Disinformation Defense

AI Inference

Karel Baloun

Keng-Shao (Ken) Chang

Matthew Holmes

Master of Information and Cybersecurity (MICS)

University of California - Berkeley

Fall 2019 | Capstone | School of Information

cybersecurity@berkeley

Contents

1. Introduction

The Internet has no physical boundary on how personal data can be collected across the global, international community. The personal data collected and generated through the digital transaction of personal life can be considered as active and passive types 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 Artificial Intelligence under high-performance computing platforms using these vast amounts of personal data, 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.

Any consideration of individual data security needs to account for the mutual tradeoff among security, utility and privacy. We use a Venn diagram to illustrate the overlapping relationship between security and privacy in the result of regulatory requirements such as GDPR

and CCPA; between security and utility in the result of least privilege best practices; between privacy and utility in the result of data minimization from data science domain.

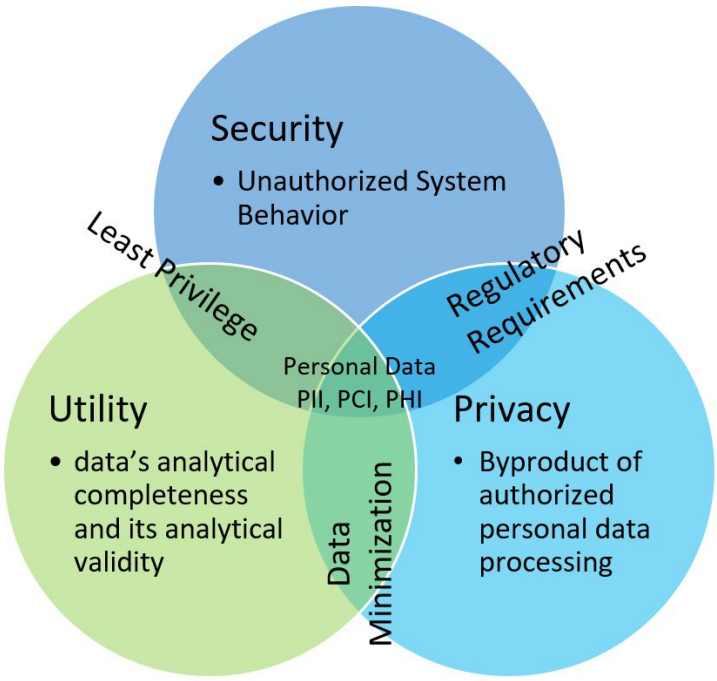


Figure 1: Security vs. Privacy vs. Utility[1, 2]

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

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 focus on exploring 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

1. Pillars of Personal Data

Our analysis of voter influence looks at four general types of individual data, which together we view as the 4 pillars of a potential organized attack on individual political autonomy.

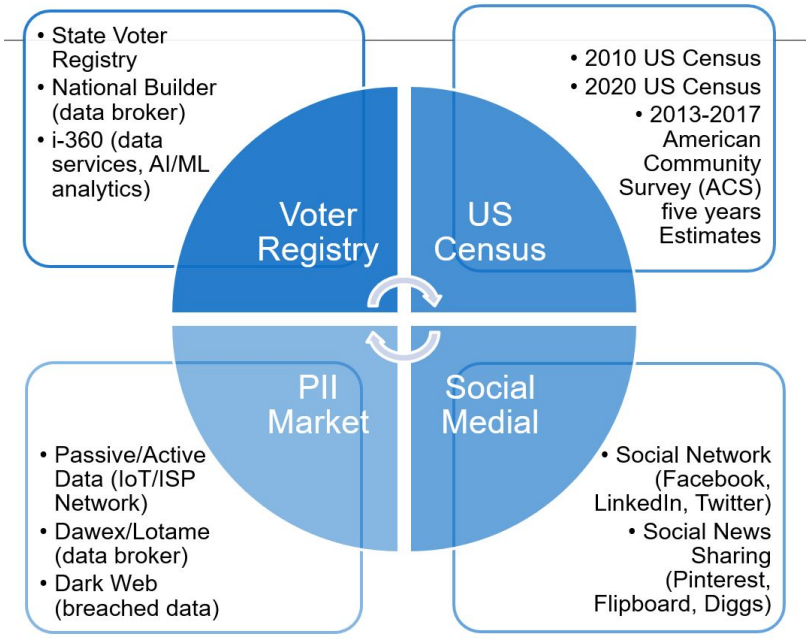


Figure 2: Pillars of Personal Data Sources

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.

After the Cambridge Analytica data scandal, the social media pillar has tightened up the access to social media user profile data. However, targeted advertisements to individuals based

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

on user profile and content-based behavior analysis remains the major source of revenue. The growing market for data brokers reselling personal and voter data has been on the rise.

2. Public Voter Records

Voter files identify registered voters, to ensure the legitimacy of all votes in an election. Name, Age, Sex, Residential Address are sufficient to definitively identify a person, and designate the local area in which they should vote. Party affiliation and history of voting help candidates campaign effectively, and also ensure that votes are counted credibly. Each individual who voted, can check that their vote contributed to the official tally. Several states add demographic fields, such as Florida and North Caroline including the race of voters.

Voting data does not reveal the actual vote, since all modern, fair elections are expected to offer a secret ballot. However, as more people self-report voting by party line, the confidentiality of each private vote is compromised. For example, if an area reports 80% of ballots were cast for a Republican, in an area of 70%+ Republican party registrations, and your public party affiliation is Republican, then an observer could correctly claim high likelihood that you voted republican. Exit polls of self-reported voting patterns further erode such anonymity. Keep in mind that the secret ballot is a cornerstone of a free election.

Since public vote counts are of course reported for every election, in small local polls, the confidentiality of a vote may be especially low. In the 90% partisan voting example above, if the republican candidate is reported as winning many more votes than his partisan constituency, then the likelihood that each republican voter casts his or her ballot with the party is further elevated.

3. Census Data

For any region in America, down to regions as small as a zip code or town or voting district, detailed demographic statistics are available, online via web interface or programmatically with a free developer API key. The canonical counts for each decade are fuzzed (in a way that is not publicly disclosed), since one legal directive on the census bureau is to guarantee a protection of personally identifiable information. K-anonymity is almost certainly enforced, in that data is not displayed on a small region in the case where the diversity of responses is not enough to guarantee anonymity. The degree of K-anonymity which the reporting guarantees is also not disclosed, nor is it easily discoverable.

