Disinformation Defense

AI Inference

Karel Baloun

Keng-Shao (Ken) Chang

Matt Holmes

Master of Information and Cybersecurity (MICS)

University of California - Berkeley

Fall 2019 | Capstone | School of Information

cybersecurity@berkeley

Table of Contents

Disinformation Defense

AI Inference

1 Introduction

Internet has no physical boundary on how personal data can be collected across the international community. The personal data collected and generated through the digital transaction of personal life can be considered as an active and passive type of data collection channels. Active user data collection requires a user to spend time entering information and requires that a user feel comfortable with providing information over the Internet. Passive user data can be obtained and unintentionally left behind by the users of the internet and digital devices without explicit consent on how it is being collected, processed and stored.

With the advance of Machine Learning and Artifactual Intelligence under high-performance computing platform using the vast amount of personal data, the adversaries sponsored by nation-states can perform data linkage and aggregation to analyze and infer the targeted person's personal and professional activities in the critical infrastructure sector which might pose a threat to national security. Furthermore, based on the leaked personal data, the adversary can spread disinformation or misinformation on the targeted person held a critical position at the private or public sector which might eventually lead to the disruption of business and government functions. The Russian government interfered in the 2016 U.S. presidential election with the goal of harming the campaign of Hillary Clinton, boosting the candidacy of Donald Trump, and increasing political and social discord in the United States.

	Security: the traditional information security focus on confidentiality, integrity and availability.

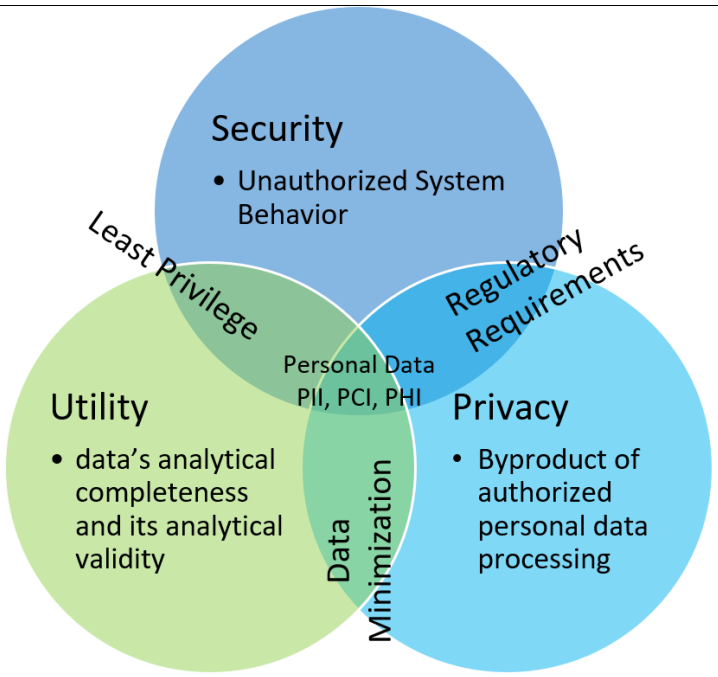


Figure: Security vs. Privacy vs. Utility[1][2]

Utility: for the data scientist community, full unvetted data access is considered the necessary evil for analytical completeness.

Privacy: the fundamental rights for privacy of individual in the digital space.

For our research, we are focus on exporing the balance of utility and privacy when access personal data attributes available on the public space and legalled collected and shared.

