of a weapon, or the admission by the individual.

16. Any additional circumstances, including: proximity to the scene of an offense, evasive response to questioning, associating with known criminals, changing direction at the sight of an officer, being in an area with a high crime incidence, being seen at a time of day that fits a crime incidence, sights or sounds of criminal activity, a report by a victim, witness, or officer, or an ongoing investigation.

These sensitive attributes are used to assess how much sensitive information is contained in each stop. The results are presented in Figure 1.1, which reports the
average number of sensitive attributes per stop, with error bars to indicate the variance.

1.3 Outline

The rest of this paper is organized as follows: Chapter 2 discusses related work in the field of data privacy, including other work done with the Stop-and-Frisk dataset as well as foundational work on data linkage attacks and data privacy. Chapter 3 discusses the methods used for assessing the data’s uniqueness. The stop-and-frisk dataset is rich with information, so deciding which information to use to examine the uniqueness and re-identification potential of the dataset is an open question that this paper explores. Likewise, the dataset is different from those explored by previous authors in that an individual might be stopped multiple times, and Chapter 3 presents the methods used to account for this. Chapter 4 then presents the results of these uniqueness studies. Chapter 5 goes on to discuss some methods for re-identifying the dataset, as well as the results of these methods. Finally, Chapter 6 concludes the paper with a discussion of the findings.
Figure 1.1: Average number of sensitive attributes per stop, 2012
Chapter 2

Background and Related Work

This chapter provides an overview of previous research and work related to this paper. Specifically, it examines earlier work on de- and re-identification of datasets, the development of data privacy techniques, and other studies involving New York’s stop-and-frisk data.

2.1 De-Identification, Re-Identification

Previous research has shown that de-identifying data by removing names and other characteristically identifying attributes, like social security numbers and addresses, is not enough to protect the privacy of individuals in datasets. The term “quasi-identifier” was introduced by Tore Dalenius in 1986 [5] to refer to pieces of partially identifying information, which themselves are not uniquely identifying, but can be combined with other outside information to uniquely identify an individual. Quasi-identifiers have since been used to re-identify a number of different datasets, showing that the practice of scrubbing only traditionally identifying attributes from datasets is insufficient to protect the privacy of the individuals in the data.

In 1997, Latanya Sweeney showed that it was possible to re-identify de-identified hospital records of state employees released by The Massachusetts Group Insurance Commission by linking them with publicly available information. She demonstrated this by locating then-governor William Weld’s records, who had been admitted to
the hospital after a much publicized collapse at a college commencement ceremony, and presented them to him [25] [2]. Similarly, in 2013 Sweeney showed that some de-identified hospital records could be linked to newspaper stories [27]. Additionally, in 2013 Sweeney, Abu, and Winn re-identified individuals in the Personal Genome Project by linking their publicly available profile information to voter lists, as well as mining the documents for names [28]. Moreover, in 2008 Narayan & Shmatikov showed that it was possible to re-identify data released by Netflix as part of a competition by linking it to reviews on the movie review site IMDB [17].

Other researchers have also exposed the insecurity of deidentified data. The work of de Montjoye et al. [7] found that “mobility traces are highly unique,” since the vast majority people in a dataset can be uniquely identified by four spatiotemporal points. This paper replicates some of the methods of de Montjoye’s paper, while also adding new methods to examine the uniqueness of subpopulations in the data.

On the other hand, Solomon et al. find that merely identifying unique combinations of attributes does not necessarily pose a re-identification threat [23]. This paper examined a large, sparse social science dataset and found individuals are similar to their nearest neighbors in the dataset, which mitigates the risk of re-identification through unique attributes. Solomon et al. use a cosine similarity test to measure the difference between individuals in the dataset, which is similar to the test used by Narayan & Shmatikov in their attack on the Netflix data [17]. This paper borrows methods from this paper to explore how “uniquifiable” the stop-and-frisk data is.

In previous work, this author, Mir and Schlenker examined data from tollgates in the center of Milan and conducted tests to examine the uniqueness of tollgate usage [14]. We also conducted an inference attack identifying 8.5% of vehicles in December based on their November traces [14]. The first part of this paper uses similar methodology to this process. Additionally, since we do not have a “ground truth” on which to build an inference attack, this paper attempts to individually re-identify a small number of people based on the data.
2.2 History of Data Privacy Techniques

Foundational work in the area of data privacy has proposed countermeasures to protect the privacy of individuals in data releases. In 2002, Sweeney proposed “k-anonymity” [26], which is a method for redacting and generalizing data. The aim of Sweeney’s “k-anonymity” procedure is to ensure that no record is unique, and that each record’s personally identifying attributes (which Sweeney considers to be things like zip code, gender, or race) resemble at least $k - 1$ other records in the dataset.

In 2007, Machanavajjhala et al. revealed that $k$-anonymity provides insufficient protection when the sensitive attributes of a data lack diversity: for instance, if every person with the same zip code, gender and race in the data has the same medical condition, say heart disease, then it’s possible to figure out that your target has heart disease [12]. Likewise, Machanavajjhala et al. show that $k$-anonymity fails when an adversary has background knowledge: for instance, if the adversary knows their target is female and every other matching record in the database has prostate cancer as the sensitive attribute, then the adversary can figure out what condition their target has through a process of elimination. To address these limitations, the paper proposes $l$-diversity, which ensures that the sensitive attributes, like a medical condition, are also diverse.

Li et al. found weaknesses in $l$-diversity, however, and proposed $t$-closeness in its stead [11]. Specifically, the paper criticised $l$-diversity as difficult to achieve and argued that if the distribution of sensitive attributes is skewed or sensitive attributes are similar to each other, $l$-diversity is insufficient to protect privacy. For example, if 99 out of 100 people sharing a zip code, race and gender with a target have heart disease, it’s very likely that the target does too. Additionally, if all the sensitive conditions for those people who match the target are related to the stomach (the paper cites gastric ulcers and gastritis), then an adversary will learn that their target has an illness related to this part of the body. $t$-closeness aims to ensure that all sensitive attributes are evenly distributed and dissimilar for all people sharing the same personal information (like gender, age, race).
In contrast to this “alphabet soup” of ad hoc proposals for data privacy, Cynthia Dwork proposed mathematically rigorous definition of data privacy, called differential privacy [10]. Unlike previous works, differential privacy does not attempt to distinguish between background information and the sensitive attribute and does not depend on the content of the data. Instead, differentially private techniques ensure that the results of data analysis are about the same with or without any one individual’s participation. In doing so, they protect unusual individuals against a knowledgeable adversary, who might have an unlimited amount of outside information about a participant. To achieve this, differential privacy techniques add a small amount of noise to the data. The trade off then is between utility and privacy in the dataset.