For most census data tables, annual estimates of are projected from the canonical data, and are reported with margins of error. Beyond census data, other personal information is publicly available from other government records, for aggregation or combination with the voter records. Notably birth, marriage and death certificates are retrievable, along with other public records, including real estate transactions and court filings.

4. Personally Identifiable Information (PII)

American law and consumer habits support a well-developed marketplace for personally identifiable information. Data aggregators can match records on phone number, email address, device hardware identifier, internet tracking pixel, financial account identifier, physical address, vehicle or other registration records, and countless other loose identifiers. Each of these matched

records may contain unique financial transactions, social interactions, messages or posted text, opinions or preferences, or interpersonal relationship markers.

Financial transaction data, internet browsing behavior, and communication metadata are all available for purchase from the corporations that provide customers these services, and it is not possible as far as we know to opt out of these sold data collections. Even medical and DMV records are available for purchase or "sharing". Notably, none of these data are protected by online privacy legislation CCPA and GDPR, which focus on securing privacy from online services, not from all other businesses who store our data.

5. Social Media posts and profiles

If the registered email or phone number of a social media profile is public, then a data aggregator can join all of the public social media account information with the other personal information pillars. In the case of Facebook, just the public Likes (especially if sorted by recency) or public friends list of users would enable substantial classification. In the case of Twitter where every post is public, the entire publication history of a user can be attached, and not all Facebook users have been always aware of the distinction between friends/non-public and public posting.

Additionally, through the sophisticated ad targeting products of Facebook, Google and Twitter, it is possible to target advertising to specific people identified by email or phone number, as well as to visitors to your online presence identified by tracking pixel (e.g. retargeting). These ad APIs also offer targeting to "people similar to" any definable groups (e.g. Facebook's Custom Audiences).

2. Threat Modeling

Weaponized disinformation is already damaging election integrity, and campaign regulation and privacy law have not kept elections fair and secure. Our ML model prototypes behavioral prediction using public census data and estimates benefits of anonymizing interventions. AI speed and optimization utilizing personal information will soon pose dramatic threats to free and fair elections, and regulation disallowing personalized advertising and messaging in campaigns is essential.

1. Data Breach

Collective data breaches are leaking personally identifiable information (PII), Payment Card Information (PCI), and Protected Health Information (PHI) in the United States, and in other developed countries including Taiwan and Singapore. The personal data can be sensitive or not sensitive and 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. Such 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.

[5] https://www.techopedia.com/definition/14222/personally-identifiable-financial-information-pifi

The private sector has been under repeated, offensive attacks, resulting in countless 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 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 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.

Public sector institutions are also frequent victims of data breaches. 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 21 million people. This includes records of

[6] https://en.wikipedia.org/wiki/Yahoo!_data_breaches

[7] https://www.theguardian.com/technology/2017/nov/21/uber-data-hack-cyber-attack

[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/

people who had undergone background checks, but who were not necessarily current or former government employees. The same malware, Sakula, was used to attack other US companies which led to the conclusion that Chinese adversaries were behind the attack.

Social Media platforms have been targeted as mega data hubs of personal active and passive data. In 2018, one Facebook[10] data scandal revealed that Cambridge Analytica had harvested the personal data of millions of people's Facebook profiles without their consent and used it for political advertising purposes. While the FTC investigation did determine that both the developer and the platform were both responsible for this release of personal information beyond what members of the community were authorizing, comprehensive reports[11] on the 2016 election found many forms of election interference, especially Russian activity including the GRU and IRA, swung the election to Trump, not primarily this data leak. Indeed, a data scientist at Cambridge Analytica[12] at the time reported that their most important election contribution was high quality probabilistic matching between personal information profiles and voter records.

Online service providers, including broadband and wireless carriers, search and aggregation portals, purveyors of professional tools, merchants and retailers, and endless other commercial entities also store active PII and generate passive PII.

For analyssi of Assets vs. Threats, we applied threat modeling techniques using the CIA triad (Confidentiality, Integrity, and Availability) to analyze the pillars of personal data as assets

---

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

[11] e.g. political historian Jane Mayer's summary of Kathleen Hall Jamieson's "Cyberwar: How Russian Hackers and Trolls Helped Elect a President—What We Don't, Can't, and Do Know," in https://www.newyorker.com/magazine/2018/10/01/how-russia-helped-to-swing-the-election-for-trump

[12] personal video conference interview, with anonymous source on Thursday, Sept 26, 2019.

in addition to personal privacy to understand the possible attack vectors. We apply a holistic view on the whole data lifecycle, including collection, distribution, processing, and retention, highlighting potential vulnerabilities. We see personal privacy as an abstract asset with an increasing awareness of privacy as a basic human right.

Assets	Threats	Description
Voter Database	Confidentiality, Integrity	Public static database contains personal attributes such as full lega name, birth date, gender, residential address and registered party affiliation. Voter data broker provides services for distribution wi fee and might subject to data poisoning attack.
US Census Database	Confidentiality	Public statistical database contains summary statistics of person attributes such as ethnicity, income, education, marriage, and num of children. With combination of synthetic data generation, data aggregation and linkage, it is very likely to infer sensitive persona attributes such as vote preference.
Social Media	Confidentiality	Data leakage via public post with personal attributes, personal association with degrees of separation, and subject to targeted marketing advertisement.
Personal Data Market	Confidentiality, Integrity	Collective Data breach events creates black market of personal da Dark Web. Consumer service provider collects and resell the personal attributes for profit or in exchange of free service. Perso data exchange market promotes monetized model for individual. The personal data broker might subject to data poisoning attack.
Personal Privacy	Integrity	Personal service right might subject to perpetuate discrimination because machine learning result are trained on biased data. An individual might not receive the service they needed.