[1] USENIX Enigma 2019 - Privacy Engineering: Not Just for Privacy Engineers

[2] NIST 8062

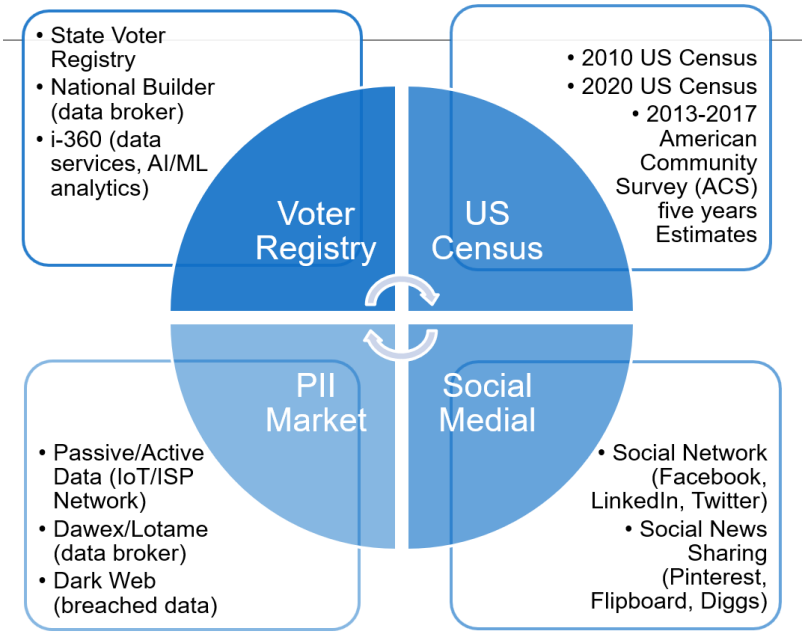


Figure: Pillars of Personal Data Sources

US Census: The Census Bureau's primary mission is conducting the U.S. Census every ten years.

Voter Registry[3]: Voter registration in the United States takes place at the county level, and is a prerequisite to voting at federal, state and local elections.

Social Media[4]: An interactive computer-mediated technologies that facilitate the creation and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks

PII Market: Legal or illegal acquired personal data available for sell on the market.

[3] https://en.wikipedia.org/wiki/Voter_registration#United_States
[4] https://en.wikipedia.org/wiki/Social_media

2 Threat Modeling

Collective data breaches leaking personally identifiable information (PII), Payment Card Information (PCI), and Protected Health Information (PHI) in the United States and other developed countries like Taiwan and Singapore. The personal data can be sensitive or not sensitive collected via active and non-active channels. Non-sensitive PII is information that can be transmitted in an unencrypted form without resulting in harm to the individual under normal circumstances. The normal circumstances imply data being collected and processed in isolation without linking other sources of personal data. Non-sensitive PII can be easily gathered from public records, phone books, corporate directories, and websites. Sensitive PII, PCI and PHI data are information which, when disclosed, could result in harm to the individual whose privacy has been breached. Sensitive personal data includes biometric information, medical information, personally identifiable financial information (PIFI[5]) and unique identifiers such as passport used for international travel or Social Security numbers.

The private sector has been under repeated offending attacks in the result of massive data breach events. In October 2017, Yahoo[6] disclosed 3 billion accounts breached in 2013 incident; up from the original estimate of 1 billion accounts. Specific details of material taken include names, email addresses, telephone numbers, encrypted or unencrypted security questions, and answers, dates of birth, and hashed passwords. In November 2017, Uber [7]shocked consumers when it admitted that it failed to notify victims for over a year after paying $100,000 to hackers who had

[5] https://www.techopedia.com/definition/14222/personally-identifiable-financial-information-pifi
[6] https://en.wikipedia.org/wiki/Yahoo!_data_breaches
[7] https://www.theguardian.com/technology/2017/nov/21/uber-data-hack-cyber-attack

stolen data on 57 million users and drivers. In 2017, one of the largest CRAs, Equifax [8]Inc. ("Equifax") announced that it had suffered a data breach that involved the PII of over 145 million Americans in the result of 230 million revenue loss quarter over quarter. In November 2018, Marriott first revealed it had suffered a massive data breach affecting the records of up to 500 million customers. Information accessed included payment information, names, mailing addresses, phone numbers, email addresses, and passport numbers. In February 2015, Anthem[9], Inc. disclosed that criminal hackers had broken into its servers and potentially stolen over 78.8.5 million records that contain personally identifiable information. The compromised information contained names, birthdays, medical IDs, social security numbers, street addresses, email addresses and employment information, including income data.

