# Angels and Daemons: Is More Knowledge Better Than Less Privacy? An Empirical Study on a K-anonymized Openly Available Dataset

*Completed Research Paper*

**Ferdinando Pennarola**
Bocconi University
Via Roentgen 1, Milan, Italy
ferdinando.pennarola@unibocconi.it

**Luca Pistilli**
Bocconi University
Via Roentgen 1, Milan, Italy
luca.pistilli@uniboconi.it

**Michael Chau**
The University of Hong Kong
Pokfulam, Hong Kong
mchau@business.hku.hk

## Abstract

*Many organizations are starting to make datasets, such as customer review data and service usage logs. To protect the privacy of involved individuals, these datasets are usually pseudonymized or anonymized before they are released. A method called k-anonymization is widely used in such open datasets. Recent literature showed that this method, however, can be unsafe and compromise individuals' privacy. In this paper, we address this problem by analyzing the New York Citi Bike dataset. Through our analyses, we show that given some generalized and payload data, it is possible to recover other payload data of an individual in the k-anonymized dataset. We also demonstrate that it is possible to achieve a high success rate in re-identification of records. These findings shed additional light on the weakness of the k-anonymization method, thus evidencing a trade-off between data availability and privacy protection. We finally provide some implications for both academics and practitioners.*

***Keywords:*** *Data privacy, Information security/privacy, Privacy/information privacy, Open Database*

## Introduction

The idea of an information society has never been as trendy as it is today: data, literally, powers everything and drives our everyday life. Moreover, in the upcoming blast due to IoT (Internet of Things) adoption, we envisage the information society at its upmost potential, as never before, to the extent that people start thinking that "we are no longer individuals, but clusters of data". Even though this quote may appear too strong, it holds some truth in it: humans leave continuous traces of their behaviors, which are carefully collected in datasets, and some of those are shared openly on the Internet.

This is the starting point of our research. We embarked on open data research project with a large dataset, and we asked ourselves to what extent this data holds some secrets. This is a very relevant issue that has attracted a political debate: how should our society guarantee more sophisticated privacy protection (Bennett, 1992)?

When data concerns individuals, anonymization represents a feasible technique to balance both need of information and privacy related problems (Chen et al., 2012). Specifically, k-anonymization is used to assess data degree of security. Generalization and suppression are used to ensure k-anonymity. These methods are generally applied on identifiers and quasi-identifiers only, since payload data is regarded as anonymous per se. In addition, suppressing payload data would leave nothing to analyze. Recent literature has already acknowledged that k-anonymization is not hundred percent safe: it can reveal much about single individuals, of groups of them, disclosing secrets in their behaviors, and thus breaching the privacy protection walls (Nergiz and Clifton, 2007). Therefore, current anonymization models put people's privacy in danger, even more so when payload data is made publicly available. On the other side of the coin, massive datasets have been made public with the openly declared, often legitimate, intent of disseminating research opportunities to academics, statisticians, engineers, and developers to discover more, and ultimately contribute to the society as a whole. We accurately embraced this tradeoff, by conducting a test of what it is known about k-anonymization techniques. Our objective is to make a preliminary assessment of the tradeoff between research opportunity enhancements offered by dataset owners (angels) and potential privacy protection violations carried out by dataset hackers (daemons).

The rest of this paper is structured as following. First, we provide a careful review of the concepts of data privacy and anonymization, two issues that represent the starting points of our research. Second, we present a set of general hypotheses, aimed at re-testing the k-anonymization limitations. For such purpose, we used an exemplifying database, the New York Citi Bike System Data. Third, we explain our methodology and show our results. Lastly, we come back to the original starting point: does the noble aim of contributing to science prevail over the consequences of privacy leakage? Finally, we conclude this paper by discussing key takeaways, both from an academic and a managerial perspective

## Theoretical Background

### *Data Privacy*

Privacy protection has been widely recognized as one of the hottest topics for the information systems community (Bélanger and Crossler, 2011; Smith et al., 2011). In this section, we provide an overview on privacy related issues. It must be noted that not only open datasets, but also those of more private nature must deal with all the due requirements to protect users' sensitive information (Lee et al., 2011). Driving forces originate from a multitude of sources: from political influence to data scientists' breakthroughs. The definition of privacy is not straightforward, especially when political parties are involved. In fact, apparently similar concepts can be framed in many different ways to accommodate different objectives. A study argues that some political parties tend to favor certain privacy definitions as a way to reach their goals and shape the general public's opinion (Epstein et al., 2014). The analysis has been carried out by observing discussion patterns at the Internet Governance Forum (IGF). The IGF is an organization hosting both state and non-state actors, whose decisions concerning privacy are consultative and non-binding (Epstein, 2013). During IGF debates, it has been noted that state parties are inclined to frame privacy as a security issue: the less the state knows, the less citizens are protected. On the other hand, civil parties favor rights-based arguments: freedom and openness are empty words without real privacy. Whatever equilibrium these opposing forces will reach, it will heavily affect how privacy is both perceived and legislated in the future. Besides political dialectics, the concept of privacy is more and more connected with technical advancements. In fact, its boundaries have been blurring as technology armed analysts with new tools. In the past, one of the most consolidated definitions has been the following: *"Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others"* (Westin, 1968).

In turn, informational privacy has been structured as the individual ability to control which information is given and who has access to them (Malhotra et al., 2004). This approach presents several flaws, ranging from the notion of information to more philosophical issues, such as human dignity. However, the main problem is that society needs to shift its attention from data-collection to data-processing (Mai, 2016). The researcher points out that while the public has been focusing on consent-based strategies (i.e. whether or not to provide personal data), data science evolved to allow organizations to produce new information, independent from the consented one. Target, the discount store retailer in the U.S., stands as the most exemplary case (Hill, 2012). By collecting voluntarily provided purchase data, Target was able to infer that one of its customers was pregnant. When a specific customer found a congratulation note in her mailbox, it turned out the company was the first one to know. Paradoxically, following Westin's privacy definition, one may argue that this piece of information belongs to Target as much as to the mother. In fact, the company produced it anew. Therefore, the existing privacy definitions and models, based on data-collection and consent strategies, must be integrated with a new "datafication" model that takes into account predictive analytics. It must not be regarded as a substituting approach, but as a supplementing one.

However, technical advancements generate also positive implications. In fact, if it is true that technology poses new threats to informational privacy, it also helps creating new protections, the so-called privacy-enhancing techniques. For example, pseudonymization is a procedure according to which personal identifiers (such as full name, date of birth, IP, etc.) are replaced by pseudo-IDs (Claerhout and DeMoor, 2015). Pseudo-IDs are random identifiers that have no link to the original information. Thus, information undergoes a process of irreversible cryptography. Although it may sound as little technology is involved, the actual implementation is very sophisticated. In their paper, Claerhout and DeMoor (2015) outline two existing methods: Batch Data Collection and Interactive Data Storage. The former envisages the installation of servers at the lower level of the data chain, where data suppliers operate (e.g., treating doctor). On a timely basis, data suppliers will send different data packages to two separate entities. First, payload data (i.e. non-identifying data) are sent directly to data collectors (higher end of the data chain, the actors that gather and store data for re-use). Secondly, identity data are pre-pseudonymized and sent to a TTP (Trusted Third Party) that is in charge to complete the pseudonymization process and transfer the data to the collector. By doing so, the TTP will never own both identity and payload data and, therefore, it will not be able to perform any analysis. In the integrated data storage, data is not stored at the local level, but directly in the collector's server. This server allows data suppliers to display nominative information, while collectors see pseudo-IDs only. The pseudonymization process is still executed by the TTP, which ensures pseudo-IDs cannot be reversed. This solution is highly practical when there are two parties that use the same data: practitioners and researchers. Although these two methods are regarded as highly secure, re-identification is still a possible threat, in some fields more than others. For example, genomics payload data must deal with the high correlation between genotype and phenotype, a detail that may easily uncover the identities behind pseudo-IDs (Claerhout and DeMoor, 2015).