3. Design of Experiment

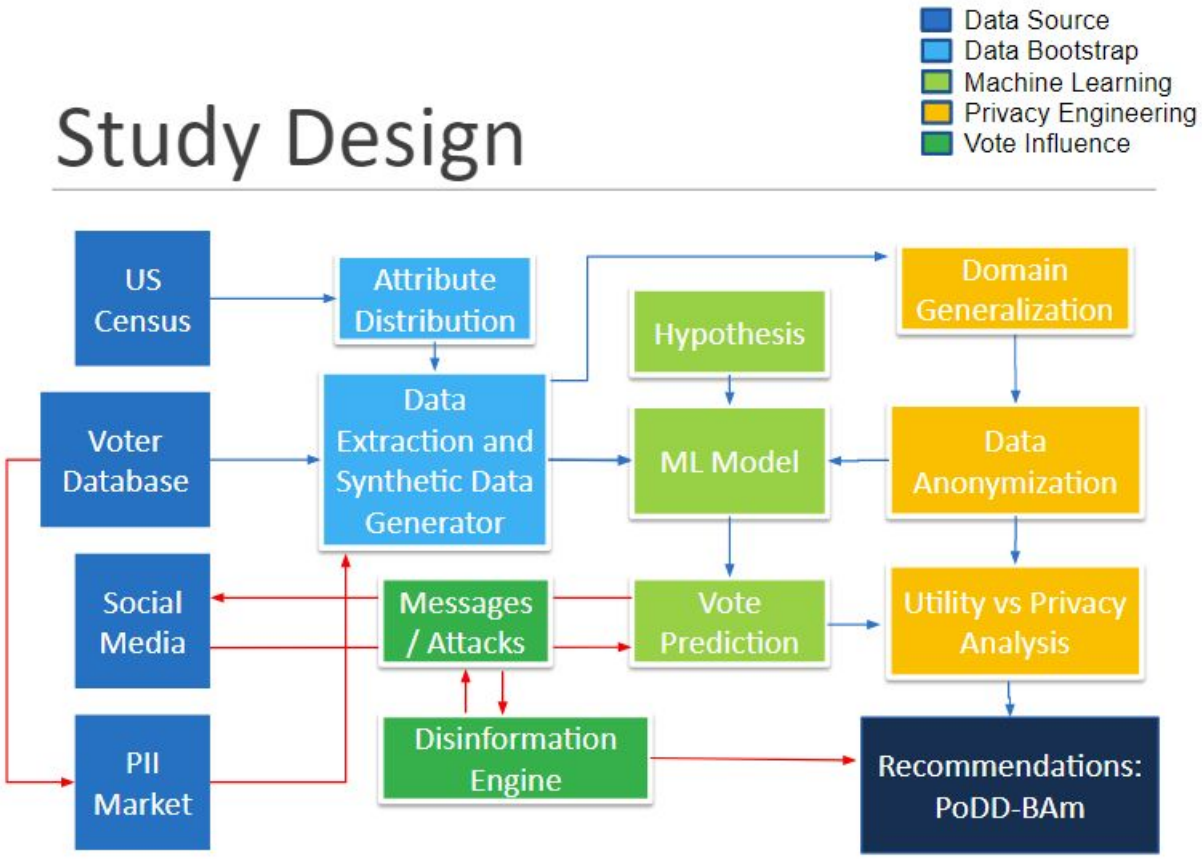


Figure 3: Design of our AI Inference Project

1. Data Bootstrapping

We designed and developed a Voter and Synthetic data generation algorithm to take input from sample voter databases, for IA-01 and FL-07, acquired as free academic research samples from Nation Builder. Linking these voter records to social media profiles, PII data from data broker, and dark web data were all out of scope for our research, due to budget and other

resource constraints. Such linkages are completely feasible, and well-resourced entities like Cambridge Analytica have already achieved them. US Census summary statistics of synthetic attributes were used to populate voter data output to simulate real data. If we had the budget to purchase real data from PII data market to avoid the use synthetic attribute, the overall experiment accuracy shall improve greater. Of course, to obtain and match 500K personal record from dark web would be another channel, but this would violate basic ethical research principles.

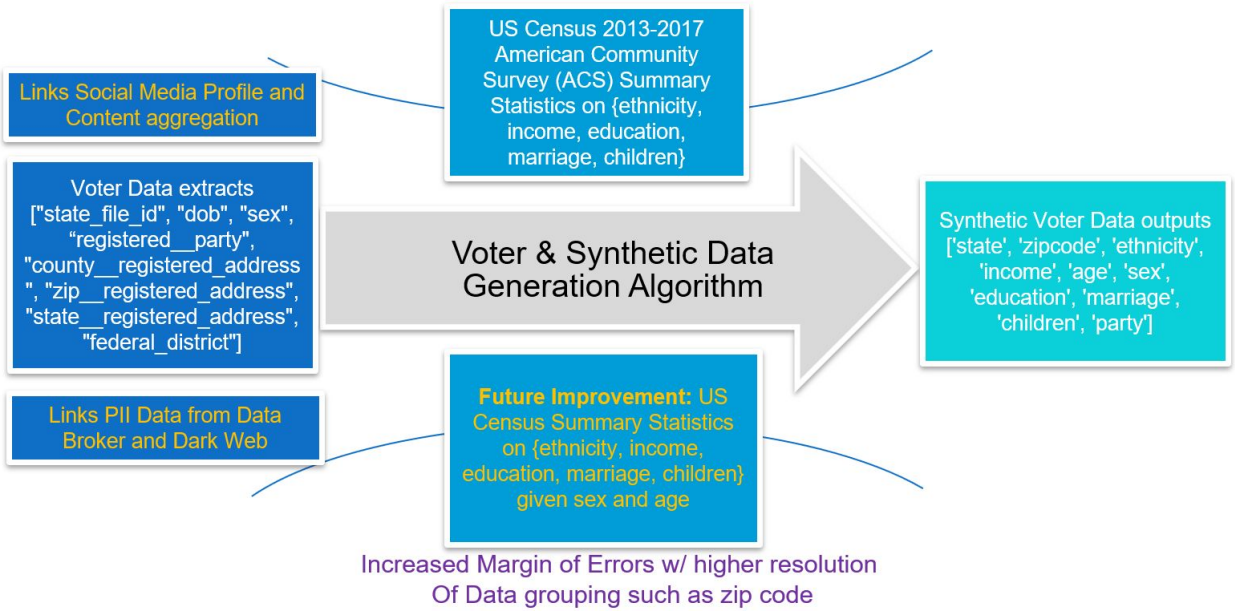


Figure 4: Voter Data Bootstrapping

2018 estimates are available from 2010 census data, for income, education, family size, and ethnicity. These can be filtered by geographic area, gender, and age. Data is cumbersome to extract from the web interface. Margin of error can rapidly overwhelm any information content,

as filters lower the number of class members. Due to margin error, it is 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.