Differential privacy has been applied to many different fields and types of data [15]. In 2013, Mir et al. constructed a differentially private model of human mobility based on Call Detail Records [16]. Similarly, this author, Mir, Lu and Sanchez created a differentially private, synthetic model of the data that preserved two of its key characteristics with little error [13]. Future work could involve generating differentially private, synthetic models of the stop-and-frisk data.

2.3 Re-Identifying the Stop-and-Frisk Data

Previous research provides techniques useful for de-anonymizing the stop-and-frisk data. Luis Daniel created a quasi-identifier based on the reported date of birth, gender, race and exact height of the individuals stopped, along with the precinct in which each stop occurred [6]. Yet this approach may be flawed: we know from the NYCLU’s report that the most frequently stopped demographic are Latino and African American men of age 14-24 (41%)[19], yet Daniel’s approach led to a median age of 24.7 for people stopped more than 5 times [6]. Daniel notes that his quasi-identifier will tend to under count young people (who make up the bulk of the stops) because his quasi-identifier only identifies people as being the same person if their height is consistent. However, the height of teenagers is likely to increase over time. Similarly, the inclusion of precinct in the quasi-identifier is misleading: precinct refers
to the precinct where the stop was made, not the precinct where the victim lived. Daniel’s method finds that the top five precincts where people are stopped more than five times are all relatively remote to the city center (where most stops occur). It is possible that this is because these locations might have less easy access to public transportation and their residents may be less mobile. Accordingly, the method misses victims who were stopped multiple times in multiple precincts. Finally, the author assumes that this quasi-identifier can lead to easy re-identification. This is based on a finding made by Latanya Sweeney, who showed that “87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex” [1, 24]. Yet this is far from clear for a number of reasons: we do not have the victim’s zip code, the victims tend to be homogeneous in terms of race, sex, and age, and Sweeney’s findings rely on voter records which may be less useful in poor, minority communities where people are more likely to be disenfranchised. This paper will examine Luis Daniel’s quasi-identifier throughout, and compare it with other possible quasi-identifiers.

A team of Harvard students also considered ways to re-identify the stop-and-frisk data [30]. They proposed a quasi-identifier based on sex, race, date of birth, height, eye color, hair color, and build. Yet it is possible that this quasi-identifier is overly specific, since information such as hair color and build may not be consistently coded throughout the data, and thus might under-count the number of stops per individuals in the data. This paper examines this quasi-identifier as well. The Harvard team also proposes a method for re-identifying the 100 most frequently stopped individuals using New York State voter records, which this paper examines. They recognize the limitations of this approach: the young, urban, African American and Latino men disproportionately targeted by the stop-and-frisk program may be underrepresented in the voter records. Likewise, they recognize that individuals’ most frequently stopped location may not necessarily correspond to their home zip code. Nevertheless, the data does not contain information about the victims’ home location, so it is necessary to approximate it through other means in order to apply the results from Sweeney [24]. Finally, they suggest that arrest records or mug shots might be used
to re-identify individuals who were arrested. This is explored in Chapter 5 in the discussion of re-identification methods.
Chapter 3

Uniqueness Methods

In order to examine the risks of re-identification posed by the stop-and-frisk dataset, this paper considers the uniqueness of the stop-and-frisk data. As explained in Chapter 2, to re-attach names to de-identified records in a dataset the records must be unique in some way. For example, if an adversary knew that a 75 year old Native American woman was stopped and frisked, and knew nothing else about her or the stop, in order to determine with certainty when and why the police had stopped her and what the outcome of the stop was, there would need to be only one such woman in the dataset. If there were two such records, the adversary would learn about the conditions of the stop with a varying degree of certainty: if both stops had the same exact conditions, then the adversary would know for certain why the woman was stopped, but if they diverged then the adversary would have narrowed it down to two possible sets of conditions.

For all parts of these uniqueness tests, records that are missing any of the traits under consideration are excluded. This is significant in the case of date of birth, which is missing in 40.2% of all stops (it is coded as 12/31/1900 in the data).

Additionally, in all parts of the uniqueness tests, this paper examines the number of matches for nine different groups:

1. All people

2. People aged 25 and older
3. Women

4. Asian and Pacific Islanders (abbreviated as “Asian” throughout this paper)

5. Black people

6. Black Hispanics

7. American Indians and Native Alaskans (shortened to ”Native Americans”)

8. White people

9. White Hispanics

These constraints were selected in order to examine the threat posed to outliers, like women and older individuals, as well as to highlight the different threats posed to individuals of different races. The racial categories correspond to those found in the NYCLU and original NYPD data. This paper explores the hypothesis that people who are not young Black and Latino men would be more likely to be unique in the data, and thus have a higher risk of re-identification.

3.1 Challenges

The data presented some challenges not found in other studies, such as that of de Montjoye et al. [7] and [14]. First, while we would like to examine the uniqueness of an individual in the dataset, each record in the data corresponds to a specific stop rather than an individual. Thus, if an individual is stopped multiple times, they will have multiple records associated with them. The data itself does not contain a ground truth for identifying which records correspond to the same individual. To compensate for this, we examine different combinations of demographic traits and hypothesize that definitions with more specific combinations will do a better job of singling out an individual without including other individuals. Additionally, we attempt to address this issue by reporting the average number of matches for each
quasi-identifier, in addition to the percent of stops involving a unique quasi-identifier. This is intended to capture the fact that individuals may be stopped multiple times.

On the other hand, we recognize that there may be errors in the adversary’s knowledge of their target as well as in the reporting or encoding of the data. For example, an adversary may know only a subset of their target’s attributes, or may only have a rough idea of their target’s height or weight. Likewise, police may guess the target’s height or weight, or the target may misreport it. Additionally, as was mentioned in the discussion of the dataset, youth are disproportionately targeted by this program, and their height, weight, and build is likely to change over time. So, whereas we prefer more specific quasi-identifier combinations in order to single out individuals, there is a limit to how specific such combinations can be: if we make the quasi-identifier too specific it may count a large number of individuals more than once.

Similarly, some attributes are more likely to be accurately coded than others. Whereas, for the most part, it might be easy to determine someone’s sex reliably and consistently, their build (“heavy,” “muscular,” “medium,” or “thin”) might be inconsistently coded depending on who is doing the coding and the shape the person happens to be in at the time of the stop. Likewise, some hair and eye colors (like “brown” and “black”) may be easily confused. Height and weight, as previously mentioned, are also likely to change over time and may not be coded accurately if, say, the police officer guesses, the person stopped refuses to participate, or the person stopped does not know their exact height or weight. Race too may present some challenges, since it is not always possible to accurately classify an individual by their appearance and an individual may fail to self-report their race or report it in a way that does not align with the available categories. Finally, without photo identification or the participation of the person stopped, it might be hard to reliably determine their exact date of birth. In the 2012 NYCLU data, it seems that most unknown birthdays are coded as 12/31/1900 (40.2% of all stops), yet the two most frequent birthdays are 1/1/1992 and 1/1/1990, with 156 and 153 stops, respectively. This is slightly higher than the 149 stops for the third most frequent birthday, 9/16/1995. While the fact
that these two most frequent birthdays both occur on the first of the year suggests that some stops may have just used these dates as convenient placeholders, it is not clear how to separate these placeholder dates from real data.

