Analyzing Members of Differentially Private Datasets Exposed by Attacks
(Technical Topic)

Analyzing Utilitarianism and Virtue Ethics Perspectives of Privacy Protection
(STS Topic)

A Thesis Project Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Youssef Errami

Fall, 2019

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signature: _____Youssef Errami_____

Approved _____ Date _____
Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Approved _____ Date _____
David Evans, Professor of Computer Science, Department of Computer Science

Introduction:

Over the last year, people have been starting to become aware of the lack of privacy with respect to their online data. Privacy discussions did not enter the mainstream media until the 2018 Facebook-Cambridge Analytica scandal came to light with many companies updating their privacy policies as a result (Lage, 2018, n.p). Due to these events, people started to care about their privacy with respect to the personal information that they have online, whether or not they voluntarily allowed their information to be there. With that said, the more important concern is that there are people attacking datasets and stealing private user information (Jayaraman: A, 2019, n.p). Currently, it is not clear what can be done to preserve people's privacy to the fullest extent because every dataset protected by some privacy mechanism so far are somewhat vulnerable to attacks where an attacker can acquire information from that dataset. People have been up in arms about their personal data online, to the point where cyber law experts, such as Omer Tene and Jules Polonetsky (2012), suggest that "policymakers must address some of the most fundamental concepts of privacy law" in order to protect user's privacy (Tene & Polonetsky, 2012, n.p). People's private data is at a higher risk of exposure than ever before. As we create better ways to protect our information from hackers, ethical dilemmas will rise when it comes to protecting users private information. The thought-process behind choosing what types of ethical practices should be followed leave plenty of room for ambiguity as opposed to a solid conclusion.

This prospectus is split into two sections: the technical and the STS approach to resolution. The technical approach is that currently, we suspect that all the members in the

dataset who are exposed by attacks share some set of characteristics but we do not know what these characteristics are. The STS approach would be that once these characteristics (if any) are found, it is crucial to evaluate the importance of the results from an ethical perspective and whether privacy mechanisms that protect datasets are good enough even if it does not protect everyone.

Technical Topic: Analyzing Members Exposed by Attacks from Differentially Private Datasets

The gold standard to ensure the most amount of privacy for any given dataset has been a principle called differential privacy. According to Nissim (2018), differential privacy is defined as "a strong, mathematical definition of privacy in the context of statistical and machine learning analysis". According to Jayaraman (2019: B), "Differential privacy has become a de facto privacy standard, and nearly all works on privacy-preserving machine learning use some form of differential privacy" (Jayaraman: B, 2019, pg 1). It is important to realize that differential privacy is not a specific *tool* but instead a *criterion*, which tools meant for analyzing private information have been devised to satisfy (Nissim, 2018, pg 2).

Differential privacy provides a guarantee of privacy protection from many types of privacy attacks, where here privacy attacks mean attempting to learn users private information from some sort of database, such as their credit card information or their social security number. The differential privacy guarantee states that "Differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis" (Nissim, 2018, pg 2). This guarantee is done through adding some random noise during your differentially private computation (Jayaraman:

B, 2019, pg 2). What adding this noise does is that the output of the differentially private computation done on a dataset will be the same whether that dataset contains person X or not. For example, let's say attackers are trying to figure out if you were a member of some dataset. They do not have access to the raw dataset but they do have access to the output from the differentially private computation that was done on the dataset. Due to the computation done on the dataset, the attacker would not be able to confidently say whether you are in said dataset because the output from the computation is the same whether you are in the dataset or not (Desfontaines, 2018, n.p). This idea is highlighted below in Figure 1 (Papernot, 2018).