Another important aspect concerns international regulations. In fact, the need to develop privacy laws goes beyond national boundaries. As the flow of data across private organization increases, so does the flow across countries. Therefore, international authorities long for an integrated privacy bill, which would make implementation and control far easier. Global convergence does exist, although mild. The so-called "First Principles" define a set of guidelines for fair information practice. However, the implementation of such principles differs widely as inherent philosophies vary across societies. Reidenberg (2000) anticipates that liberal markets would react in different ways as compared to socially protective systems. According to this study, convergence must start as co-regulation, and then evolve into legal transplantation, namely incorporating foreign bodies of law into one's system. Finally, the privacy issue is also an ethical matter. Organizations involved in the big data industry must acknowledge their responsibilities towards individuals. In fact, three potential sources of ethical conflict have been identified (Martin, 2015). First, there are issues linked to the data supply chain, both upstream and downstream, such as the creation of prejudices and price discrimination. Second, negative externalities not evident in the immediate data transaction might be generated, like in the case of predictive analytics used to infer behavior of people that opted out of data collection. Third, fostering of destructive demand for consumer data, such as pushing customer-facing companies to adopt deceitful means to collect data, could represent a critical issue as well.

Specifically, to this last point, as most data originate from transactions in customer-facing industries, many companies are now exploiting their datasets to make financial gains. Problems would arise if the data business would end up being more profitable than the original one. At that point, the pressure to collect more and more information may instigate unethical behavior. It must be noted that educating individuals could be as important as educating organizations. Fortunately, although they may have only a vague understanding of what big data is, internet users seem to be aware and worried about companies utilizing their information (Turow & Hennessy, 2007). In the next sub-section, we will take a closer look at one of the most renowned privacy model: anonymization. Also, focusing on the concept of unique identifier, anonymization and pseudonymization will be indirectly compared.

## *Anonymization and Unique Identifier*

In the previous section, we outlined the pseudonymization process. Such method allows full analyses at the individual level without observing real identities. Although it is a powerful procedure, it is not always implementable, especially with regard to open data. Therefore, another privacy-enhancing technique has been engineered: anonymization. Such methodology often envisages the removal of unique identifiers, namely those indexes that represent single individuals (Sutanto et al., 2013).

Privacy and anonymity are not exactly the same: the former obscures the facts, while the latter obscures the identities (Gritzalis, 2004). Nevertheless, organizations tend to treat these two concepts as one. As a consequence, anonymity has become an integral part of data protection. In any anonymization process, the first step requires the removal of all identifiers, such as name, mail address, and telephone number. As a second step, it is necessary to determine quasi-identifiers, namely other information that, alone or combined, allow the unique identification of a certain individual (Daries et al., 2014). As an example of the power of quasi-identifiers, Sweeney (2002) proved that 87% of the American population is uniquely identified by a combination of date of birth, ZIP code, and gender. In order to evaluate the degree of anonymity in a dataset, k-anonymization is the most common model. We take the following formal definition (Daries et al., 2014): *"A dataset is k-anonymous if any one individual in the dataset cannot be distinguished from at least k–1 other individuals in the same dataset. This requires ensuring that no individual has a combination of quasi-identifiers different from k–1 others."*

According to Sweeney (2002), when a dataset is not k-anonymous, data scientists can resort to two methods to make it so: suppression and generalization. While suppression simply entails deleting fields until k-anonymity is reached, generalization hinges on decreasing the level of detail of such fields. For example, if income were the crucial quasi-identifier, we would either remove it from the variable set or transform it into income class (e.g. <10k, 10k-50k, >50k). It must be noted that suppression can also be implemented by deleting single cells for certain variables. The effects of these methods are plainly visible when the datasets contain small groups of individuals that are identical along the quasi-identifiers. In fact, in these cases there are higher chances to find k or less people with the same features. When it comes to academic research, it must be noted that both suppression and generalization decrease the usefulness of data.

K-anonymization saw a fair deal of success in recent years. Even private organizations put some efforts to improve the usability of this model. For example, IBM developed a theoretical system to apply k-anonymization without using time-expensive procedures, such as sorting. Their goal is to provide a model such that optimal k-anonymization is achieved at the lowest cost (Bayardo & Agrawal, 2005).

However, the k-anonymization technique has its own downsides. First, since it does not include any randomization process, hackers may still infer about sensible individual data. For example, if a given human being is known to be in the database containing all patients affected by specific illnesses, then this implies that such person has one of the included diseases. Further, k-anonymization does not represent a good methodological choice when high-dimensional datasets have to be anonymized (Aggarwal, 2005). For example, De Montjoye et al. (2013) evidenced how, given 4 points, the unicity of mobile phone datasets may reach 95%. Moreover, Angiuli et al. (2015) revealed that k-anonymity technique might have the unintended consequence of messing results of a dataset in case of a suppressing in a disproportionate way data points with non-representative features. Such last problem, anyway, might be solved by altering algorithms used to k-anonymize datasets (Angiuli and Waldo, 2016). In the digital world, many organizations claim that they keep only anonymous records, while they are actually performing pseudonymization only (Nissenbaum & Barocas, 2014). They do indeed remove persistent identifiers (such as the physical addresses), but they are still able to recognize one specific user from the others (e.g. through pseudo-IDs). Although this makes it more difficult for an organization to send a salesman to our doorway, it does not impair the power of such company within its internet platform. For example, Amazon may obscure our private details but still suggesting new books based on our purchase record. In addition, our data may reveal information that was not initially disclosed. In fact, companies may be able to estimate correlation indexes between variables they have and information they want.

Anonymization is commonly considered more secure than pseudonymization. In fact, this latter permits to single out individuals, which may be especially relevant with datasets that have more than one row per person. On one hand, there is the need to enhance privacy protection by removing identifiers. On the other, both researchers and practitioners insist on the importance of individual-specific data to conduct useful analyses. Existing literature shows many attempts to fill the informational gap between anonymization and pseudonymization: for example, Butler (1982) tried to build a model to retrieve individual identities from available non-unique quasi-identifiers. Such model hinges on the trade-off between certainty/redundancy and efficiency. Namely, the more variables one includes in the artificial ID, the higher the chances to identify an individual. However, as the number of variables rises, the algorithm efficiency falls and the process becomes slower. In order to optimize the model, the analyst must estimate the impact of each quasi-identifier in yielding uniqueness. Other models focus on matching individuals across different databases (e.g. Du Bois, 1969). The researcher tried to match death-certificates with before-death questionnaires. About 99% of the observations were correctly linked. In extreme cases, when retrieving unique identifiers proves to be impossible, there are methods to minimize bias. For example, some longitudinal population studies are repeated over time and neglect that a portion of data may come from the same people. Clearly, observations from the same individuals are not independent and their correlation index is different from zero. In these cases, sandwich variance estimator can be used to correct for bias in variance (Mountford et al., 2007).