The public sector can't be exempted as the victim of the data breach. In June 2015, the United States Office of Personnel Management (OPM) announced that it had been the target of a data breach targeting the records of as many as twenty-one million people. This includes records of people who had undergone background checks, but who were not necessarily current or former government employees. Same malware, Sakula was used to attack other US companies which led to the conclusion of Chinese adversaries behind the attack.

Social Media has been targeted as mega data hubs of personal active and passive data. In 2018, the Facebook [10] data scandal was revealed that Cambridge Analytica had harvested the personal

[8] https://investor.equifax.com/~/media/Files/E/Equifax-IR/reports-and-presentations/events-and-presentation/investor-relations-presentation-december-4-2018.pdf

[9] http://money.cnn.com/2015/02/04/technology/anthem-insurance-hack-data-security/

[10] https://www.theguardian.com/technology/2018/apr/10/facebook-notify-users-data-harvested-cambridge-analytica

data of millions of people's Facebook profiles without their consent and used it for political advertising purposes.

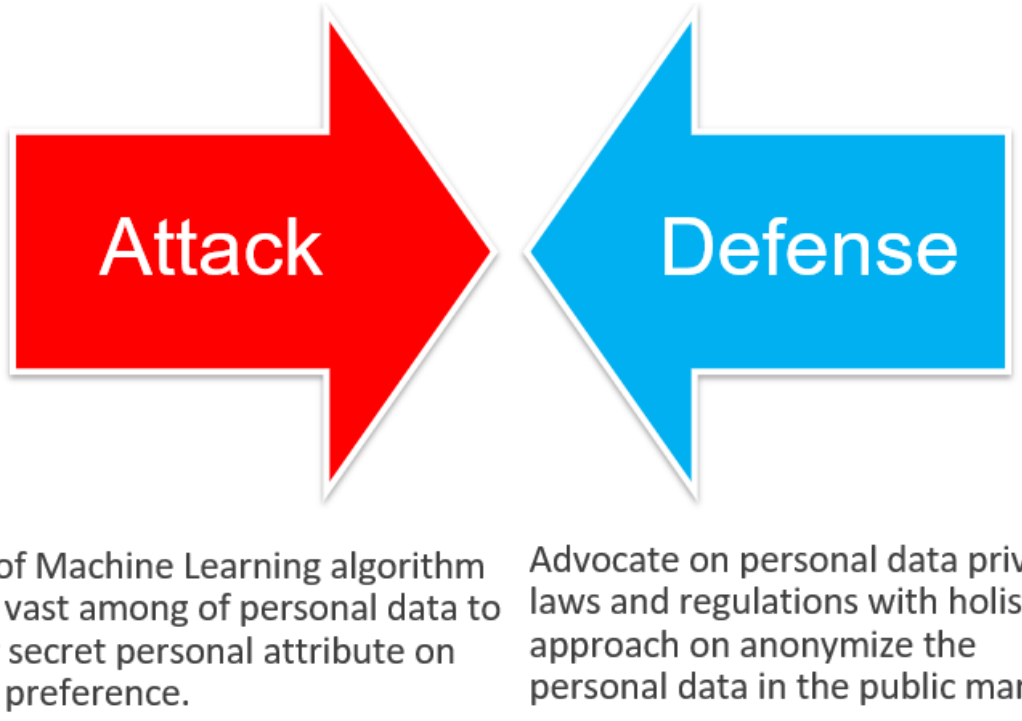


Figure: Voter Preferences Disclosure Attack and Defense

Assets	Threats	Description
Voter Database	Confidentiality, Integrity	Public static database contains personal attributes such as full legal r birth date, gender, residential address and registered party affiliation data broker provides services for distribution with fee and might sub data poisoning attack.