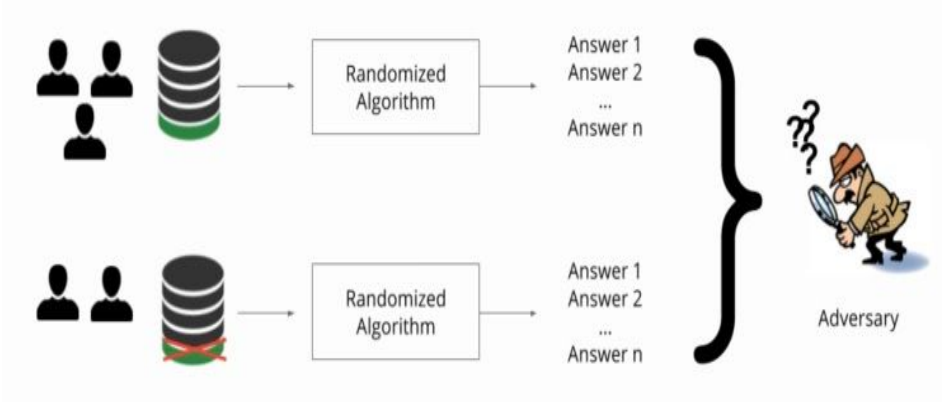

Figure 1, Differential Privacy Control Flow: This shows the control flow of differential privacy through inserting the random noise so that the output would be the same whether or not a certain person is in the original dataset (Papernot, 2018, n.p)

As seen through the experiments done in "Evaluating Differentially Private Machine Learning in Practice" (Jayaraman: B, 2019, pg 9-16), there are still members in datasets that are exposed by membership inference attacks even when differentially private computations are done onto the datasets. Membership inference attacks are when attackers use machine learning models to be able to infer an individual's membership in some dataset (Truex, 2019, pg 1). What I am working on finding out is are certain members of a dataset exposed strictly because of some characteristic about their data. If so, why are those the set of characteristics that allow these members to get exposed. If exposure is based on some set of characteristics, such as a person's

age or race, this could be a massive blow to user's privacy. It would be unfortunate if certain individual's data were more prone to being exposed compared to other people's data just because of some characteristic about themselves. Currently, privacy researchers in the field of computer science suspect that there are characteristics that these exposed members share. My capstone research project will be to investigate these exposed members by analyzing these differentially private datasets through computational geometric models and coming to a conclusion about what characteristics, if any, these members share.

STS Topic: Analyzing the Different Ethical Perspectives of Privacy Protection

Through viewing the results of the experiments done by Jayaraman and I in "Evaluating Differentially Private Machine Learning in Practice", we can see that for most experiments involving differential privacy, a majority of people in a particular dataset do not get exposed. These experiments raise an important ethical question of whether we as software engineers are okay with a majority of the people being protected even if the minority are at risk of their private data being exposed or should we strive to protect everyone no matter what? This destabilizing condition brings us to two ethical schools of thought which can help us better understand our two options along with the pros and cons that come with the two. These two schools of thought are known as utilitarianism and virtue ethics.

Utilitarianism is an ethical theory that determines right from wrong by focusing on outcomes. Utilitarianism holds that the most ethical choice is the one that will produce the greatest good for the greatest number (McCombs: A, 2019, n.p). The option that coincides with utilitarianism is to be content with our current methodology of differential privacy where we protect a *majority* of members in a dataset and allow a *minority* amount of members in a dataset

to be at risk of exposure through privacy attacks. On the other hand, virtue ethics emphasizes the "virtuous habits" such as honesty, bravery, fairness and generosity when making decisions, so it tends to bring normative decision making into question a lot more than utilitarianism (McCombs: B, 2019, n.p). One way to look to see if a decision is more virtue ethics-centered would be to see if said decision would be representative of developing the habits and character traits of a good person, and if so, then said decision embodies the ideals of Virtue Ethics (Raicu, 2013, n.p). The option that would coincide with Virtue ethics would be if software engineers and researchers were not okay with a minority amount of members of a dataset being exposed and the only acceptable option for them would be for no one to be at risk of exposure from a privacy attack. Some key differences between the two schools of thought are highlighted below in Figure 2.

Utilitarianism	**Virtue Ethics**
Reason-based approach to ethics	*Character-based* approach to ethics
Determines right from wrong by focusing on outcomes	Determines right from wrong by focusing on if the actions are moral and honorable
Endorses specific rules such as producing the greatest good for the greatest number	Guide for living life without giving us specific rules for resolving ethical dilemmas

Figure 2, Utilitarianism vs Virtue Ethics: This chart highlights some differences between Utilitarianism and Virtue Ethics that are important for understanding the two ethical schools of thought (Created by Author)