To summarize, we have shown that open data has many barriers to overcome in order to achieve its full potential. Among such barriers, privacy stands out as one of the most crucial factors, attracting both political and technical attention. In order to ensure privacy protection, several models exist: we focused primarily on pseudonymization and anonymization. Anonymization, and the deletion of unique identifier, is considered the most secure one. It is often put in place as k-anonymity, exploiting the techniques known as generalization and suppression. In the next empirical part, we will show how the current applications of k-anonymity, or similar frameworks, can still leave many traces about individuals. Such model must be expanded to account for the identifying power of apparently anonymous data. Borrowing vocabulary from the literature above, we will refer to identity data (i.e. data containing individual-specific details) as identifiers and quasi-identifiers, while apparently anonymous variables are to form the so-called payload data. As already debated in the literature, we bring one more evidence to the issue of k-anonymization: it does not fully preserve privacy protection, and in particular by knowing given variables, it is possible to recover another one. More formally:

> *H1: Given a set of publicly available variables of a dataset, it is possible to recover another variable.*

Moreover, we consequently claim that re-identification rate correctly accomplished through de-anonymization methods is significantly superior to simple random assignment:

> *H2: De-anonymization methods offer a significantly higher re-identification rate as compared to a random assignment.*

## Data and Methodology

The New York Citi Bike Dataset is used in this study. Citi Bike is New York's bike sharing provider, the organization in charge of managing shared cycling in New York. Since 2013, the company has moved the first steps toward open-data and started releasing much of its system information. In line with the latest government trends, the end-goal is to guarantee transparency while grasping improvement opportunities. A direct quote from the open data project website clearly says:

"*Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular? What days of the week are most rides taken on? We've heard all of these questions and more from you, and we're happy to provide the data to help you discover the answers to these questions and more. We invite developers, engineers, statisticians, artists, academics and other interested members of the public to use the data we provide for analysis, development, visualization and whatever else moves you.*"

And indeed, the availability of the dataset has improved ancillary services. For example, although regarded as an unofficial source, the Google Group, called "BikeNYC and CitiBikeNYC Hackers", regularly offers insights on how to perfect business operations. We suppose that the company shares the data hoping that a group of independent data scientists, also known as bike hackers, will provide them with useful data-driven insights, such as station saturation times and bike sharing health impact. No unique identifier was available in the data. The data contains two quasi-identifiers: gender and year of birth. On its internet page, Citi Bike publishes multiple databases. Unless otherwise stated, in the following sections the word "dataset" will refer to Citi Bike Trip Histories, namely the data containing the following variables. It is important to notice that every entry carries information about a specific trip.

- trip duration: how long the trip took, measured in seconds
- start time: at what time the trip started
- end time: at what time the trip ended
- start station id: unique identifier for the station where the trip started
- start station name: name of the station where the trip started
- start station latitude: latitude coordinate for the start station
- start station longitude: longitude coordinate for the start station
- end station id: unique identifier for the station where the trip ended
- end station name: name of the station where the trip ended
- end station latitude: latitude coordinate for the end station
- end station longitude: longitude coordinate for the end station
- bike id: unique identifier for the bike that was used
- user type: type of contract owned by the rider: daily or yearly contract
- year of birth: rider's year of birth, a blank entry when it is not disclosed
- gender: rider's gender, a value of 0 was assigned to riders who did not wish to disclose this information

Data comes supposedly error-free. According to the Citi Bike website, trips taken by maintenance staff are removed. In addition, trips lasting less than 60 seconds are deleted as they are likely to reflect erroneous attempts to re-dock a bike. Citi Bike distributes the dataset on a monthly basis. Given the peculiar nature of this research, short-term results are easily generalizable to broader time-horizons. Therefore, July 2015 has been selected as the reference month, as it contains the highest number of trips (i.e. 1,048,575). On the other hand, May 2015 will be used as validation sample to check for overfitting issues. It can be noted that most fields are trip-specific, only few variables (i.e. user type, year of birth, gender) refer to the person who took the ride. However, it is still possible to deduce some characteristics of Citi Bike's customer base. Firstly, 83% of the trips are taken by annual subscribers, while 17% stem from daily passes. The average trip duration is 970 seconds. As far as gender is concerned, the majority of riders are male, accounting for 63% of the trips. Riders born in the 80s seem to be the most avid bike sharing users as they take almost 33% of the trips. Interestingly, Citi Bike does not require its customers (owning daily tickets) to share private information such as gender and year of birth. Conversely, subscribers must actively opt out if they do not want these details to be divulged. Looking at Table 1 below, it is clear that there are more subscribers concerned with their gender than their year of birth.

| **Category** | **# of rows** |
|---|---|
| Trips without gender | 1,012 |
| Trips without birth | 3 |
| Trips without both | 0 |

**Table 1 – Missing Variables**

In July 2015, Citi Bike counts 7,040 bikes and 330 stations (most of them in Manhattan), which gives an average of 23 vehicles per docking point. By analyzing the trip distribution by station, we can notice that the median (2,882) is slightly smaller than mean (3,177), meaning the distribution is marginally skewed. The 10-top stations count 97,951 trips, while the 10-bottom stations stand at 1,972. Therefore, station usage is not uniform, as some stations are clearly preferred to others. Peak hours seem to occur at around 9am and between 5 and 7pm, while late night clearly sees much fewer trips.

In order to facilitate analysis, the available datasets need to be rearranged in several ways. We used MySQL to remodel data for further investigation. This preparatory part is structured according to Schneider (2016). The first step entails deleting every customer trip. Namely, only trips coming from annual subscribers have been retained. This step has a dual purpose. First, it decreases the dataset inherent noise by eliminating unsystematic riders and, second, it removes a large share of observations for which no personal details are provided. Subscribers who refused to divulge their private information have not been deleted, so that one may observe their potential impact on the following analyses. Therefore, the new database contains 873,433 rows, the number of trips taken by subscribers only. In the following paragraphs, we will refer to this database as 'starting dataset', as it will be modeled in multiple ways to test different assumptions. In order to test H1, the starting dataset has been filtered to extract those trips that are unique along the following dimensions: (a) start station ID, (b) gender, (c) year of birth, (d) day, (e) rounded hour (e.g. 9:30 = 10, 9:29 = 9). Therefore, a unique trip is a combination of such variables that appears only once in the starting database. This step returns a list of unique trips.