To account for the possibility of error in the data, this paper also examines those stops where the individual presented photo identification (though this population does differ slightly from the population as a whole, see Tables 4.1, 4.2 and 4.3). In the state of New York, drivers licenses and non-driver identification cards include gender, height, eye color, and date of birth fields [22, 4]. While other types of identification exist, such as permanent resident cards and passports, we assume that these two would be the most frequent forms of identification presented. Under this assumption, we pay special attention to the uniqueness properties of individuals stopped who presented photo identification using gender, height, eye color, height, date of birth, and race as a quasi-identifier. This does assume that police officers are filling at the form according to the identification they’re presented with, and also that an individual’s race would be recorded consistently.

3.2 Quasi-Identifiers

To account for these challenges, this paper examines up to 28 different combinations of traits, representing different levels of outside knowledge and accuracy in the coding and reporting of the data. It also reports on those stops where the person stopped presented photo identification.

The trait combinations analyzed are enumerated below. Numbers 1 and 2 correspond to the quasi-identifiers proposed by Harvard [30] and Luis Daniel [6], respectively. 3-9 examine quasi-identifiers that correspond to what is found on New York identification (sex, date of birth, exact height, eye color), as well as race, which we assume could be determined consistently for most individuals, and weight with various tolerances to account for errors in judgement, reporting, or natural fluctuations over time. 10-15 examine sex, race, date of birth, height and weight, with varying levels of tolerance on both height and weight. These were considered because, for those stops
that do not involve photo identification, it seems possible that eye color may not be coded consistently. 16-20 examine sex, race, date of birth, and height (with different levels of tolerance), whereas 21-27 examine sex, race, date of birth, and weight (also with different levels of tolerance) to account for cases where height or weight may be missing, coded incorrectly, or unknown by an adversary. Finally, 28 examines only sex, race, and date of birth under the assumption that these three characteristics are perhaps the least subject to change, error or uncertainty: in most cases, sex and race can be determined by an individual’s appearance, and date of birth, as previously mentioned, is typically coded as 12/31/1900 if it is unknown.

1. Sex, race, date of birth, exact height, eye color, hair color, build (as in [30])

2. Precinct, sex, race, date of birth, exact height (as in [6])

3. Sex, race, date of birth, exact height, eye color, exact weight

4. Sex, race, date of birth, exact height, eye color, weight +/- 10 pounds

5. Sex, race, date of birth, exact height, eye color, weight +/- 20 pounds

6. Sex, race, date of birth, exact height, eye color, weight +/- 30 pounds

7. Sex, race, date of birth, exact height, eye color, weight +/- 40 pounds

8. Sex, race, date of birth, exact height, eye color, weight +/- 50 pounds

9. Sex, race, date of birth, exact height, eye color

10. Sex, race, date of birth, exact height, exact weight

11. Sex, race, date of birth, height +/- 1 inches, weight +/- 10 pounds

12. Sex, race, date of birth, height +/- 2 inches, weight +/- 20 pounds

13. Sex, race, date of birth, height +/- 3 inches, weight +/- 30 pounds

14. Sex, race, date of birth, height +/- 4 inches, weight +/- 40 pounds
15. Sex, race, date of birth, height +/- 5 inches, weight +/- 50 pounds

16. Sex, race, date of birth, height +/- 1 inch

17. Sex, race, date of birth, height +/- 2 inches

18. Sex, race, date of birth, height +/- 3 inches

19. Sex, race, date of birth, height +/- 4 inches

20. Sex, race, date of birth, height +/- 5 inches

21. Sex, race, date of birth, exact weight

22. Sex, race, date of birth, weight +/- 10 pounds

23. Sex, race, date of birth, weight +/- 20 pounds

24. Sex, race, date of birth, weight +/- 30 pounds

25. Sex, race, date of birth, weight +/- 30 pounds

26. Sex, race, date of birth, weight +/- 40 pounds

27. Sex, race, date of birth, weight +/- 50 pounds

28. Sex, race, date of birth

Additionally, under the assumption that photo identification would allow an accurate, consistent recording of the subject’s height, we omit those quasi-identifiers that involve a tolerance on height for those studies involving stops with photo identification. Finally, an exact height is added to the quasi-identifiers that examine sex, race, date of birth and weight for studies involving only those stops where photo identification was presented.

To examine the uniqueness of the data, this paper uses three methods: sampling, binning, and “uniquification,” which are described below. With sampling and binning, we obtain information about how unique stops in the dataset are using the quasi-identifiers listed above for a sample of the data (sampling) and the entire dataset
(binning). With uniquification, we examine how uniquely identifying any number of attributes are, without regard to the quasi-identifier categories previously mentioned.

### 3.3 Sampling

The sampling procedure begins by selecting an initial record at random that contains all of the traits under consideration (for example, sex, race and date of birth). It then searches all the data and keeps track of the number of matches for the trait combination. A set of 1,000 different initial records are selected in this way, and the average number of matches is reported along with the variance. Additionally, we examine the percentage of these 1,000 selected initial records that had only one match (themselves) and report this as the “percent unique.” Algorithm 1 shows how the number of matches for each of the sampled 1,000 records is determined. It returns a list with the number of matches, from which the average number of matches as well as the percent unique (where there is only one match, the chosen target) are determined. This is similar to the methods used in [7] and [14].

#### 3.3.1 Sampling Example

Figure 3.1 and Figure 3.2 illustrate the sampling algorithm on a small, simplified version of the dataset that contains only sex, race and date of birth fields. It assumes the quasi-identifier combination under consideration is sex, race and date of birth.

### 3.4 Binning

Like the sampling algorithm, the binning algorithm calculates the average number of matches and the percent unique, with a subtle difference: the sampled algorithm reflects the average number of quasi-identifier matches and percent unique for each individual stop, whereas the binning algorithm shows the average number of matches and percent unique for each quasi-identifier “bin.” The binning uniqueness algorithm is also more computationally efficient, making it possible to run on all individuals in
Algorithm 1 Sampled Uniqueness Algorithm

1: function UNIQUENESS(data, quasiIDList)
2:     numMatchResultList ← []
3:     i ← 0
4:     while i < 1000 do  \(\triangleright\) Sample 1,000 people
5:         target ← Random.choice(data)
6:         if isMissingQuasiID(target, quasiIDList) then \(\triangleright\) If the
7:             continue  \(\triangleright\) randomly chosen target is missing any of the quasi-identifier fields, pick another
8:         end if
9:         numMatches ← 0
10:        for all other in data do
11:            isMatch ← True
12:            for all quasiID in quasiIDList do
13:                if target.quasiID ≠ other.quasiID then \(\triangleright\) If any
14:                    isMatch ← False
15:                end if
16:            end for
17:            if isMatch then
18:                numMatches ← numMatches + 1
19:            end if
20:        end for
21:     numMatchResultList.append(numMatches)
22:     i ← i + 1
23: end while
24: return numMatchResultList
25: end function
(a) Select an initial record at random (M, Black, 9/3/1994) and compare it against the first record in the dataset for all fields under consideration (here sex, race, date of birth). This one does not match because the date of birth fields do not match.