The perspective from a utilitarian approach here would be that the results we have found from our experiments would be enough proof that we have a good enough level of privacy in order to protect at least the majority of the dataset. However, by following this approach, we are okay with alienating the minority group as they will remain at risk of exposure to privacy attacks. The alienation could become even more jarring if we were to find out that there are

certain characteristics that allow for individuals to be exposed more often than those who do not embody said characteristics. If that were the case, it could be inferred that we as a society are okay with people of certain characteristics being at risk of attacks as long as the majority is safe, which could lead to even more injustice in our society than there is today. On the other hand, one point of view from a virtue ethics approach here would be that if we were to find a way to protect everyone in a dataset and guarantee that no member of said dataset is at risk of exposure from a privacy attack, then everyone is protected. With that said, a lot more research would have to go into trying to perfect this idea of differential privacy or even possibly look into a completely new privacy criterion that would allow for such a feat. It could take years, if it is even possible, for us to find a way to protect every member of every dataset from every possible type of privacy attack from exposure. An insight we can see from analyzing these two different approaches would be that differential privacy may be okay *ethically* if we know for certain which groups are protected and which groups are not protected. This is because knowing the limits of differential privacy's protection will allow us to react accordingly in trying to determine the costs and benefits of differential privacy thereby acting in an ethical manner. Privacy viewed from a utilitarianism and virtue ethics point of view help provide a framework for what software engineers and researchers could do moving forward with respect to the field of privacy. When deciding on what to do, they need to pick what set of values they want to prioritize given the pros and cons of both schools of thought.

Conclusion:

The future research I plan on doing with respect to analyzing the characteristics of exposed individuals can only help further our understanding of how to better protect people's

information. The deliverable for the technical portion of the project will be concrete evidence on whether there are any characteristics that exposed members of differentially private datasets share and if so, why is this the case. This will be presented in a thesis paper discussing all of the steps and actions I took from the beginning to the end of the project. The deliverable for the STS portion of the project will be to further analyze the two schools of ethical thought with respect to user's privacy. The thought-process behind choosing what types of ethical practices should be followed leave plenty of room for ambiguity as opposed to a solid conclusion, which is why this field should continue to be analyzed and discussed.

Word count : 1834

Sources and Annotations

Desfontaines, D. (2018, July 30). Why differential privacy is awesome - Ted is writing things. Retrieved from https://desfontain.es/privacy/differential-privacy-awesomeness.html. [Article suggested by Bargav Jayaraman (Ph.D candidate) and Professor David Evans which speaks to the trustworthiness and referenceability of this website]

Jayaraman, B. (2019, July 14). Why you should evaluate your private model. Retrieved from https://bargavjayaraman.github.io/post/evaluating-dpml-intro/.[Written by Bargav Jayaraman (Ph.D candidate) and editorial process by Professor David Evans in the Computer Science department at UVa]

Jayaraman, B., & Evans, D. (2019, August 12). Evaluating differentially private machine learning in practice. *arXiv, arXiv:1902.08874*

McCombs School of Business. Utilitarianism. (2019). *University of Texas,* Retrieved from https://ethicsunwrapped.utexas.edu/glossary/utilitarianism.

McCombs School of Business. Virtue ethics. (2019). *University of Texas*, Retrieved from https://ethicsunwrapped.utexas.edu/glossary/virtue-ethics.

Raicu, I. (2013, February 4). The ethics of online privacy protection. *Markkula Center for Applied Ethics*, Retrieved October 14, 2019, from https://www.scu.edu/ethics/privacy/the-ethics-of-online-privacy-protection/.

Nissim, K. et al. (2018, February 14). Differential privacy :A primer for a non-technical audience. *Harvard University,* Retrieved October 23, 2019, from https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf.

Papernot, N., & Goodfellow, I. (2018, April 29). Privacy and machine learning: Two unexpected allies? Retrieved from http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html.

Lage, A. (n.d.). Why are companies updating their privacy policies? It has to do with the general data protection regulation in the EU. Retrieved from https://www.bustle.com/p/why-are-companies-updating-their-privacy-policies-it-has-to-do-with-the-general-data-protection-regulation-in-the-eu-9198972.

Tene, O., & Polonetsky, J. (2012, February). Privacy in the age of big data. *Stanford Law Review,* Volume 64 (2011-2012), n.p, Retrieved October 14, 2019, from https://www.stanfordlawreview.org/online/privacy-paradox-privacy-and-big-data/.[Stanford Law Review is an online journal article and that is why there are no page numbers.]

Truex, S., Liu, L., Gursoy, M. E., Yu, L., & Wei, W. (2019, February 1). Towards demystifying membership inference attacks. *arXiv*, arXiv:1807.09173