US Census Database	Confidentiality	Public statistical database contains summary statistics of person attri such as ethnicity, income, education, marriage, and number of childr combination of synthetic data generation, data aggregation and linka very likely to infer sensitive personal attributes such as vote preferen
Social Media	Confidentiality	Data leakage via public post with personal attributes, personal assoc with degrees of separation, and subject to targeted marketing adverti
Personal Data Market	Confidentiality, Integrity	Collective Data breach events creates black market of personal data Web. Consumer service provider collects and resell the personal attr for profit or in exchange of free service. Personal data exchange ma promotes monetized model for individual. The personal data broke subject to data poisoning attack.
Personal Privacy	Integrity	Personal service right might subject to perpetuate discrimination bec machine learning result are trained on biased data. An individual m receive the service they needed.

3 Design of Experiment

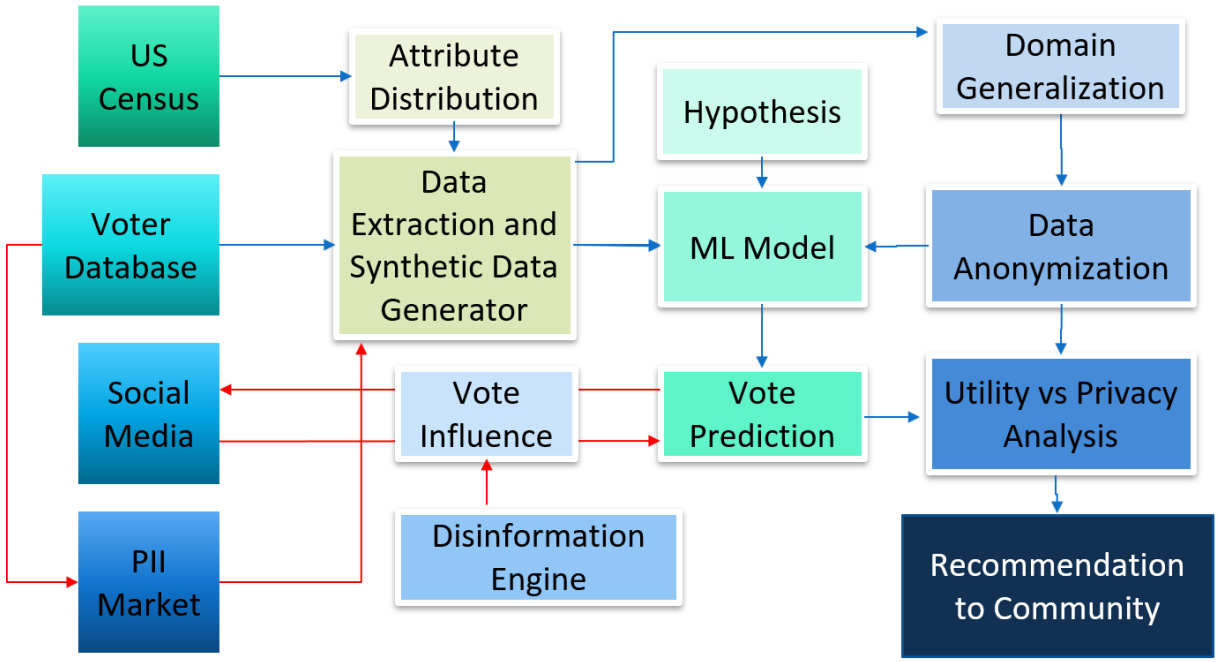


Figure: Design of AI Inference

3.1 Attribute Distribution

2018 estimates available from 2010 census data, for income, education, family size, and ethnicity. These can be filtered by geographic area, gender, and age. Data is difficult to extract from web interface. API exists but the interface for getting an apikey is broken. Margin of error can rapidly overwhelm any information content, as filters lower the number of class members. Due to margin error, not possible to map to voter database, beyond large sample proportional distributions. Real world individual voter profiles exist, with all of these attributes and many more. These exist as purchasable data, as well as advertising API targets.