2. Machine Learning (ML) Model

1. Hypothesis

Apply privacy engineering principle on k-anonymity as defense model, the risk of using machine learning algorithm to attack in result of voter preference inference shall be mitigated with much less accurate prediction on individual voting preferences from personal attributes, as illustrated by this figure:

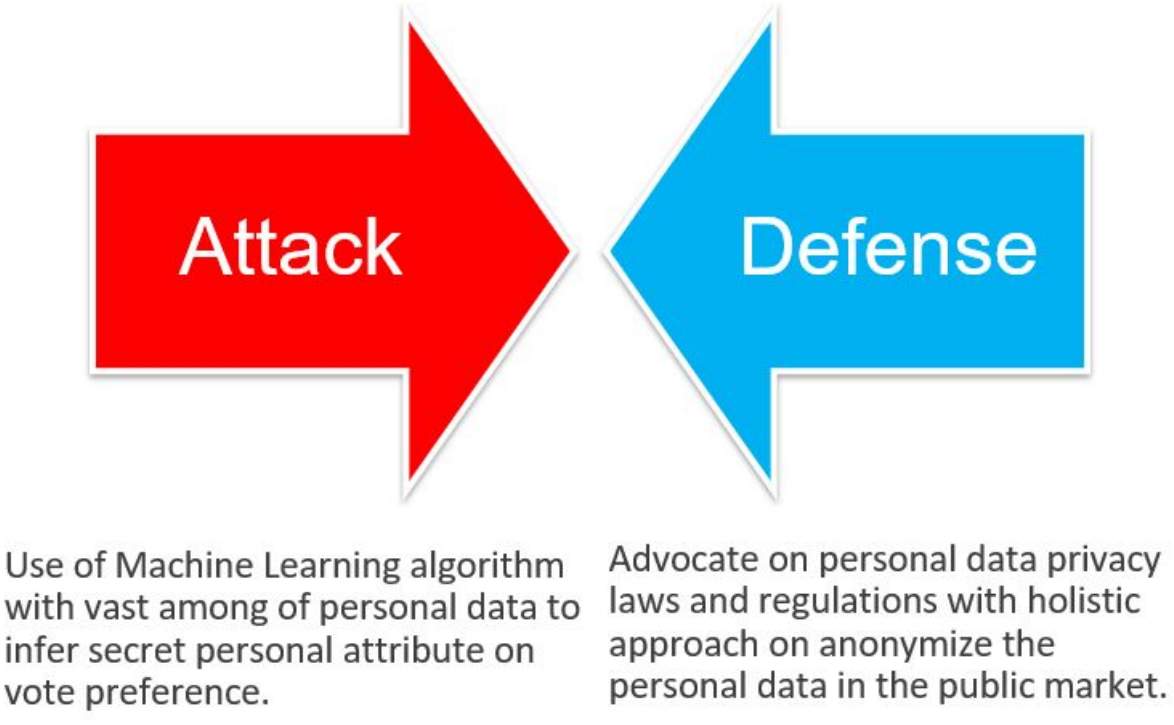


Figure 5: Voter Preferences Disclosure Attack and Defense

Personal attributes include {'ethnicity', 'income', 'age', 'sex', 'education', 'marriage', 'children'} extracted from the voter registry and linked with PII data from the US Census.

2. Supervised Learning with Classification

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

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 scores produced by using both the raw data and the anonymized data set.

3. Privacy Engineering

We illustrate the details of generalization tree for each feature that will be used in datafly algorithm with quasi-identifiers defined as ethnicity, income, age, sex, education, marriage, children. The generalization tree can be further tuned to allow fine grain data anonymization with different anonymization data set.

Ethnicity Generalization Tree:

{'NA','White','Black','Mexican','Native Indian','Asian'}
{0,1,2,3,4,5}
{0} -> {20} -> {30}
{1} -> {21} -> {30}
{3,4,5} -> {22} -> {30}

Sex Generalization Tree:

{'NA','Male',Female'}
{0,1,2}
{0,1} -> {10}-> {20}
{2} -> {11}-> {20}

Income Generalization Tree:

{'0-14999', '15000-34999', '35000-49999', '50000-74900', '75000-99999', '100000-149000', '150000-'}
{0,1,2,3,4,5,6}

Education Generalization Tree:

{'NA','elementary','middle','high','associate','bachelor','master','phd'}
{0,1,2,3,4,5,6,7}
{0}-> {20}-> {30}-> {40}

{0} -> {10}-> {20} -> {30}
{1,2} -> {11}-> {20} -> {30}
{3} -> {12}-> {20} -> {30}
{4} -> {13}-> {21} -> {30}
{5,6} -> {14}-> {21} -> {30}

{1,2,3}-> {21}-> {30}-> {40}
{4,5}-> {22}-> {31}-> {40}
{6,7}-> {23}-> {31}-> {40}

Marriage Generalization Tree:

{'widowed','not married', 'married', 'divorced'}
{0,1,2,3}
{0} -> {20} -> {30}
{1} -> {21} -> {30}
{3,4} -> {22} -> {30}

Children Generalization Tree:

{'NA', 'one', 'two', 'three', 'four', 'five'}
{0,1,2,3,4,5}
{0} -> {20} -> {30}
{1,2} -> {21} -> {30}
{3,4,5} -> {22} -> {30}

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

1. Vote Influence

Freedom is America's seminal and central value. Berkeley's own George Lakoff exhorts, "A threat to free will is a threat to freedom, the imposition of a dangerous worldview without public awareness. When free will itself is threatened, that is the ultimate threat to freedom."[13]

People do not want other people interfering with their own ability to make political decisions, in line with personal interests and values. People don't appreciate manipulation. However, at least when humans manipulate each other, the activity is both limited and fully precedented level of dishonesty. When AI manipulates us, perhaps on behalf of a rich benefactor or for any programmed agenda outside visible control, the level of scale it can apply to the problem, and the lack of auditable values driving its activity, are uniquely new and nefarious. Finally, soon arriving levels of AI sophistication could be so high, such that none of us would

[13] George Lakoff (2006) in Whose Freedom?, p62

become aware of its activity and influence. The AI may not even produce or leave any accounting or audit trail of its activity, beyond whatever is needed to train and update its models, which can be fully encrypted and hidden.