(b) Continue by comparing the selected record in the same way against every other record in the dataset.

(c) Eventually there will be at least one match for all the quasi-identifier fields under consideration when the randomly selected record matches itself. This is recorded as a match.

Figure 3.1: First three figures of the sampling algorithm example using sex, race and date of birth as the quasi-identifier.
(a) The selected record also matches another in the dataset for all considered quasi-identifier fields (here sex, race and date of birth), so the count of matches is incremented.

(b) When the randomly selected record has been compared against all others, record the number of records that matched all of the considered quasi-identifier fields (sex, race and date of birth in this case). In this example, the number of recorded matches is 2.

Figure 3.2: Continued sampling example.
the data. The procedure for determining the number of matches per quasi-identifier bin is presented in Algorithm 2, which returns a list from which the average number of matches per bin, the variance in this average number, and the percent of bins holding only one individual (“percent unique”) are obtained.

Algorithm 2 Binning Uniqueness Algorithm

1: procedure UNIQUENESS(data, quasiIDList)
2:   qidDict ← dict()
3:   for all stop in data do
4:     if isMissingQuasiID(stop, quasiIDList) then ▷ If the stop is missing any of the quasi-identifier fields, skip this stop and don’t report it in the final results.
5:       continue
6:   end if
7:   quasiIDString ← “,”.join(quasiIDList)
8:   if quasiIDString in qidDict then
10:    else
11:     qidDict[quasiIDString] = 1
12:   end if
13: end for
14: return list(qidDict.values())
15: end procedure

3.4.1 Binning Example

Figure 3.3 illustrates the binning algorithm on a small, simplified version of the dataset that contains only sex, race and date of birth fields. It assumes the quasi-identifier combination under consideration is sex, race and date of birth.

3.5 Uniquification

Finally, this paper examines the percent of stops that can be made unique, or “uniquified” by any one, two or three combinations of attributes. This is similar to the work presented by Solomon et al. [23], and “uniquify” as a verb comes from this work. Rather than using all 101 attributes, this paper limited the attributes to the follow-
(a) Select the first record in the dataset and enter the quasi-identifier fields under consideration (M, Black, 2/4/1990) as a key in a hash table, setting the value to 1.

(b) Continue in the same way for every other record in the dataset. Whenever a new quasi-identifier instance is encountered (here F, Black, 3/9/1994), add a new key to the hash table.

(c) When a record is encountered whose quasi-identifier instance already exists in the hash table, increment the count of its value.

Figure 3.3: An illustration of the binning uniqueness algorithm using sex, race and date of birth as the quasi-identifier.
ing, which we assumed might reasonably be known by an adversary wishing to learn more about the details of their target’s stop:

1. The precinct of the stop.

2. The borough of the stop.

3. The sex of the individual stopped.

4. The race of the individual stopped.

5. The date of birth of the individual stopped.

6. The age of the individual stopped.

7. The height of the individual stopped.

8. The weight of the individual stopped.

9. The hair color of the individual stopped.

10. The eye color of the individual stopped.

11. The build of the individual stopped.

12. Whether an arrest was made.

13. Whether a summons was issued.

The first two attributes are specific to the stop, yet might be inferred by an adversary who knows the general whereabouts of their target. The last two attributes are also specific to the stop, yet would result in a public record of an arrest or a summons under the targeted individual’s name. All of these attributes are numbers that can take on a limited range of values, so no one will be trivially uniquified as they would be by a string value that could vary due to typos. This paper does not include other attributes specific to the stop itself, because we assume that the adversary does not know these specifics and is attempting to learn more about the stop. An
adversary who witnessed a stop, however, would likely be able to locate the record in the dataset using the time, date, and location of the stop.

The uniquification algorithm is presented in Algorithm 3. It takes as input the data as well as a list of all possible combinations of the $n$ attributes under consideration, where $n$ ranges from 1 to 3. For example, if we were only considering sex, race, and date of birth (“dob”), the $attributeCombinationList$ when $n = 2$ would consist of $[[sex, race], [sex, dob], [race, dob]]$. Meaning, with the above 13 attributes and the length of combination $n$ (ranging from 1-3), the length of the $attributeCombinationList$ is length $\binom{13}{n}$. It returns the number of individuals in the dataset for whom some attribute combination in $attributeCombinationList$ is uniquely identifying. This algorithm differs from [23], however, as unique attributes which by themselves uniquify an individual are counted toward the unique-by-2-attributes uniquification percentage. The unique-by-3-attributes uniquification percentage likewise includes those attributes that, by themselves or as a pair, would uniquify an individual.

### 3.5.1 Uniquification Example

Figure 3.4 and Figure 3.5 illustrate the uniquification algorithm on a small, simplified version of the dataset that contains only sex, race and date of birth fields. It assumes the number of quasi-identifier options under consideration is two.
Algorithm 3 Uniquification Algorithm

1: procedure Uniquify(data, attributeCombinationList)
2:     numUnique ← 0
3:     for all target in data do
4:         for all attributeCombination in attributeCombinationList do
5:             match ← 0
6:                 for all other in data do
7:                     allAttributesMatch ← True
8:                         for all attribute in attributeCombination do
9:                             if target.attribute ≠ other.attribute then
10:                                allAttributesMatch ← False
11:                                   break
12:                             end if
13:                         end for
14:                 if allAttributesMatch then
15:                     match ← match + 1
16:                 end if
17:             end for
18:         if match == 2 then \(\triangledown\) If there are 2 matches for this attribute combination, it is not a unique combination so break to save time and move onto the next attribute combination.
19:             break
20:         end if
21:     end for
22:     if match == 1 then \(\triangledown\) If there’s only 1 match for this attribute combination, increment the number unique and break. It is not necessary to look at any more attribute combinations, since the goal is to see if any combination can uniquely identify a person.
23:         numUnique ← numUnique + 1
24:     break
25: end if
26: end for
27: return numUnique
28: end procedure
(a) Select the first record in the dataset and the first combination of 2 quasi-identifier options (here sex, race). Compare the selected record against every other record in terms of the 2 chosen quasi-identifier options. This record matches itself, so the count of matches is incremented.

(b) Continue in the same way for every other record in the dataset. As soon as a second match is encountered, the quasi-identifier combination (here sex, race) is not unique, so the next combination can be considered without comparing the selected record against the rest of the records for the chosen (sex, race) combination.