3.2 Data Extraction and Generation (Bootstrapping)

Input: Voter Data File, vd_file_name, Voter File Table, VFT: vf_attributes = ["state_file_id", "dob", "sex", "party", "county__registered_address", "zip__registered_address", "state__registered_address", "federal_district"], Synthetic Attribute Distribution, p_[ethnicity, income, education, marriage, children]

Output: Data Output Table: dt_attributes = ['state', 'zipcode', 'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children', 'party']

Algorithm:

```
vf_data = load_VF_data(vd_file_name, vf_attributes)

sd_ethnicity = generateEthnicity(len(vf_data), p_ethnicity)
sd_income = generateIncome(len(vf_data), p_income)
sd_education = generateEducation(len(vf_data), p_education)
sd_marriage = generateMarriage(len(vf_data), p_marriage)
sd_children = generateChildren(len(vf_data), p_children)

 for i, row_ in enumerate(vf_output):
                for index in range(0, 10):
                                dt_row = generate_columns(index, row_, dt_attributes)
                append_row(dt_row)

output(dt_output_file)
```

Figure: Data Bootstrapping Algorithm

3.3 Machine Learning (ML) Model

3.3.1 Hypothesis

Use of Machine Learning Algorithm to infer voter preferences (secret attribute) from personal attributes includes {'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children'} extracted from voter registry and linked with PII data from US Census, PII data market.

3.3.2 Supervised Learning with Classification

Please reference to section 4 for more detail on Attack with Machine Learning Model.

3.3.3 Vote Preference Inference (Prediction)

Please reference to section 4 for more detail on result of vote preference inference on the accuracy of prediction with machine learning model. We will compare the prediction score using the raw data and anonymized data set.

3.4 Domain Generalization

qID = {'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children'};

Ethnicity Generalization Tree:	Sex Generalization Tree:
Income Generalization Tree:	Education Generalization Tree:

Marriage Generalization Tree:	Children Generalization Tree:
Age Generalization Tree:	

3.5 Data Anonymization

Please reference to section 5 for more detail on Defense with Principles of Privacy Engineering.

3.6 Vote Influence

3.7 Disinformation Engine

4 Attack: Machine Learning

5 Defense: Privacy Engineering

In our design for applying privacy engineering, we choose k-anonymity for the simplicity and easy to implement with known effectness on defense against identity disclosure attack. For the voter preference as sensitive attribute, it is not considered as a publishable attribute. Voter registry might contain the voter party but it does not necessary reflect the true vote. To achive k-anonymity, we choose datafly algorithm with domain generalization for categoricial attributes; developed by Sweeney in 2017 is global, bottom-up, and greedy generalization algorithm. We have referenced and enhanced datafly implementation [11] by Alessio Vierti.

Core Datafly Algorithm

Input: Private Table PT; quasi-identifier QI = (A_1, …, A_n), *k*-anonymity constraint *k*; domain generalization hierarchies DGH_{A_i}, where i=1,…,n with accompanying functions f_{A_i}, and *loss*, which is a limit on the percentage of tuples that can be suppressed. PT[*id*] is the set of unique identifiers (key) for each tuple.

Output: MGT a generalization of PT[QI] that enforces *k*-anonymity

Assumes: |PT |≥ *k*, and *loss* * |PT| = *k*

algorithm Datafly:

// Construct a frequency list containing unique sequences of values across the quasi-identifier in PT, *// along with the number of occurrences of each sequence.*

1. **let** freq be an expandable and collapsible Vector with no elements initially.Each element is of the form (QI, *frequency*, *SID*), where *SID* = {id_i : ∃*t*[*id*]∈PT[*id*]⇒*r*[*id*]=id_i}; and, *frequency* = |*SID*|. Therefore, freq is also accessible as a table over (QI, *frequency*, *SID*).
2. **let** *pos* ← 0, *total* ← 0
3. **while** total ≠ |PT| **do**
   - 5.1 freq[*pos*] ← (*t*[QI], *occurs*, *SID*) where *t*[QI]∈PT[QI], (*t*[QI],__, ___)∉ freq; *occurs* = |PT| - |PT[QI] – {*t*[QI]}|; and, *SID* = {id_i : ∃*t*[*id*]∈PT[*id*]⇒*r*[*id*]=id_i}
   - 5.2 *pos* ← *pos* + 1, *total* ← *total* + *occurs*

// Make a solution by generalizing the attribute with the most number of distinct values // and suppressing no more than the allowed number of tuples.