AI and Data, the two elements which together are reshaping so many facets of modern economy and society. Rapidly improving machine learning models for political speech along with easy access to the 4 data pillars above, together are driving methods for effective voter influence, by optimizing political messaging and iteratively measuring impacts. In our model, we trained the model on prediction of voter party, since this was exact attribute immediately in the voter record file; however, any number of other intermediate outcomes are available for training influence models, such as engagement on a delivered messaging, sending of a monetarily insignificant donation to show alignment, or participation at a political event or joining an online group.

Voters themselves are not a "big data" challenge, since only 5 to 30 million registered persuadable voters exist in the United States. Such a target audience could even managed with ordinary customer relations management (CRM) tools. However, the universe of possible messages is certainly a big data opportunity! A tremendous number of possible messages and message agents and methods can be evaluated for delivery, personalized to each voter and to the current context. Disinformation may not even be necessary, but if used, it can be so personalized and surreptitiously delivered, that it would never be discovered. Disinformation may also be impossible to operationally define or regulate.

Strongest messaging attacks often reinforce existing voter beliefs, tying into confirmation bias and other cognitive weaknesses, or aim to persuade at inaction. It is empirically easiest to disempower voters by encouraging methods to waste their voter, than it is to positively persuade towards support for a position.

Probabilistic profile matching as Cambridge Analytica pioneered enables the effective targeting of these messages, with advertising on social and digital media as only one delivery vector. Coordinated media posts, group and personal messaging are all even more pervasive. Narrow AI will organize and optimize this messaging, and the continually iterate to improve.

4. Attack: Machine Learning

1. Methodology

Our attack analysis has two parts: first, establish a baseline of predictive power using our raw unprotected data. Next, run the same type of classifier against anonymized data and compare the accuracy scores from both tests. Our hypothesis is that the anonymized data should test with lower accuracy. This would demonstrate that our data protection helps to preserve the privacy of individuals, and by extension helps to prevent the targeting mechanism of a disinformation messaging campaign.

2. Decision Tree

For our decision classifier, we decided to use the Scikit-Learn Decision Tree Classifier [cite?]. We chose this particular classifier mainly for its flexibility, predictive power, and ability to work with categorical features (such as the demographic data found in the US Census). We ran several experiments using single decision trees (see Section 4.3 for Results), but ultimately found better results using a Decision Forest. We'll describe two iterations of Decision Tree testing, and then follow with the description of the Decision Forest.

1. Decision Tree: Estimated Labels

In the first iteration of our experiments, our target variable was the true vote that each individual actually cast in their election. This data exists in tallies only, and we would never be able to actually get this data on an individual basis. So our first attempt at synthetic data was to check the election results on a district-by-district basis, and make sure we had the correct number of each vote type. For example let's assume that 80% of the cast votes were for the Democratic candidate, 17% of cast votes were for the Republican candidate, and the remaining 3% of the votes were for an Independent candidate. We would proceed record by record, and assign a "vote" to that record accordingly.

This strategy allowed us to have the correct distribution of votes at the population level of the district, but many of the individual records were almost certainly mis-labeled by our method. Since the decision tree is a supervised learning model, it depends on having accurate and real labels. Randomly assigning labels, even if it was accurate at the population level, was effectively

destroying any pattern in the data that the decision tree would learn. After seeing strange results, we quickly abandoned this method in favor of one that used real labels.

2. Decision Tree: Registered Party as Proxy

There are various things we could use as a proxy for the true vote cast at polling stations, but we chose to use registered party. We chose this because it was part of the voter registration database, and something that we knew was accurate. Since the voter database does not include demographic information, we had to rely on Age, Sex, and Zipcode from the voter database and synthesize the remaining demographic data.

Using the registered party as the target variable allowed us to get more sensible results, and results that are more consistent with a supervised learning model. However, on initial testing using single decision trees, we saw the opposite of our hypothesis: the anonymized data actually allowed for greater accuracy! In order to investigate this counter-intuitive result, we developed a decision forest with privacy engineering in mind for more robustness.

3. Privacy Aware Decision Forest

To improve the robustness and accuracy of our model, we moved to develop a decision forest where each tree had privacy engineering in mind. Using 600 trees, the model behaved more as expected, and helped to support our initial hypothesis. We suspect there was either an issue with over-fitting when using a single tree, or some other phenomena dealing with the patterns in anonymized data that only emerged after several iterations of the model.

Each tree in the forest has some awareness of the privacy engineering we've done to aid in efficiency and to not overfit the model. One example of this, is to leverage the knowledge of the number of Equivalence Classes in our data to not build a tree that is too large. Since our decision tree makes decisions based on the attributes, then it follows that the individuals from the same equivalence class would wind up classified in the same node at some point. As far as the tree is concerned, those records are the same. Since splitting such a node any further would not give any extra information, that makes that node a leaf. So we can conclude that the number of leaves in each decision tree should not exceed the number of equivalence classes in the anonymized data. The maximum leaf count is something that we can control when building the model, so we can readily make this optimization. For future work, we will look for other such improvements that will help when building our model.