(c) Begin again, comparing the first record against every other record in the dataset for the next quasi-identifier combination (here sex, date of birth). This record matches itself.

Figure 3.4: An illustration of the uniquification uniqueness algorithm using all combinations of two quasi-identifiers.
(a) Continue comparing the selected record (M, Black, 2/4/1990) against all others in the dataset for the selected quasi-identifier combination.

(b) The end of the dataset has been reached and the selected record (M, Black, 2/4/1990) only matched itself in terms of sex and date of birth. Record that this record was “uniquified” by a combination of two quasi-identifiers.

(c) Begin again, comparing the second record against every other record in the dataset for the first quasi-identifier combination (here sex and race).

Figure 3.5: Continued uniquification algorithm example.
Chapter 4

Uniqueness Results

Our methods find that the stop-and-frisk data is highly uniquely identifying. That is, for all examined quasi-identifiers, only a small number (relative to the total number of stops) share all the attributes of the quasi-identifier. The results for the sampled uniqueness algorithm 1 are presented in Figure 4.1 through 4.6, and in Appendix A in Figure A.1 through Figure A.14. The results for the binning algorithm are presented in Figure 4.7 through Figure 4.11 and Figure A.15 through Figure A.17 in Appendix A.

4.1 Sampling Results: Stops With and Without Photo ID

Figure 4.1a shows that, for all people in 2012 using the NYCLU data, the average number of matches for each of the 1,000 randomly selected stops ranged from 1.1 to 10.5, depending on the quasi-identifier. The maximum variance was 98.47 for sex, race, date of birth and weight plus or minus 30 pounds. As expected, the most specific quasi-identifiers had the lowest number of average matches (sex, race, date of birth, exact height, exact weight, both with and without the inclusion of eye color, resulted in 1.1 average matches). Likewise, the highest number of matches corresponded to the most genera, sex, race, date of birth quasi-identifier. The percent of stops that
were unique with this data is presented in Figure 4.1b. Also as expected, the more specific quasi-identifiers resulted in a greater number of unique stops: for example, using sex, race, date of birth, exact height, and weight as a quasi-identifier, 90.8% of the sampled stops had no match other than themselves. Using sex, race, and date of birth, only 10.3% of stops were unique.

These trends hold throughout the all groups of data that were examined: more specific quasi-identifiers lead to a lower number of average matches, with sex, race, date of birth, exact height, exact weight, and (optionally) eye color standing out as a particularly identifying combination. Additionally, Luis Daniel’s quasi-identifier (precinct, sex, race, date of birth, and exact height [6]) uniquely identifies a large number of stops in the data and results in a low average number of matches, as seen in column 2 of all the graphs.

Figure 4.2 presents the results from sampling the NYPD’s 2003-2012 data for all people. Across the board, the average number of matches is higher than in the NYCLU’s 2012-only data, sometimes considerably so as seen in Figure 4.2a. This average number ranges from 1.8 for all three highly specific quasi-identifiers to 84.2 for sex, race, and date of birth. Likewise, the percent of stops that were unique is lower considerably lower, ranging from a high of 69.7% for Luis Daniel’s quasi-identifier to a low of 2.0% for sex, race and date of birth, as seen in Figure 4.2b. Figure 4.2a also shows that there was a large variance in the average number of matches: the maximum variance recorded was 54,995.70 for sex, race, date of birth, exact height, eye color and weight plus or minus 40 pounds. The maximum number of matches for that quasi-identifier combination was 3,332.

There are a number of factors that might be responsible for the discrepancy between the NYCLU and NYPD numbers. First of all, since the NYPD data spans ten years instead of one, there is a larger window of time for an individual to be stopped. For example, if an individual whose attributes are distinct from everyone else in the dataset were stopped once per year, they would be unique in the NYCLU data but not in the NYPD data. Additionally, there are 4,792,543 total stops in the 2003-2012 NYPD data but only 532,912 stops in the NYCLU’s 2012 data. More stops means
Figure 4.1: Sampled uniqueness results for all people in the 2012 NYCLU dataset.
Figure 4.2: Sampled uniqueness results for all people in the 2003-2012 NYPD dataset.
that there is more of a chance for multiple individuals who share the same quasi-
identifier attributes to be stopped. Finally, the NYCLU data was sanitized in a way
that captured a good deal of information, including information that wasn’t present
in the NYPD data (for example, assigning a binary sex to stops marked “unknown” in
the NYPD dataset). This information may be more accurate, and thus more uniquely
identifiable, than incorrect or filler data from the NYPD dataset.

4.2 Sampling Results: Only Stops with Photo ID

This paper also examines only those stops where photo identification was presented in
order to examine whether the standard information presented on the photo identifica-
tion increases the uniqueness of the data. Accurate coding of individuals’ attributes
in the data is essential for re-identification, since this is what an attacker would use to
locate their target in the data. We hoped that by considering only those presenting
photo identification, the attributes recorded would more accurately reflect the ground
truth and would be more consistent over time. Indeed, as seen in Figure 4.3, stops
involving photo identification do have a lower average number of matches than in the
data as a whole: the sex, race, date of birth quasi-identifier, for instance, results in
an average of 5.6 matches as opposed to the previous 10.5 in Figure 4.3a. Likewise,
17.2% of the sampled sex, race, date of birth quasi-identifier attributes are unique,
as opposed to 10.3% for all the NYCLU data in Figure 4.3b.

The sampled uniqueness results for the NYPD’s 2003-2012 data are presented in
Figure 4.4b. As with the NYCLU data, photo identification increases the uniqueness
of the data and decreases the average number of matches. The average number
of matches for Harvard’s quasi-identifier, for example, is 2.3 for stops with photo
identification and 3.5 for all stops. Likewise, for the Harvard quasi-identifier, the
percent unique increases from 44.8% to 55.4%.

Finally, this paper examines the uniqueness of sub-groups in the NYCLU data,
including all the racial categories (African American, Asian, Black Hispanic, Native
American/Pacific Islander, White and White Hispanic), women only, and stops with
Figure 4.3: Sampled uniqueness results for all people with photo identification in the 2012 NYCLU dataset.
Figure 4.4: Sampled uniqueness results for all people with photo identification in the 2003-2012 NYPD dataset.
people aged twenty-five and older. Overall, we found that the more prevalent a group was in the data, the less unique and the higher the average number of matches they had for all categories. For example, we can contrast the uniqueness results of African Americans, who make up 53.3% of all 2012 NYCLU stops and are the most frequently stopped race in the data, against women, who make up only 7.1% of all 2012 NYCLU stops (as reported in section 1.2.1). For 2012 NYCLU stops involving photo identification, the photo ID quasi-identifier (sex, race, date of birth, height and eye color) results in 1.7 average matches for African Americans (Figure 4.5a) but only 1.0 average matches for women (Figure 4.6a). Similarly, the percent of African American stops that are unique using the photo ID quasi-identifier is 62.4% (Figure 4.5b), whereas it is 96.5% for women (Figure 4.6b).