Even one single re-identified record of a database can threaten information security, especially when a "face" is attached to it. We claim that some open datasets, such as Citi Bike's datasets, have a weakness in terms of single record re-identification: data collection is observable in real time. An example will clarify this statement.  On Saturday at 22:00, Julia, your 25 years old flat mate, comes out of her room and gets ready to go out with friends. She says that she is running late and, therefore, she wants to cycle to the meeting point. She starts looking for her Citi Bike card. Thus, you expect a new trip to pop out in Citi Bike's databases. It will record a woman, born 25 years ago, who took trips at around 22.00 at some close-by station. You wonder whether you can use this information to infer where the meeting point is.  Clearly, you will be able to determine her destination if, and only if, Julia is the only 25 years old female taking a trip from that station at that time. K-anonymization would prevent you from performing a similar trick and isolating a rider, but k-anonymization applies to identifiers and quasi-identifiers only. Starting station and time are not individual-specific information and, therefore, they are treated as payload data. Therefore, coherently with Schneider (2016) methodology and trip definition, the first hypothesis becomes:

> *H1-a: Given gender, year of birth, starting station, date and rounded time of departure, it is possible to recover the end station.*

Similarly, you may ask yourself where John, your 27 years old colleague, picks up his bike to cycle to work. A second version of the first hypothesis can thus be tested:

> *H1-b: Given gender, year of birth, end station, date and rounded time of arrival, it is possible to recover the starting station.*

By testing the aforementioned hypotheses, we may find that a combination of dataset analysis and real-life observation can threaten individuals' privacy. In fact, a certain trip would cease to belong to a general 25-years-old female, and take a real face instead. Besides the stand-alone value of this hypothesis, it also sets the foundations for the following, central, analysis. In fact, if most trips uniquely identify their riders, it may be possible to build journey maps by finding some similarities among trips that belong to the same person. To put it short, trips would embed enough uniqueness to distinguish between different individuals. We will look at this point closely in the next paragraph.

Gutwirth et al. (2016) show in their book how through a pseudonymized dataset of users it is possible to retrieve all users' maps for London's bike sharing service. Such dataset would pose enormous threats to personal privacy. For example, it could be combined with the technique introduced in the previous hypotheses, providing a complete overview over one person's movements. It would take only one observed trip in real life to match a face with a unique code. In addition, social networks and other digital sources could yield information to quickly identify any rider. Therefore, it is not surprising that Citi Bike did not make the same mistake. If the pre-analysis' hypotheses were positively verified, we would find that most trips are linked to riders in a one-to-one relationship. Namely, a trip, as defined in the methodology, has only one rider. This means that a certain trip is somehow a unique identifier for its rider; we only lack a variable to link all these unique identifiers/trips to the originating individual. Payload data may give us that variable: in fact, it might contain some hidden patterns that could turn out to be individual-specific. In particular, we tried to analyze trips occurrence patterns over the month, holding the reasonable assumption that trips belonging to the same individual have higher chances to occur on the same days.

The underlying rationale is that when a person cycles to work, he/she is likely to cycle her way back too. Although it is clearly an imperfect assumption, the purpose is to test whether it possesses some explanatory power. Therefore, we aim at recovering the excluded unique identifier by combining gender, year of birth and the index that will stem from the trips pattern analysis. Hence, hypothesis 2 in this setting becomes:

> *H2: For routine trips, given gender and year of birth, trips occurrence patterns yield a higher re-identification rate as compared to a random assignment.*

By comparing our results to a baseline formed by a combination of gender, year of birth and a random index, we will be able to understand whether some explanatory power exists in trips patterns. Other analysis' details will be provided in the methodology part. Finally, it is important to notice that such analysis will be restricted to routine trips only. In fact, being Citi Bike an occasional means of transport for most of its subscribers, we would accumulate too much distortion if every ride were to be considered.

## Results

According to our hypotheses, it is possible to breach privacy starting from anonymized data. The pre-analysis shows that anonymization fails when a certain identity can be uniquely linked to specific payload data, as a result of real-time observations. In other words, private information can be recovered by observing the trip in the moment it is taken. Besides, by confirming the uniqueness of most trips in terms of rider's ownership, it sets the foundation for the mapping. We will later discuss more into deep the mapping individuals' issue, another aspect of anonymization: impeding the matching of observations from the same individual. We demonstrate that anonymous observations hide patterns that link back to the physical person. As described in the methodology, H1-a and H1-b are tested by extracting unique combinations of the reference variables from the starting dataset (including only trips from subscribers). It must be noted that no inferential technique is used and, therefore, every uniquely identified trip has a 100% probability to belong to the observed individual. The software returns a list of the unique trips along the chosen variables, as previously showed.

| **Starting Dataset** | **#** | **Start Station** | **#** | **%** | **Stop Station** | **#** | **%** |
|---|---|---|---|---|---|---|---|
| Male | 662,929 | Male | 523,615 | 79% | Male | 528,660 | 80% |
| Female | 209,492 | Female | 191,531 | 91% | Female | 192,449 | 92% |
| Unspecified | 1,012 | Unspecified | 1,002 | 99% | Unspecified | 1,008 | 100% |
| Overall | 873,433 | Overall | 716,148 | 82% | Overall | 722,117 | 83% |

**Tables 2a/2b/2c – Unique Trips along Variables**

By analyzing such output along the variable gender and comparing it to the starting dataset, Tables 2a – 2c above emerge. The start station table concerns H1-a (for which start station is given), while stop station table refers to H1-b. For each category, we computed percentages as number of unique trips (#) over total trips in the starting dataset. Said percentages stand for re-identification rates, as they represent the proportion of trips whose ownership has no ambiguity.

For both start and stop stations, the female category exhibits extremely high percentages, standing at 91% and 92% respectively. Re-identification risk appears to be even higher for unspecified gender. Paradoxically, if one were to know a person's concerns about divulging gender, this person's privacy would be increasingly compromised. Proportionally, males seem to be harder to re-identify, standing at 79% for start stations and 80% for stop stations. Overall, more than 80% of the trips can be uniquely identified for both cases.

Similar to other k-anonymity-related studies (e.g. Wang et al., 2007; El Emam and Dankar, 2008), this analysis shows how minorities are particularly exposed to re-identification risk. It can be noted that the re-identification percentage rises as the number of trips in the starting dataset decreases. In other words, the smaller is the gender sub-group in the starting dataset, the higher will be the re-identification rate. Such a pattern is consistent for both start and stop station. This phenomenon is plain when considering the definition of k-anonymity: the impossibility to isolate less then k-1 records through a combination of quasi-identifiers (and payload data in this case). Given fixed distributions for all quasi-identifiers, a smaller group has smaller chances to see repeated quasi-identifiers combinations. Namely, there is a higher probability for every trip to be the only one with specific features. For this reason, female trips (24% of the starting dataset) show higher uniqueness than male trips (76%).

The same principle applies to by-birth analyses. The chart below illustrates male and female re-identification rates by year of birth for the start station case. It can be noted that older and younger riders are more easily identified, some yielding 100% re-identification. This is due to the fact that fewer of them feature as Citi Bike subscribers. Consistently, reaching a low of 73% in 1984, riders born in the 80s show the smallest re-identification rates since they represent most of the initial database's rides. It can be clearly seen that, although showing similar trends, females constantly exhibit higher rates. As mentioned above, it is a consequence of the lower number of female trips in the datasets. The following graphs (Figures 1 and 2) provide information about start and end station re-identification rates. They show similar trends.