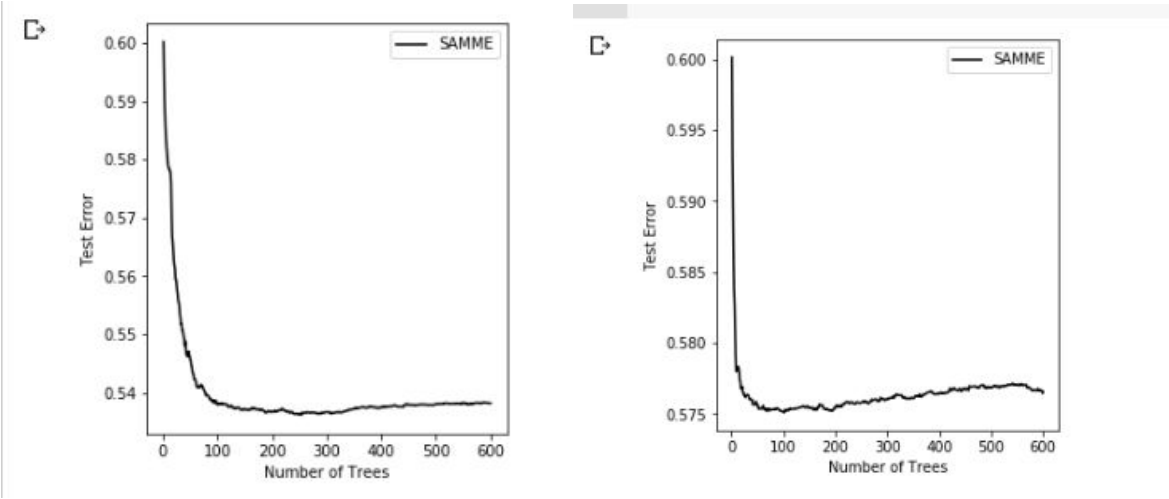
4. Results

1. Single Tree Results

Our results for raw data was around 40% accuracy, and our results for k=17 data was around 47% accuracy.

2. Decision Forest before Census Enhancement

Figure 6: k=1 test error (left) vs k=17 test error (right)



Using our decision forest, accuracy on the raw data was around 46% accuracy, and the accuracy on the k=17 data was around 43% accuracy. This evidence supports our claim that the privacy engineering reduces the ability of our ML model to predict vote preference.

3. Decision Forest after Census Enhancement

Figure 7: k=1 Full Census Data Test error (Best ~47% Accuracy)

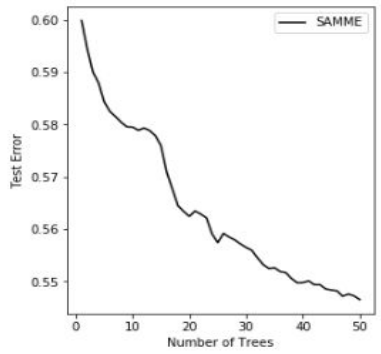


Figure 8: k=17 Full Census Data Test Error (Best ~47% Accuracy)

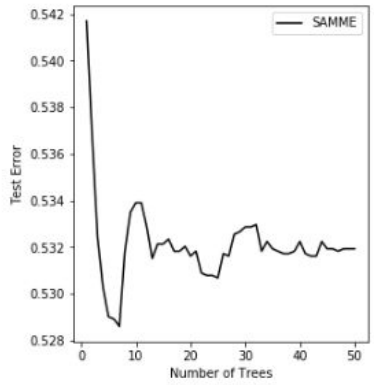


Figure 9: k=156 Full Census Data Test Error (Best ~47% Accuracy)

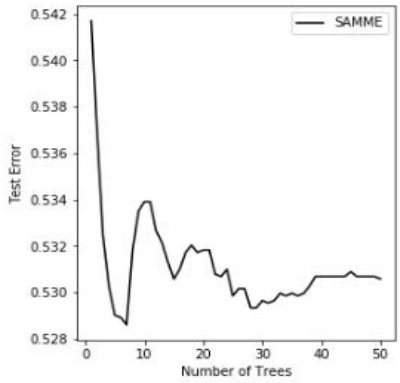
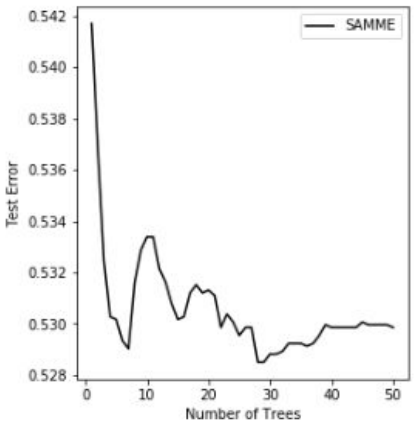


Figure 10: k=432 Full Census Data Test Error (Best ~47% Accuracy)



We suspect these results highlight a limitation of our current synthetic data creation, and we will discuss this limitation in more detail in the following section.

5. Limitations

1. Complete Data

Full demographic data that is linked to registered party does exist, it was just a limitation of this project that we couldn't acquire it or otherwise assemble it ourselves. The goal of the project isn't to obtain the highest accuracy possible for voter prediction, but rather to demonstrate the pipeline and the effect of privacy engineering. The lack of complete data lead to the next limitation.

2. Synthetic Demographic Information

Since we didn't have full true data, we decided to create synthetic demographic information. We didn't do this completely randomly, and using the census data, the population levels of each demographic are correct. However, like our initial synthetic vote generation, the specific demographics of each individual may not reflect any real person. This limits the model's overall accuracy.

In a practical sense, given a powerful enough ML model, the model may recognize the fact that the demographic data is synthetic, and therefore ignore it completely. However, since we use synthetic data in both the raw and anonymized data, we do not think that the effect of synthetic data alters our hypothesis. The accuracy score in absolute terms may change if we get real demographic data, but we expect the relationship between the raw accuracy and the anonymized accuracy to be the same.

Specifically referring to the Results section 4.3.3, we believe this is why the accuracy hovers around 47%. It's possible that the model is understanding that the demographic data is synthetic, and therefore any anonymization based on that demographic information is irrelevant.

3. Model Strength

The ML model's strength is closely related to the previous point about synthetic data. Given a powerful enough model, it may begin to learn that parts of our data are synthetic, and then ignore it. It is also possible that the decision forest model on its own is not powerful enough to capture the patterns in real data. Future work will address these concerns.

6. Next Steps

To address our first limitation, we will try to retrieve more complete data to enhance the demonstrative capability of the model. We aren't trying to get the highest accuracy possible, but it would help to illustrate our point.

To address concerns about synthetic data generation, we can be more careful in generating synthetic data based on the equivalence classes we get from the voter registration database. This will help to better capture patterns in the data. A further optimization would be to find population level data that cite our equivalence classes; this would mean we wouldn't have to synthesize as much data and would limit the population from the whole district to each individual equivalence class.