6. **let** *belowk* ← 0
7. **for** *pos* ← 1 **to** |freq| **do**
   - 7.1 (__, *count*) ← freq[*pos*]
   - 7.2 **if** *count* < *k* **then do**
     - 7.2.1 *belowk* ← *belowk* + *count*
8. **if** *belowk* > k **then do:** *// Note. loss* * |PT| = k
   - 8.1 freq ← **generalize**(freq)
   - 8.2 **go to** step 4
9. **else do**

// assert: the number of tuples to suppress in freq *is* ≤loss * |PT|

   - 9.1 freq ← **suppress**(freq, *belowk*)
   - 9.2 MGT ← **reconstruct**(freq)
10. **return** MGT.

```
~$ python3 datafly.py -pt
"IA_synthetic_USCensusDist_full_data_Output.csv" -qi "ethnicity"
"income" "age" "sex" "education" "marriage" "children" -dgh
"ethnicity_generalization.csv" "income_generalization.csv"
"age1_generalization.csv" "sex_generalization.csv"
"education_generalization.csv" "marriage_generalization.csv"
"children_generalization.csv" -k 30 -o
"k_30_anon_IA_USCensusDist_data_full.csv"
[LOG] Created output file.
...
[LOG] Current attribute with most distinct values is 'age'.
[LOG] Generalizing attribute 'age' for sequence 9111...
[LOG] Generalized attribute 'age'. Current generalization level is 3.
[LOG] 17530 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'income'.
[LOG] Generalizing attribute 'income' for sequence 4830...
[LOG] Generalized attribute 'income'. Current generalization level is
2.
[LOG] 12662 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'education'.
[LOG] Generalizing attribute 'education' for sequence 1781...
[LOG] Generalized attribute 'education'. Current generalization level
is 2.
[LOG] 7464 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'children'.
[LOG] Generalizing attribute 'children' for sequence 838...
[LOG] Generalized attribute 'children'. Current generalization level
is 2.
[LOG] 2157 tuples are not yet k-anonymous...
...
[LOG] Current attribute with most distinct values is 'ethnicity'.
[LOG] Generalizing attribute 'ethnicity' for sequence 127...
[LOG] Generalized attribute 'ethnicity'. Current generalization level
is 2.
[LOG] 0 tuples are not yet k-anonymous...
[LOG] Suppressed 0 tuples with Actual K=156.
[LOG] Writing anonymized table...
```

Figure: Datafly Algorithm[12]

[11] https://github.com/alessiovierti/python-datafly

[12] Computational disclosure control : a primer on data privacy protection - Sweeney, Latanya

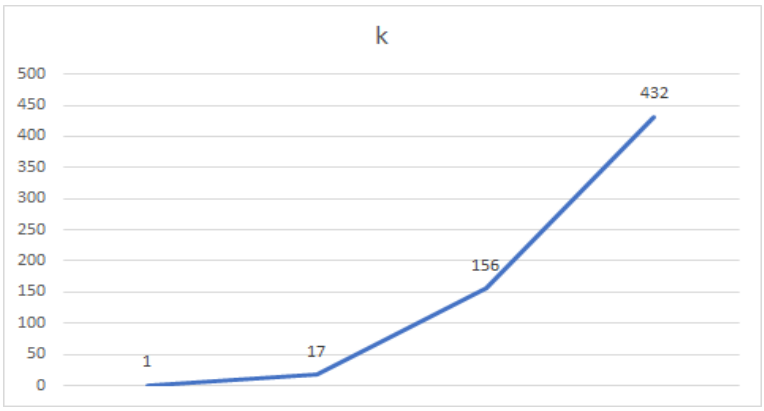


Figure: Data Anonymization

qID = {'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children'};

k=1 on the raw data mixed from voter data and synthetic attributes.