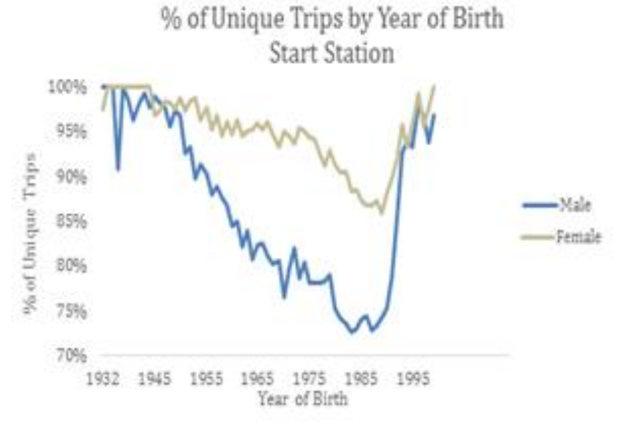

**Figure 1 – Start Station Re-identification Rates**

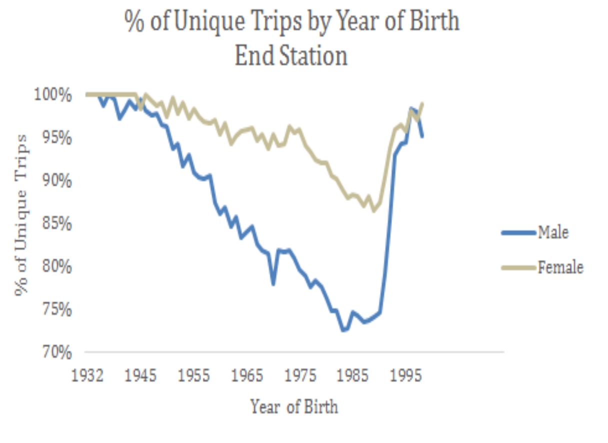

**Figure 2 – End Station Re-identification Rates**

We previously mentioned Julia and we wondered where she was meeting her friends. We now know that there is a 91% probability that we can recover the surroundings of the meeting point. In other words, there is a 91% probability that Julia is the only 25 years old female taking a ride from the closest station at around 22.00 today. Same is true for our colleague John, although the rate stands at 80%. One may believe that these pieces of information hold little significance. However, it must be noted that trips can reveal far more than simple movements from X to Y. For example, it is straightforward that by checking John's starting point, we probably discover his residential area. Similarly, by looking up her destination, we may able to find out which friend Julia is dating. This means that both start and end station are often correlated with other (sensitive) personal information. In addition, it is possible to increase the accuracy of said findings by performing the same checks multiple times. For example, we may control John's start station over a one-week period, in order to consolidate our results.

As mentioned above, no statistical inference is involved in this analysis. Therefore, there is not any p-value-like measure to confirm the significance of our outcomes. However, we can safely state that a major part of the dataset is uniquely identifiable and, therefore, privacy can be considered substantially breached at the single record level. Given the high re-identification rates, hypotheses H1-a and H1-b are positively verified. In addition, we now have enough proof of trips' uniqueness to proceed with our main analysis. If single-record re-identification rates were low, it would be impossible to extract individual riders from trips' similarities. In fact, there would not be enough difference in riders' behavior to use trips occurrence as a distinguishing factor. It must be noted that the trip definition used for H2 can only increase single-record re-identification, since an additional variable is added. In fact, it would be as high as 99%.

### *Mapping Individuals*

This part of the analysis aims at broadening the scope of re-identification as compared to the former hypothesis. While the previous paragraph focuses on the de-anonymization of single records, the ideal objective is now to re-build journey maps (Siddle, 2016). Therefore, even if no face will be attached to the rider, a larger portion of such rider's movements can be identified. Anonymization is breached as we manage to re-create a unique identifier that matches single trips that belong to the same individual. In short, we recreate a pseudonymized dataset. This analysis is based on a set of matrices that have one column per day and one row per routine trip. Therefore, they contain only binary variables that display 1 when a certain trip occurred on a certain day, or 0 otherwise. For each birth-gender combination, a different matrix is created and submitted to a hierarchical clustering algorithm. Supposedly, every cluster corresponds to a rider. Thus, our unique identifier is formed by gender, year of birth, and cluster membership.

As outlined in the hypothesis, only routine trips have been taken into account. The underlying assumption is that routine trips correspond to routine riders. Namely, we assume that there is a group of subscribers that use Citi Bike as their main means of transport and, in turn, they are responsible for the more frequently observed trips. In the next section, we take one step forward and develop an analysis to infer the average number of trips per rider in the dataset. We intend to find out how many rides, on average, belong to the same individual. It must be noted that we keep such figure constant for every birth-gender combination. Given the fact that occasional trips have been removed, we expect this number to stand between 3 and 5.

## Internal Validation

It is common for machine learning algorithms to lack external validation tools. Namely, there is no independent dataset that provides references for testing the accuracy of the algorithm. Clustering techniques are no exception and, therefore, many internal validation systems have been lately engineered. Internal validation resorts to the inherent characteristics of the dataset to determine whether consistent patterns exist or clusters are just the result of random matching. However, such techniques are often used to compare different models rather than develop benchmarks to either retain or discard a single solution.

In this case, we use internal validation to decide upon the number of clusters. Such choice is done indirectly by choosing the average number of trips per rider. In turn, this figure determines how many individuals (and clusters) are represented in the dataset. In order to simplify the analysis, we assumed an equal number of trips per rider across all birth-gender combinations.

In order to choose the number of clusters, it is common practice to look at the between-cluster to within-cluster ratio. Although such index is not optimized for binary analyses, bias should be under control as we are using a distance matrix.  By plotting the distance ratio against the number of clusters, a downward curve is returned. It is known as elbow curve, since the number of clusters is chosen by looking at the inflection point. In fact, the ratio monotonically decreases as the number of clusters increases. However, from a certain point onwards, the fall is far less evident. The following chart in Figure 3 plots the average ratio by average number of trips across all birth-gender combinations. It is upward sloping because the average number of trips per rider is inversely proportional to the number of clusters.

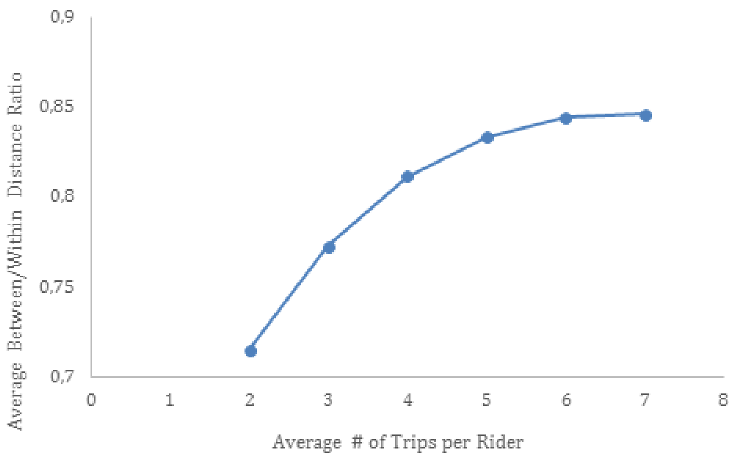

**Figure 3 – Average Ratio by Average Number of Trips across all Birth-Gender Combinations**

By interpreting the chart, we can conclude that 4 is the optimal average number of trips per rider. This is in line with our expectations. Said result is used to cut the dendrogram so that each cluster contains approximately 4 observations. In the dendrogram below (Figure 4), different colors represent different clusters and, thus, different riders.