To address concerns about the strength of the ML model itself, we can continue to optimize model parameters using grid search when applicable, and privacy engineering analysis

otherwise. Additionally, we can research whether a Neural Network or SVM would be better suited to learn the patterns in our data.

5. Defense: Privacy Engineering

In our design for applying privacy engineering, we choose k-anonymity for simplicity and ease of implementation, with known effectiveness on defense against identity disclosure attack. Real time voter preference is sensitive attribute, and it would not be published in any voter record, but it can be inferred from other voter activities. Voter registry might contain the voter party, but it does not reflect the true vote.

To achieve k-anonymity, we choose datafly algorithm with domain generalization for categorical attributes; developed by Sweeney in 2017, it is a global, bottom-up, and greedy generalization algorithm. We have referenced an enhanced datafly implementation[14] by Alessio Vierti. The figure below illustrates the idea of datafly algorithm with inputs of raw data and domain generation tree (for example, ethnicity domain tree) with iterative process to apply domain generation or suppression on personal attributes selected as quasi-identifiers.

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

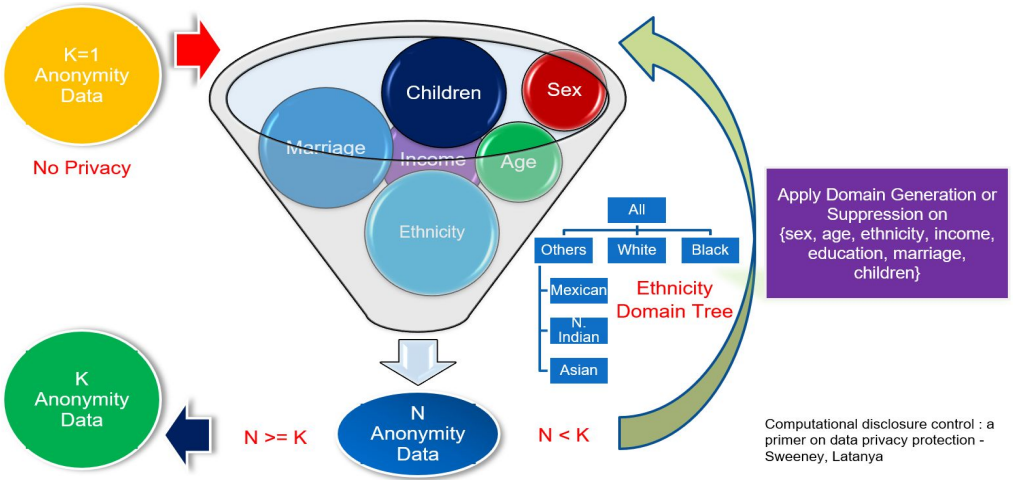


Figure 11: Voter Data Anonymization - Datafly Algorithm with desired K-Anonymity

The program output with design of algorithm illustrated below where you can see different levels of domain generation tree applied in different stage of algorithm until desired k has been met. The output of the program represents the desired K or greater data set as anonymized data set.

Core Datafly Algorithm

Input: Private Table PT; quasi-identifier QI = (A_1, …, A_n), k-anonymity constraint k; domain generalization hierarchies DGH_{Ai}, where i=1,…,n with accompanying functions f_{Ai}, 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*]⇒t[*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*]⇒t[*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 12: Datafly Algorithm[15]

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

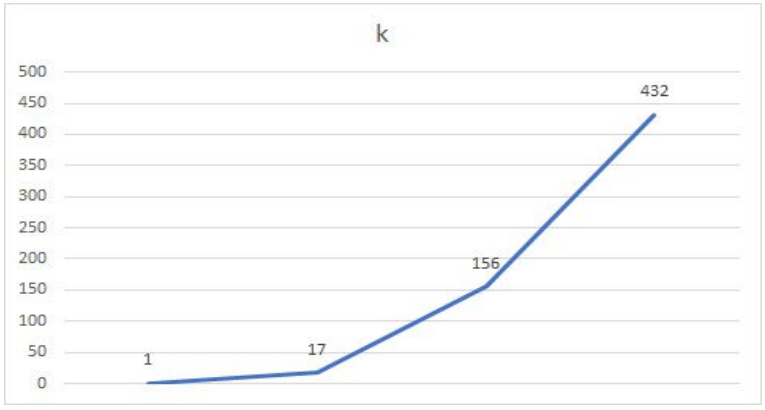
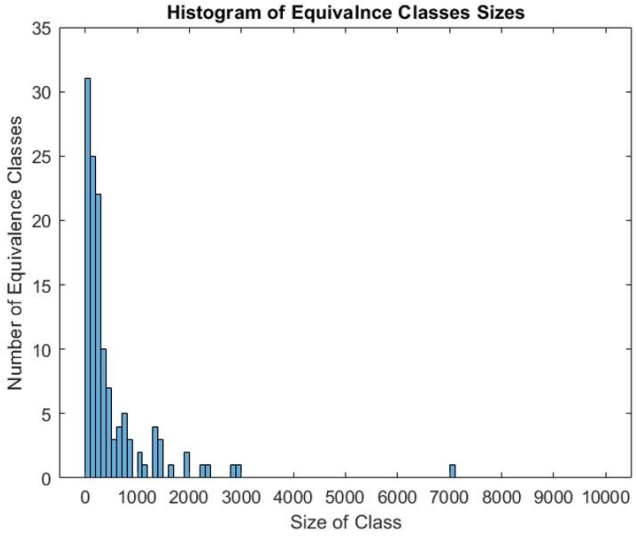


Figure 13: Data Anonymization

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.