Section A.1 presents additional results from the sampled uniqueness study, which bear out the trends described above. These graphs present the results for all races individually, as well as for women only and only people twenty-five years of age or older in the 2012 NYCLU data.

4.2.1 Discussion: Sub-Populations and Photo Identification

As expected, sub-populations such as women, Native Americans and Pacific Islanders, and Asians, who make up only a tiny number of the total stops in the NYCLU data are more unique than those groups that are over-represented in the data, such as African Americans (Table 4.1). This makes sense, since there are fewer individuals in the data for them to match. It is also possible that individuals from groups that do not fit a traditional profile, such as women, may be less likely to be seen as suspicious and thus would not be stopped repeatedly. Overall, this means that groups that are disproportionately targeted by police violence may be the least likely to be harmed by the release of police data.

The increased uniqueness of stops with photo identification might be the result of several different factors. First, individuals may be stopped repeatedly and only present identification during some of those stops. This would lead to a lower number of average matches and a higher chance of a unique stop due to the fact that their
Figure 4.5: Sampled uniqueness results for African Americans with photo identification in the 2012 NYCLU dataset.
Figure 4.6: Sampled uniqueness results for women with photo identification in the 2012 NYCLU dataset.
non-photo identification stops aren’t included in the photo ID studies. Additionally, individuals who presented photo identification when stopped tended to be older (and thus more unique, as will be discussed below): on average photo ID holders were 30 years old, as opposed to 28 years old for all people, as seen in Table 4.3. Moreover, Blacks and Black Hispanics made up a slightly smaller percentage of the photo identification data than they did in the overall data, as seen in Table 4.1. On the other hand, women (who make up a minority of the data and whose stops are thus more often unique) accounted for fewer of the stops in the photo ID data than in the overall data (Table 4.2).

<table>
<thead>
<tr>
<th>Race</th>
<th>Black</th>
<th>Black Hispanic</th>
<th>White Hispanic</th>
<th>White</th>
<th>Asian</th>
<th>Native American</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Photo</td>
<td>51.60%</td>
<td>6.54%</td>
<td>25.49%</td>
<td>11.77%</td>
<td>4.08%</td>
<td>.53%</td>
</tr>
<tr>
<td>All</td>
<td>54.80%</td>
<td>6.87%</td>
<td>24.90%</td>
<td>9.71%</td>
<td>3.29%</td>
<td>.43%</td>
</tr>
</tbody>
</table>

Table 4.1: Racial makeup of all people vs. only those with photo ID in the 2012 NYCLU data.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Photo</td>
<td>6.97%</td>
<td>93.03%</td>
</tr>
<tr>
<td>All</td>
<td>7.22%</td>
<td>92.78%</td>
</tr>
</tbody>
</table>

Table 4.2: Gender breakdown of all people vs. only those with photo ID in the 2012 NYCLU data.

<table>
<thead>
<tr>
<th>Average Age</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>With Photo</td>
<td>30.00</td>
</tr>
<tr>
<td>All</td>
<td>28.12</td>
</tr>
</tbody>
</table>

Table 4.3: Average age of all people vs. only those with photo ID in the 2012 NYCLU data.

Nevertheless, while it is not possible to be certain, it seems plausible that some of the increased rate of uniqueness comes from the fact that the photo ID-only data more accurately reflects the ground truth of those stopped and involves less guessed information. Whereas the two most common birthdays in the data as a whole were reported to be January 1 (of 1992 and 1990), the four most common birthdays in
Table 4.4: Most common birthdays of all people vs. only those with photo ID in the 2012 NYCLU data and the number of people sharing those birthdays.

<table>
<thead>
<tr>
<th>Most Common Birthdays</th>
<th>All Stops</th>
<th>Photo ID Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>1/01/1992: 156</td>
<td>10/30/1992: 53</td>
</tr>
<tr>
<td>5th</td>
<td>11/02/1991: 126</td>
<td>1/01/1990: 48</td>
</tr>
</tbody>
</table>

The data involving those with photo ID are seemingly random dates (10/30/1992, 3/04/1991, 8/12/1994, and 8/30/1993). The fifth most common birthday in the photo ID set, however, is 1/01/1990. Additionally, in the data from all people there is a larger gap in the number of people stopped with the most common birthdays, whereas this gap disappears in the photo ID set (Table 4.4). This birthday diversity suggests that police officers might be using the photo identification to record information for the stop, meaning this information might be less likely to vary over time due to misreporting by the individual stopped or differences in the police officer’s judgment. Similarly, it seems possible that police may guess the value of other attributes in systematically biased ways: for instance, they may be more likely to record that someone is 6 foot as opposed to 6 foot 1 inches or 5 foot 11 inches. Due to this, it is possible that the stops where photo ID was provided might be more accurate and might provide a better basis for identifying a persistent individual in the data, and thus might provide a better ground truth for re-identification.
4.3 Binning Results

Figures 4.7, 4.8, and 4.9 present the binning results using the Luis Daniel, Harvard, and photo identification quasi-identifiers, respectively. These graphs exhibit the same trends as noted in the above section 4.1, and serve to present the information in a more concise way. From these graphs, we see that, using Luis Daniel’s quasi-identifier, every group has close to only one average match (Figure 4.7a) and every group is over 90% unique (Figure 4.7b). The Harvard quasi-identifier resulted in slightly fewer unique matches, with a minimum of 83.7% unique for African Americans with and without photo identification (Figure 4.8b). Finally, the photo identification quasi-identifier results in the lowest uniqueness, and only 68.6% of all African American stops are unique (Figure 4.9b). Binning results from other quasi-identifiers are found in Appendix A.2.
Figure 4.7: Binning uniqueness results for Luis Daniel’s quasi-identifier (precinct, sex, race, date of birth and height) in the NYCLU dataset.
Figure 4.8: Binning uniqueness results for Harvard’s quasi-identifier (sex, race, date of birth, height, eye color, hair color and build) in the 2012 NYCLU dataset.
Figure 4.9: Binning uniqueness results for sex, race, date of birth, height and eye color in the 2012 NYCLU dataset.
4.4 Uniquification Results

The uniquification results are presented in Table 4.5. While any one attribute rarely uniquifies any record in the data, any two attributes “uniquify” nearly half of the data (42.75%) and with any three attributes, more than half (64.31%) of entries are uniquified. Since this is a computationally expensive algorithm, we only examine up to three attributes.

<table>
<thead>
<tr>
<th>Number of attributes</th>
<th>Percent Unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46%</td>
</tr>
<tr>
<td>2</td>
<td>42.75%</td>
</tr>
<tr>
<td>3</td>
<td>64.31%</td>
</tr>
</tbody>
</table>

Table 4.5: Percent unique with different numbers of attribute combinations in the 2012 NYCLU dataset.