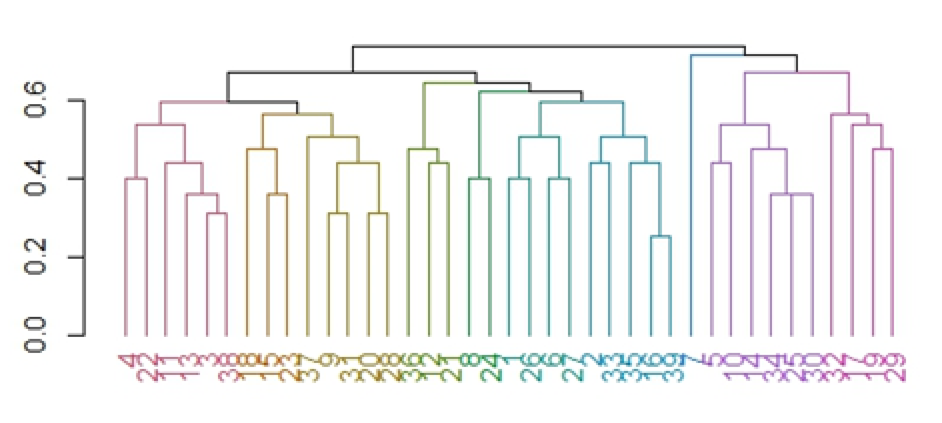

**Figure 4 – Dendrogram Colored according to Different Clusters**

It is important to notice that not all clusters contain exactly 4 observations. In fact, every run of the clustering algorithm must adapt to a specific configuration and, therefore, small differences in cluster size can be seen. However, the "complete" hierarchical clustering ensures approximately even sizes across clusters.

## *External Validation*

Clustering solutions are known to be hard to validate. In the best-case scenario, an ex-ante dataset provides external references. As mentioned in the methodology, we were not given such testing sample, but we built one on the basis of specular trips. These are the trips with the same stations and demographics, but different directions. Among the routine trips, we assumed that mirroring trips belong to the same person. Therefore, the accuracy rate was tested by counting how many specular couples were correctly matched by the clustering analysis. Although more sophisticated external measures exist, the nature of our testing sample calls for simplicity. In fact, it does not replace a fully labeled external dataset.

Similar to the previous hypothesis, it is possible to split results along the given demographics. Table 3 summarizes our findings for the accuracy rates.

| **Random Assignment** | | **Cluster Analysis** | |
|---|---|---|---|
| Male | 13% | Male | 35% |
| Female | 43% | Female | 93% |
| Overall | 18% | Overall | 45% |

**Table 3 – Random Assignment Comparison**

Precision rate stands at 93%, thus we can trust accuracy rates to be reliable.

As described earlier, clustering membership was used as third element to build our unique identifier, along with gender and year of birth. In the control case, membership was replaced by a random number between 1 and the number of trips in the birth-gender subgroup divided by 4 (by doing so, the range of possible clusters is as wide as the range of random numbers). It can be noted that the clustering results are superior to the random assignments (RA) under all categories. For females, clustering scores 93%, while RA stands at 43%. For the male group, results are less striking, clustering and RA stand at 35% and 13% respectively. Overall, the gap between the algorithm and RA equals nearly 30%.

Analogous to the previous hypothesis, small groups seem to facilitate re-identification. It can be seen that, in proportion, female trips in the routine dataset are even fewer than in the starting dataset. They account for only 16% of the most frequent trips. Consistently, in the testing sample, only one female trip was mismatched. On the other hand, since male trips form the bulk of the routine rides, the matching process is less accurate.

Conversely, the by-birth analysis does not show any interpretable pattern (see Figure 5). It is probably due to the fact that by-birth subgroups contain few observations and, therefore, are not statistically representative.

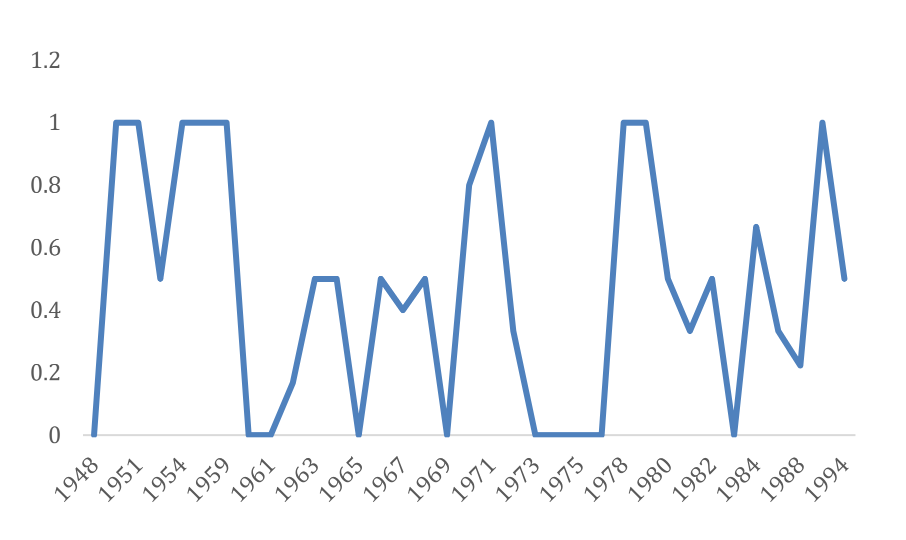

**Figure 5 – Re-Identification Rate by Year of Birth**

Three assumptions are crucial for our study: first, that frequent trips belong to routine riders; second, that the average number of trips per rider can uniformly extend to all birth-gender combinations; third, we highlight our leading assumption, i.e. that trips belonging to the same individual have higher chances to occur on the same day.

It is evident that although these assumptions are reasonable and likely to happen in most cases, they cannot be perfect. This is the reason why re-identification rate stands at about 50%, and journey maps cannot be considered as fully rebuilt. However, our goal was to demonstrate that hidden patterns in payload data can reveal person-specific information. By comparing our process to a random assignment, we were able to show that apparently anonymous information has predictive power with regard to individuals' habits. Therefore, to a certain extent, payload data can be exploited to transform anonymization into pseudonymization. In addition, minorities should be considerably concerned with this sort of analysis, as their exposure is far greater than average.

## Discussion and Conclusion

The analysis of Citi Bike's datasets uncovered some threats that can stem from even anonymized data. As expected, k-anonymization techniques, based on quasi-identifiers alone, have proved to be insufficient strategies to fully protect personal information.

We have seen that, given a small number of details, most trips are uniquely identifiable. Thus, we were able to recover one's starting (end) station from just few demographics: the end (starting) station and some date/time variables. Re-identification rates were particularly high for female riders, representing a minority in our dataset. Some of the recoverable pieces of information can be correlated with sensitive data, such as one's home or work address.

Our study contributes to a mainstream literature on the topic of record identification in large datasets. Much of the debate took place on operations research journals since the early eighties, but also *MISQ* and the *Communications of the ACM*, which are in the information systems research domain, hosted the discussion. In the meantime, more and more organizations are considering releasing public versions of their data gold mines: a recent example (summer of 2017) comes from the initiative undertaken by Uber, named "Uber Movement". Uber publicly declares that the initiative will broaden the knowledge on research issues like (for example): impact of metro shutdown on traffic congestions, assessment of road infrastructure based on usage, understanding of holiday traffic trends. All honorable and valuable research issues: answering to such challenges could significantly improve citizens' lives, given the more and more congested urbanization trends experienced in all continents. Nonetheless, research alerts, coming from our findings and similar previous investigations, have not been strong enough to raise a debate on openly available datasets: the proof is that data is more available than before. Specifically, our study adds the following points to the debate:

1. There is a profound difference between non-human related datasets and payload data that leads to human behaviors. Most of the initial debate on this topic has been focused on the algorithm and the mathematical demonstration of the re-identification possibilities, given a set of data, regardless the nature of this data. We argue that this is acknowledged enough, but in our case the Citi Bike data does not talk about weather conditions in Manhattan, rather it talks about people's behavior in transportation needs.
2. Moreover, data like ours is a free downloadable resource for anyone that is surfing the right web page: no prior profiling or registration is required. One more time, anonymized dataset is not enough to fully protect privacy, but organizations are still sharing data online.
3. This brings to the third point: when payload data has to do with people's preferences the tradeoff between more knowledge in exchange of less privacy is even more important. This research should bring the debate to a next upper level. What does the academic community suggest to face this issue? It is hot and burning.
4. Our study highlights that minorities are particularly affected: their re-identification possibilities are higher; this makes the debate even more urgent.