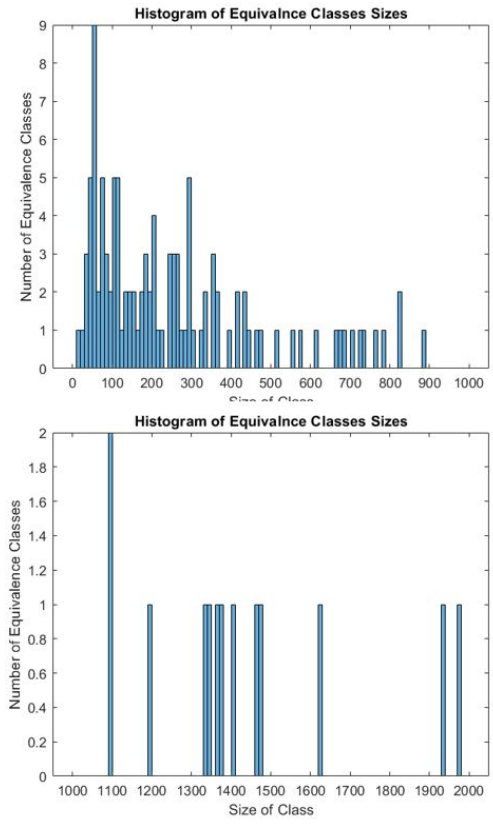


Figure 14: Equivalence Classes Distribution for Anonymization dat set with K=17

We first verified that the raw data set extracted from the voter registry and generated from our attribute distribution indeed has k=1. We have run the data fly algorithm with different desired k outputted with actual k graphed in the 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 anonymized 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.

$$C_{DM} = \sum_{EquivClasses\ E} |E|^2$$

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 15: Utility vs. Privacy Analysis[16]

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

6. PoDD-BAm: Recommendations

We propose 5 independent solutions, **PoDD-BAm**, with the first 3 for securing the privacy of voter information, and 2 to restrict the use of voter profiles and personal information in campaign outreach and messaging, as follows:

Po	Further study of **Privacy by Design** principles **on** collection, processing, and disclosure of personal voter and public census data..
D	Regulate the **data broker** industry with consumer privacy law, so everyone can view and opt-out any sensitive information.
D	Every Secretary of State should mandate **disclosure requirements** for access and use of **voter record files**.
B	FEC must **ban use voter personal profiles** in voter messaging, and ban use completely by campaigns and PACs.
Am	Large **audience minimum** size for advertising.

Any profile data that could be matched to a voter registration record, must be secured at all stages of the lifecycle of such data. To the extent that such personal data profiles continue to exist, it is at least necessary that every voter is able to opt-out of those data collections. The value of such data profiles could also be diminished if they could be corrupted with a large volume of false information, sufficient to make high probability joins impossible. Voter

registration data, while necessarily public to protect election integrity, should be controlled with maximal disclosure for access and use, so additional threats can be visible and evaluated.

We recognize that personal data profiles will likely continue to exist, due to their high commercial value, and that the public nature of voter records will make probabilistic matching possible, as Cambridge Analytica achieved. Even if these voter matched data profiles exist, the FEC must ban political campaigns, political action committees, and other political issue advertisers from accessing, storing or using any personal data. Most specifically, the FEC must not allow personalized messaging, in part because it can devolve to persuasive disinformation all too easily. Some advertising platforms[17] have announced that they plan to enforce significant minimum audience sizes (elevated from 100 up to several thousand) on all political advertising, since this prevents targeted personalization.

The FEC already regulates campaign finance and voter messaging, and the FEC has since 1992 practiced salting databases with fictitious records for the purpose of tracking sharing and use.[18] So expanding the FEC's mandate in this direction would be both feasible and normal regulatory activity.

[17] https://www.marketwatch.com/story/facebook-is-considering-changes-to-its-policy-on-political-advertising-wsj-2019-11-21

[18] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&ved=2ahUKEwiC2OeylK_mAhWPJTQIHRVLBjEQFjAJegQIAxAH&url=https%3A%2F%2Fwww.fec.gov%2Fpages%2Fbrochures%2Fsale_and_use_brochure.pdf&usg=AOvVaw1Br73Vob5lH1KYUUmpIpzg (google cache, due to an ongoing migration of the entire FEC document catalog)

7. Continuing Research

Further research would show stronger results to the extent that more sensitive data are made available to the model. NationBuilder sells complete voter profiles joined to voter records, and other companies likely already have even more complete profile data. To **Buy or Build such private profile data**, would enable iterations of the model to show precise benefits of data hiding with privacy engineering.

We have claimed that from a large big data universe of possible voter messaging, AI will optimize to choose the ones that are deemed most influential by the model. Further, we claim that AI can improve on this task over time. This claim would be strengthened further by **demonstrating Machine Learning generated messages**, for example of easy belief-conforming, attention-attracting voter spam.

Finally, any number of iterations on improving our initial machine learning **voter sentiment and behavior prediction model**, would result in improved measurement of the impact of these information attacks.

8. Conclusion

We claim that machine learning techniques are improving quickly, and that narrow AI for political messaging is close. Unfortunately, voters are easily to manipulate, even with simple bias confirming disinformation, and the value of predictably manipulating voters is high.

Elections must be protected by anonymizing sensitive voter data, restricting data accesses from campaigns and PACs, and regulating personalized messaging and unambiguous disinformation. Further study of various forms of branding and messaging on voter behavior

requires defensive academic study, since we claim that soon AI will be conducting live, real world experiments on influencing voter behavior.

With personal data profiles, by projecting the use of unlimited labor, we can foreshadow potential danger of AI power. If we imagine what a campaign could achieve with detailed voter profiles, given unlimited money, time and human participation, we could start to imagine how an AI could do similar things, completely within the practical constraints of an ordinary modern campaign. For example, an unlimited campaign could assign a personal, fully trained "political enthusiast" to every single voter, so that this one person's entire life is dedicated to managing the vote of a single voter. Since even a national campaign hinges on fewer than 5 to 30 million voters who's voting pattern can be influenced, such a level of investment is not hard to imagine. At $50/vote ($1B/20m), it is not impossible to assign a dedicated representative, including a ride to the polls. It would more than pay for operation of a narrow AI, which would digitally present itself as exactly the "kind of person" most influential to each individual voter. The AI could even order a Lyft to deliver a voter to the polls to seal the deal.

Our model, when taken to its algorithmic and data limit, shows how votes of every profiled voter could be predicted, and message influences on that prediction evaluated. With a huge universe of possible messages, including disinformation, AI can select and deliver exactly the most effective personalized messages for each person, thereby destroy the free will of each voter that is the basis for informed, fair democratic elections. We offer PoDD-BAm, our 5 solutions, to save democracy.

Democracy requires voters to have informed free will, and AI with personal data profiles of voters could capture the voting decision, without voters and observers even being aware that they were taken.

9.

10. Appendix

Github, https://github.com/adjprof/AI-Inference