4.5 Discussion

Overall, the data contains enough information to uniquely identify a majority of the stops based on demographic characteristics alone. Of the quasi-identifiers examined, more specific quasi-identifiers and more unique data lead to fewer average matches observed, which is to be expected. We can approximate the specificity of the quasi-identifier by considering how many possible values it can take on: there are 123 different police precincts, 6 different races, 2 different sexes, 11 different hair colors, 8 different eye colors, and 4 different builds. The Harvard study assumes birthdays can fall in a 30 year span and that height can take on 15 different values [30], and it seems reasonable to assume that weight could take on $300/5 = 60$ different values, since it is most frequently coded in increments of 5 and we can say that most older teenagers and adults (who tend to be stopped) probably weigh between 80 and 380 pounds. Using these approximations, Table 4.6 displays the number of possible values that each quasi-identifier can take on. The sex, race, dob, height, eye, weight is by far the most specific, followed by the Harvard study’s quasi-identifier [30], which is followed by Luis Daniel’s quasi-identifier [6].
Figure 4.10: Binning uniqueness results for sex, race, date of birth in the 2012 NYCLU dataset.
Figure 4.11: Binning uniqueness results for the 2003-2012 NYPD dataset.
Table 4.6: Number of possible values for some of the examined quasi-identifiers.

<table>
<thead>
<tr>
<th>Quasi-identifier</th>
<th>Number of possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex, race, dob, height, eye, hair, build</td>
<td>693,792,000</td>
</tr>
<tr>
<td>precinct, sex, race, dob, height</td>
<td>242,433,000</td>
</tr>
<tr>
<td>sex, race, dob, height, eye, weight</td>
<td>946,080,000</td>
</tr>
<tr>
<td>sex, race, dob, height, eye</td>
<td>15,768,000</td>
</tr>
<tr>
<td>sex, race, dob, height, weight</td>
<td>118,260,000</td>
</tr>
<tr>
<td>sex, race, dob, height</td>
<td>1,971,000</td>
</tr>
<tr>
<td>sex, race, dob, weight</td>
<td>7,884,000</td>
</tr>
<tr>
<td>sex, race, dob</td>
<td>131,400</td>
</tr>
</tbody>
</table>

The risk of an overly-specific quasi-identifier, however, is that it will fail to identify all of an individual’s stops, which may have been coded differently from stop to stop. In the case of Daniel’s quasi-identifier, an individual will not be connected to stops occurring in different precincts. A broader quasi-identifier, such as sex, race, and date of birth, however, may mistakenly attribute other people’s stops to one individual. For a compromise that is potentially less prone to error (as noted above), we suggest that sex, race, date of birth, eye color, and height be used to identify an individual, since its specificity falls in the middle of the range. Also, all of these attributes (except for race) can be found consistently coded on common forms of photo identification in New York. When examining stops involving photo identification, this quasi-identifier may perform the best. Nevertheless, we do not have a ground truth, so all we can observe is the relative impact of the different quasi-identifiers on the percent unique and average number of matches per stop.
Chapter 5

Re-identification Methods and Results

This chapter presents some possible methods for re-identifying, or attaching names, to the entries in the stop-and-frisk dataset.

5.1 Nosy Neighbors and Employers

The first and most straightforward method of re-identification would be personal knowledge of someone who has been stopped and frisked. A nosy neighbor or employer, for instance, might learn that an acquaintance was the subject of a stop-and-frisk incident. Such a person, being familiar with their target, might know their sex, race, date of birth, eye color, hair color, build and approximate height and weight. The results from Chapter 4 show that this information would likely be enough to identify an individual in the data and get a handful of stops associated with them. This would in turn reveal all the sensitive information associated with the stop discussed in Chapter 1.
5.2 Social Media

Social media and the Internet can also be used to re-identify individuals in the data. For instance, the New York Times wrote an article about Tyquan Brehon. Mr. Brehon lives “in one of the most heavily policed neighborhoods in Brooklyn” and was stopped more than 60 times before his 18th birthday [9]. The pictures in the article suggest that he is African American, and we know from the article that he is male. There is only one profile belonging to someone named Tyquan Brehon on Facebook, and the publicly visible images match the man in the article. His birthday is also publicly visible on Facebook. Using his visible characteristics, those mentioned in the article, and the information gleaned from the Internet we can query for stops matching Tyquan Brehon in the dataset.

The results of querying for Mr. Brehon’s characteristics are presented in Table 5.1. It assumes his hair and eye color were both recorded as either black or brown and only queries the data from 2003 to 2008 (before his 18th birthday, the period in which the article tells us he was stopped multiple times). In addition, we search for stops in Brooklyn for some of the quasi-identifiers, assuming that he was most likely stopped in his “heavily policed” neighborhood (though he may have been stopped in other cities). We note that city is not recorded in the 2003 dataset, so the considered quasi-identifiers involving city will undercount the number of matches. Nevertheless, it may be safe to assume that most of his stops would have occurred after 2003, since Mr. Brehon was only 12 or 13 in that year and thus outside of the age range in which people are usually stopped.

We do not know for sure that all of these stops belong to Mr. Brehon. However, if at least some of the time his relevant information was coded correctly, we can assume that some of his stops are captured in the matching stops. This significantly narrows the field of stops that could belong to Mr. Brehon. Perhaps more importantly, though, even if none of these stops are Mr. Brehon, an adversary may falsely believe that they do belong to him and discriminate against him accordingly. An incorrect assumption based on this data may still lead to negative effects for those in the dataset.
Table 5.1: Number of matches for quasi-identifier instances that match Tyquan Bre-hon’s characteristics during the time period (2003-2008) in which we know he was stopped.

<table>
<thead>
<tr>
<th>Quasi-identifier</th>
<th># of matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOB, race, sex, hair, eye, city</td>
<td>29</td>
</tr>
<tr>
<td>DOB, race, sex, city</td>
<td>32</td>
</tr>
<tr>
<td>DOB, race, sex, hair, eye</td>
<td>70</td>
</tr>
<tr>
<td>DOB, race, sex</td>
<td>76</td>
</tr>
</tbody>
</table>

5.3 Criminal Records

Another method of re-identifying individuals in the data would be to search for criminal records or mugshots, as suggested in the Harvard study [30]. This has some limitations, however. We know that 88.9% of the stops in the data do not result in a summons or arrest. Additionally, arrest records and other police information are not made public in the state of New York. What would be public, however, are the court records associated records. “Court records are subject to disclosure under Section 255 of the Judiciary Law” according to O’Connor, who used them to map all the summonses in New York City [20]. While this would re-identify those stops resulting in court proceedings, it seems that there would be more information in court proceedings themselves than in the the stop-and-frisk dataset, and thus the dataset might not significantly contribute to the harm faced by the individual. This paper does not pursue this method of re-identification.