Having said this, a number of other important implications can be considered, if one may want to consider the following potential circumstances:

- *What if payload data can be rearranged to reveal individual-specific patterns*? This is the most evident threat since it allows analysts to match observations that belong to the same user. Much information can emerge by grouping multiple records from the same individual.
- *What if the event that originates the dataset's record is observed in real time*? As a consequence, an identity can be linked to a single row and used to recover unobservable data. Apparently innocuous details can be highly correlated to sensitive information, with real time implications.
- *What if an increasing number of datasets will be released for free public access?* This is an even more threatening circumstance, already in place. A simple internet search will reveal the magnitude of the problem. "Connecting the dots" of people's behaviors, by linking different datasets, will make it even simpler to control what we do in real life. This brings up to the surface, one more time, the price paid to the information society: the loss of people's privacy.

Thinking like a "criminal mind" helps discovering the disruptive effects that dataset disclosure could have on society. This is a good exercise that could worry anyone.

All players in the (open) data field must work together to create new privacy models: systems that take into account these issues and ensure a greater degree of protection. In the case of our study, it is clear that society needs to strike a balance: taking the fruits of broadening knowledge by encouraging the open community to carry on creative investigations on datasets once made public, and protecting individuals from being chased out by dataset hackers for other purposes, either commercial nature or, worse, criminal intent. Most of these tradeoffs can be assessed only with ex-post evaluations: once the dataset is public, an appropriate time window should be designated to look out for the results. At the same time, unlawful use of the dataset should be strictly monitored, by incentivizing the dataset owner to run both the initiatives.

In the next paragraphs, we will introduce the implications for both private organizations and academic institutions.

## *Practical Implications*

As a result of our analysis, it is possible to draw some qualitative conclusions that private organizations and companies should consider when endorsing open data. First and foremost, one must know his/her data, its openness to predictive analytics and the less obvious patterns. For instance, this paper brings up two additional angles of analysis: the ability to control data-collection in real life and the use of big data to create unique identifiers. Every firm or institution releasing data should consider these elements as well as others that may be peculiar to their dataset. Deep knowledge of the data at hand is a necessary condition for better privacy shields. Once data has been thoroughly explored, it is possible to decide to what extent it must be obscured. It is important to notice that, generally speaking, discrete/continuous variables are more dangerous that dichotomous/categorical ones, as they split users in finer sub-groups. However, dichotomous factors can be highly threatening when they isolate certain minorities. For example, in our analysis, female sub-groups are far more exposed to re-identification rates. Therefore, open data releases should favor categorical variables, unless they feature highly non-uniform distribution. In that case, some minorities may be endangered as well.

As we saw earlier, organizations can resort to two main methodologies to strengthen k-anonymization: generalization and suppression. When payload data shows dangerous patterns, these methods act on quasi-identifiers to reduce re-identification risk. If the privacy level is still unacceptable, even payload data may undergo some sort of obscuring process. Clearly, the equilibrium between generalization and suppression, and their absolute levels, must account for the trade-off between privacy and data completeness. As anonymization becomes stricter, the depth and breadth of available information falls. Case by case, managers must figure out how to maximize the usability of data without endangering society. Frameworks to assess risk are left to further research.

If all the above attempts fail, and the dataset disclosure still offers possibilities of re-identification of individuals, we suggest that data should not be openly disclosed, instead restrictions could apply. For example, a preferred practice could require prior registration for all interested in dataset download. With appropriate profiling, the data owner could track accurately who is doing what, and could also ask for something in exchange, like sharing the fruits of the data usage.

## *Research Implications*

Given the fast-technological advancements and the rising pressure from the public opinion, important challenges rest on privacy scholars' shoulders. As open data elbows to make its way to the maturity stage, it is their responsibility to create models that both protect privacy and grant releasers enough freedom. As aforementioned, the trade-off between privacy and usability is likely to be the crucial point to determine open data's future. Although it features several weak points, k-anonymization remains a well-thought model. Having proved the effectiveness of this method, we believe that k-anonymity may live long if its supporters will update it to account for the incoming threats that big data poses. However, it goes without saying that new efforts must be taken to develop new models. We need solutions of both technical (e.g. cryptography, pseudonymization) and structural (e.g. k-anonymization) nature. Ideally, releasing institutions should be able to pick the best-suited model within a range of available options. As aforementioned, awareness and true understanding of data will draw the line between failure and success.

There are two technology trends that deserve to be examined in light of our analysis: real-time data and predictive analytics. Real-time data refers to "information that is delivered immediately after collection". Most papers on this topic focus on the potential applications of such technology, especially for navigation and tracking services. For instance, Citi Bike may have a mobile app that allow users knowing whether there is a bike in a certain station. However, when single records can be easily de-anonymized, real-time access to data may severely threaten personal security. Therefore, it is necessary to dedicate a fair share of attention to understand privacy implications of real-time data and how to weaken negative externalities. The second hot topic concerns predictive analytics and, in general, the techniques associated to big data.

Data science has been gaining ground, mainly because of the new opportunities it offers. Analysts can now build algorithms that automatically interpret unstructured data, such as textbooks, e-mails, audio files. In addition, the new computational mechanisms are structured so that they improve as their usage increases. For example, neural networks, computer systems that replicate the functioning of a human brain, learn "how to think" as they grow older (i.e. new input data is absorbed). As far as privacy is a concern, scholars have been pointing out that these techniques can produce new knowledge that exposes sensitive information. For example, Target foresaw a girl's pregnancy. However, another scenario arises. Big data's bullets can be shot to neutralize anonymization strategies, rather than creating new knowledge from given information. In our analysis, we used a clustering algorithm to recover hidden identifiers. Therefore, new anonymity models must take into consideration the revealing power of big data, in terms of both knowledge generation and defense disruption.

## *Limitations and Future Research*

The limitations of our analysis mainly concern the "Mapping Individual" part. In fact, this is the statistic-intensive section, which requires both imperfect assumptions and sophisticated technical solutions. The following paragraphs outline the limits of our research and provide some ideas that may be implemented in future papers to consolidate our findings.

Due to the structure of our analysis and the available dataset, we were forced to proceed by taking reasonable, but unavoidably imperfect, assumptions. For example, we assumed that routine rides are taken by frequent riders and that the average number of trips per rider is constant across birth-gender combinations. Although they are all sensible expectations, some degree of error would come with no surprise. Even the basic assumption, trips belonging to the same individual have higher chances to occur on the same days, is an expression of probability rather than supposed certainty. The effects are clearly visible on the final re-identification rate: 45%. Although it is more than enough to show the revealing power of payload data, it does not suffice to rebuild accurate and complete riders' maps.