With current design of domain generation on personal features, the Actual K={1, 17, 156, 432} can be anonymized using Datafly algorithm.

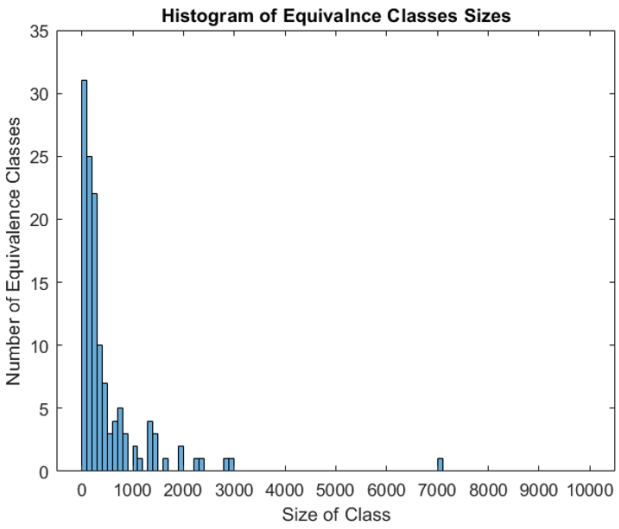
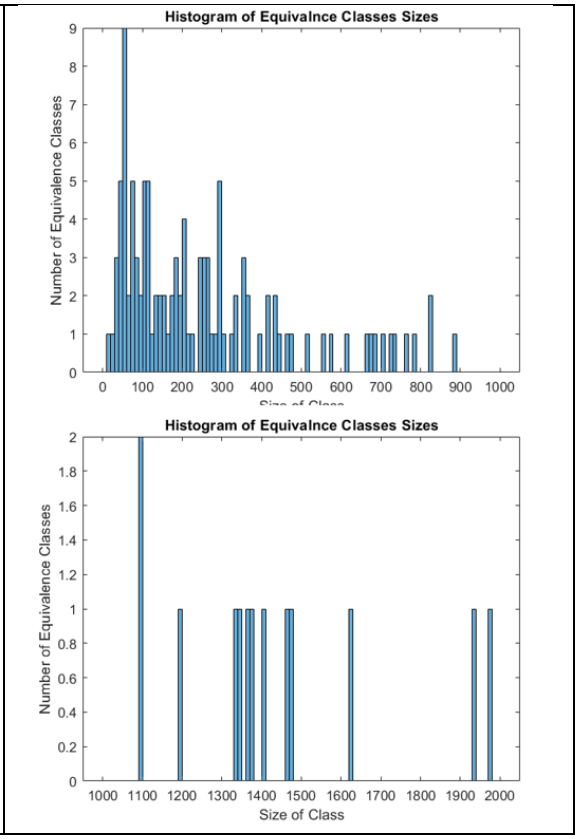


Figure: Equivalence Classes Distribution for Anonymization dat set with K=17



We first verified the raw data set extracted from voter registry and generated from attribute distribution has k=1. We have ran data fly algorithm with different desired k outputted with actual k graphed in Data Anonymization figure. On the actual k=17 data set, we have graphed the histogram of equivalence classes distribution in different resolution to understand quality of anonymizated data for the entropy value. In order to understand the quality of anonymization in respect to different choice of k, we graphed the discernibility metric and normalized average equivalence class size metric to determine the anonymized data set with k=17 will be the best choice to maintain best privacy while allow maximum utility.

$C_{DM} = \sum_{EquivClasses\ E} \lvert E \rvert^2$	Discernibility metric (CDM) assigns to each tuple t in V a penalty, which is determined by the size of the equivalence class containing t
$C_{AVG} = \left(\frac{total\_records}{total\_equiv\_classes} \right) / (k)$	The normalized average equivalence class size metric C_{AVG} measures how well partitioning approaches the best case. C_{AVG} is to reduce the normalized average equivalence class size

Figure: Utility vs. Privacy Analysis[13]

[13] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243250/

6 Recommendation

-

7 Conclusion