5.4 Voter Records

Finally, the Harvard study suggests using public voter information to re-identify individuals in this dataset [30]. They suggest using the sex, race, date of birth, height, hair color, eye color, build quasi-identifier to identify the 100 most frequently stopped individuals, and then to assume that the zip code where they were most frequently stopped is their home zip code. The combination of sex, date of birth, and zip code should be enough to uniquely identify 87% of individuals in the voting records [24]. There are a number of shortcomings to this, which the authors address. For instance,
people may be stopped more often in locations that do not correspond to their home location. Also, the young African American men that make up the bulk of the stops are less likely to be registered voters, both because they are often too young to be eligible and because low-income People of Color are less likely to vote [30].

By submitting a FOIL request, we were able to obtain the voter records for New York State. We tested both the Harvard quasi-identifier as well as the photo ID quasi-identifier, using the methods suggested by Harvard on the 2012 NYCLU data. This involved only using entries with photo identification in order to increase the likelihood that the information was correctly coded in the stop-and-frisk data.

A complication arose from the fact that the zip code field is never filled out in the stop-and-frisk data. Instead, we used the State Plane Coordinate System of 1983 x- and y-coordinates from the data, which were automatically generated at each stop. We converted them to latitude and longitude using Earth Point [3] and got the corresponding zip codes using Pygeocoder [31]. Occasionally, one of the top 100 most frequently stopped individuals had more than one most common stop zip code. In that case, we searched for all of the most common stop zip codes associated with the quasi-identifier instance.

Of the top 100 most frequently stopped individuals in the 2012 NYCLU data who had photo identification, using the Harvard quasi-identifier, 28 had only one match for their date of birth, sex, and zip code in the voter records. 75 had one or more matches in the voter records, with the average number of matches being 3.01 (Table 5.2). This shows that voter records may be a reasonable way to re-identify individuals in this data, though it is possible that this process incorrectly links people, since it can only correctly identify individuals who have registered to vote who were also stopped in their home zip code. Likewise, since we used the specific Harvard quasi-identifier, it is more likely that all the stops selected as belonging to one of the top 100 most frequently stopped individuals do indeed correspond to the same person. However, this method may miss some of an individual’s stops, if, for example, their build was coded different across the dataset.
| Percent with 1 [sex,dob,zip] Match in the Voter Records | 28% |
| Percent with any number of [sex,dob,zip] Matches in Voter Records | 75% |
| Average Number of [sex,dob,zip] Matches | 3.01 |
| Maximum Number of [sex,dob,zip] Matches | 13 |

Table 5.2: Results of the re-identification experiment with the top 100 most frequently stopped individuals with photo ID in the 2012 NYCLU data, using Harvard’s quasi-identifier [30].
Chapter 6

Conclusion

This paper examined the risks associated with the release of New York City’s stop-and-frisk data. While this data release was instrumental in ending the racially biased program, the release itself may bring further harm upon the people stopped under the program.

This paper found that each stop contains an average of 7.7 pieces of sensitive information. Moreover, an adversary who knows only three well-chosen pieces of basic demographic information can uniquely identify 64% of all stops. The vast majority of people can be uniquely identified in the 2012 NYCLU data by an adversary who knows enough information about their target: for instance, knowing a target’s sex, race, date of birth, eye color and exact height and weight makes 89.9% of the 2012 NYCLU stops unique and leads to just 1.1 average matches. In the NYPD data from 2003 to 2012, this combination of attributes makes 67.5% of the stops unique and leads to 1.8 average matches.

Moreover, this paper demonstrated that there are real ways for an adversary to re-identify this unique data: the most straightforward example of such an attack is a nosy neighbor who, by virtue of being acquainted with their target, would know basic information about the target. Since the data contains sensitive information that might not otherwise be known, such as the reason a police officer believed someone was suspicious, there is a risk of harm associated with its release.

Due to these risks, care must be taken when releasing data containing sensi-
tive information about individuals, particularly information about individuals from marginalized groups who might suffer the effects of discrimination more keenly than people of a higher socio-economic status. The NYPD’s 2013 data mitigates some of the threat by not including the date of birth field. Without this field, the studies in this paper would be much less successful in uniquely identifying individuals. Nevertheless, removing the date of birth field is not sufficient: the data contain additional information that could be used by a nosy neighbor to identify their target, such as the exact location where a stop was made (which might correspond to the target’s home address), and the time and date of the stop (which could be observed). More work must be done to assess the best way to this release data in a way that preserves its utility for activists seeking to challenge racist police practices. Techniques such as differential privacy, used by Mir et al. to release cellular records, might prove promising [16].
Bibliography


[27] Latanya Sweeney. Matching known patients to health records in washington state data. *Available at SSRN 2289850*, 2013.

[28] Latanya Sweeney, Akua Abu, and Julia Winn. Identifying participants in the personal genome project by name. *Available at SSRN 2257732*, 2013.


Appendix A

Additional Figures

The following sections contain additional graphs from the sampled (A.1) and binning (A.2) uniqueness studies. They provide additional information about the uniqueness of stops for each racial groups, women, and individuals who are at least twenty-five years of age.

A.1 Sampled Uniqueness Additional Figures
Figure A.1: Sampled uniqueness results for African Americans in the 2012 NYCLU dataset.
Figure A.2: Sampled uniqueness results for people aged 25 and older in the 2012 NYCLU dataset.
Figure A.3: Sampled uniqueness results for people aged 25 and older with photo identification in the 2012 NYCLU dataset.
Figure A.4: Sampled uniqueness results for Asians in the 2012 NYCLU dataset.
Figure A.5: Sampled uniqueness results for Asians with photo identification in the 2012 NYCLU dataset.
Figure A.6: Sampled uniqueness results for Black Hispanics in the 2012 NYCLU dataset.
Figure A.7: Sampled uniqueness results for Black Hispanics with photo identification in the 2012 NYCLU dataset.
A.2  Binning Uniqueness Additional Figures
Figure A.8: Sampled uniqueness results for Native Americans in the 2012 NYCLU dataset.
Figure A.9: Sampled uniqueness results for Native Americans with photo identification in the 2012 NYCLU dataset.
Figure A.10: Sampled uniqueness results for White Hispanics in the 2012 NYCLU dataset.
Figure A.11: Sampled uniqueness results for White Hispanics with photo identification in the 2012 NYCLU dataset.
Figure A.12: Sampled uniqueness results for whites in the 2012 NYCLU dataset.
Figure A.13: Sampled uniqueness results for whites with photo identification in the 2012 NYCLU dataset.
Figure A.14: Sampled uniqueness results for women in the 2012 NYCLU dataset.
Figure A.15: Binning uniqueness results for sex, race, date of birth, height and weight in the 2012 NYCLU dataset.
Figure A.16: Binning uniqueness results for sex, race, date of birth and height in the 2012 NYCLU dataset.
Figure A.17: Binning uniqueness results for sex, race, date of birth and weight in the 2012 NYCLU dataset.