Besides assumptions, limitations derive from some technical aspects of our research. The binary cluster analysis is a relatively new technique, whose methods are being continuously improved by scholars. The ongoing literature expansion makes it hard for software packages to keep it up. Therefore, some of the most recent techniques are not readily available on many common statistical platforms. For example, few binary internal validation measures are accessible through R, let alone Python. In turn, we had to slightly alter our analysis to align with the scope of the available tools (e.g. within-between distance was used instead of entropy to find optimal number of clusters).

Another limitation to our research is the absence of a real testing sample. In order to build ours, we had to reshape the dataset and take further assumptions. We assumed that two mirroring trips belong to the same rider. Again, although it is a reasonable one, it cannot be 100% accurate. Most external validation measures are based on fully labeled datasets, in which every observation is attached to the "right" cluster. However, we had to deal with couples of trips, rather than actual clusters. In turn, the re-identification rate might be biased. On a comparative basis, it is sufficient to prove the superiority of trips' patterns over random assignments. However, the real re-identification rate, tested against a fully labeled dataset, may differ from the one we found.

In order to strengthen the validity of our findings, our analysis may be repeated in several ways, varying input data and methodology. For example, one may try a similar analysis with a dataset that was not so strongly k-anonymized in the first place. Optimal datasets may be the ones that do not directly refer to individuals' activities or characteristics, for example a dataset that lists different items. One may use multiple variables to recover the selling point, or other geolocation-related information. Although privacy implication would be milder, it would be a good opportunity to confirm the technical validity of our findings.

Another possible improvement would be basing the analysis on a dataset that calls for standard cluster analyses, such as k-means. Not only is the effectiveness of these methods more consolidated in the relevant literature, but also the related evaluation systems are up to date on the most common statistical packages. Therefore, one should use continuous variables instead of binary, and discrete instead of categorical variables. Further, the accuracy may increase if the number of clustered variables decreases. Clearly, these contingencies depend on the initial database, rather than analysis' design decisions. In our case, there was no variable that, alone, could group trips by rider, neither there was a continuous or discrete variable that offered much explanatory variable; a large, complex, binary matrix was the only tool at our disposal.

Moreover, the re-identification method could be further improved by using other techniques, such as neural networks. It would be interesting to conduct a study on which method may achieve the best result given the contingent characteristics of each single dataset. We are also interested in developing new methods for anonymization that can protect personal privacy while achieving good data mining utility.

Finally, we suggest scholars employ datasets that come with complete testing samples. Besides proving payload data's revealing power, they might succeed in finding out the real re-identification rate. It must be noted that such sample does not need to be the full counterpart of the original dataset. As long as it shows complete clusters, it could also contain only a fraction of the initial observations. However, the sample should be statistically representative of the whole population, and the original proportions should remain unchanged.

## References

Aggarwal, C. C. 2005. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international Conference on Very Large Data Bases* (pp. 901-909). VLDB Endowment.

Angiuli, O., Blitzstein, J., & Waldo, J. 2015. How to de-identify your data. *Communications of the ACM*, (*58*:12), pp. 48-55.

Angiuli, O., & Waldo, J. 2016. Statistical Tradeoffs between Generalization and Suppression in the De-identification of Large-Scale Data Sets. In *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual* (Vol. 2, pp. 589-593). IEEE.

Barocas, S., & Nissenbaum, H. 2014. Big data's end run around procedural privacy protections. *Communications of the ACM*, (*57*:11), pp. 31-33.

Bayardo, R. J., & Agrawal, R. 2005. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering 2005* (pp. 217-228). IEEE.

Bélanger, F., & Crossler, R. E. 2011. Privacy in the digital age: a review of information privacy research in information systems. *MIS Quarterly*, (*35*:4), pp. 1017-1042.

Bennett, C. J. 1992. *Regulating privacy: Data protection and public policy in Europe and the United States*. Ithaca, NY: Cornell University Press.

Butler, A. R. 1982. A Note on the Power of Personal Identifying Information. *The Journal of the Operational Research Society, (33*:1), pp. 73-76.

Chen, H., Chiang, R. H., & Storey, V. C. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, (*36*:4), pp. 1165-1188.

Claerhout, B., & DeMoor, G. 2015. Privacy protection for clinical and genomic data. The use of privacy-enhancing techniques in medicine. *International Journal of Medical Informatics, 74*, pp. 257-265.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., ... & Chuang, I. 2014. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, (57:9), pp. 56-63.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Nature Scientific Reports*, (*3:*1376), pp. 1-5.

Du Bois Jr, N. D. A. 1969. A solution to the problem of linking multivariate documents. *Journal of the American Statistical Association*, (*64*:325), pp. 163-174.

El Emam, K., & Dankar, F. K. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, (*15*:5), pp. 627-637.

Epstein, D. 2013. The making of institutions of information governance: the case of the Internet Governance Forum. *Journal of Information Technology*, (*28*:2), pp. 137-149.

Epstein, D., Roth, M. C., & Baumer, E. P. 2014. It's the Definition, Stupid! Framing of Online Privacy in the Internet Governance Forum Debates. *Journal of Information Policy*, *4*, pp. 144-172.

Gritzalis, S. 2004. Enhancing web privacy and anonymity in the digital era. *Information Management & Computer Security*, (*12*:3), pp. 255-287.

Gutwirth, S., Leenes, R., De Hert, P. *Data Protection on the Move: Current Developments in ICT and Privacy/Data Protection*. Springer: Berlin, Germany.

Hill, C. 2012. How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. *Forbes*. Retreived on 2017/03/10 from: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#7143eee46668.

Lee, D. J., Ahn, J. H., & Bang, Y. 2011. Managing consumer privacy concerns in personalization: a strategic analysis of privacy protection. *MIS Quarterly*, 423-444.

Mai, J. E. 2016. Big data privacy: The datafication of personal information. *The Information Society*, (*32*:3), pp. 192-199.

Malhotra, N. K., Kim, S. S., & Agarwal, J. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, (*15*:4), pp. 336-355.

Mountford, W. K., Lipsitz, S. R., Fitzmaurice, G. M., Carter, R. E., Soule, J. B., Colwell, J. A., & Lackland, D. T. 2007. Estimating the variance of estimated trends in proportions when there is no unique subject identifier. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, (*170*:1), pp. 185-193.

Nergiz, M. E., & Clifton, C. 2007. Thoughts on k-anonymization. *Data & Knowledge Engineering*, (*63*:3), pp. 622-645.

Schneider, T. W. 2016. *A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System*. Retrieved on 2017/03/15 from: http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system

Smith, H. J., Dinev, T., & Xu, H. 2011. Information privacy research: an interdisciplinary review. *MIS Quarterly*, (*35*:4), pp. 989-1016.

Sutanto, J., Palme, E., Tan, C. H., & Phang, C. W. 2013. Addressing the Personalization-Privacy Paradox: An Empirical Assessment from a Field Experiment on Smartphone Users. *MIS Quarterly*, (*37*:4), pp. 1141-1164.

Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, (*10*:5), pp. 571-588.

Wang, K., Fung, B. C., & Philip, S. Y. 2007. Handicapping attacker's confidence: an alternative to k-anonymization. *Knowledge and Information Systems*, (*11*:3), pp. 345-368.

Westin, A. F. 1968. Privacy and freedom. *Washington and Lee Law Review*, (*25*:1), pp. 166-170.