Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

Diveplane Corporation

Response to Request for Information on Advancing Privacy-Enhancing Technologies

July 2022

Office of Science and Technology Policy (OSTP)

Document Number: 2022-12432

Re: Diveplane Corporation's Response to Request for Information on Advancing Privacy-Enhancing Technologies, OSTP 2022-12432

Introduction and Overview of Diveplane's Response

Diveplane applauds the OSTP's attention to this important topic and welcomes the chance to respond to the RFI on the use of privacy-enhancing technologies. Privacy and AI fairness are at the heart of Diveplane's (www.diveplane.com) mission and are the focus of its product development. Diveplane is led by multiple computer science PhDs and a team that has decades of relevant AI, technical, business, and legal experience.

The aim of this response is to present a summary of the potential risks associated with using a single method or metric for privacy-enhancing technologies ("PETs"), and to present an alternative approach using a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory that can significantly increase the potential to preserve privacy and accuracy while enabling the safe distribution of data. You will see below in our response the theme that current PETs, like differential privacy and k-anonymity can be unraveled or "attacked" in a way that can quickly erode the very privacy they are trying to protect. This is to say nothing of the possibility of accidentally synthesizing data that recreates PII or PHI, a mildly statistically improbable, but very grave, risk. As such, these single method or metric PETs can be false security blankets. We at Diveplane believe that a more comprehensive approach to creation of synthesized data is needed, including using a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory, as well as building tools to check the privacy of the synthetic data created vis-à-vis the original data. We believe that PETs as we describe promote continued innovation in emerging technologies in a manner that supports human rights and shared values of democratic nations.

Responses to the questions from the RFI:

The bulk of our response will be directed to question number 2, *Specific technical aspects or limitations of PETs*. After our response to question number 2, we briefly respond to a few of the other questions.

2. Specific technical aspects or limitations of PETs:

2(A) BACKGROUND: PRIVACY IS NOT ONLY ABOUT CHECKING THE BOX TO MEET CURRENT LAWS

Ensuring privacy of data is about more than masking data, whether that masking is done with differential privacy or another method, such as those discussed herein. The risk of reidentification of masked data is real. Consider the issues Netflix had releasing anonymized data in 2007. "Anonymizing data still enjoys a good reputation despite an abundance of evidence that it is too easy to defeat. In 2007, Netflix offered a million-dollar prize to the first algorithm that could outperform their collaborative filtering algorithm. The dataset they supplied was anonymized, but one group de-anonymized it by joining it with information from the IMDb database. An anonymized database can happily expose the PII (personally identifiable information) by combining it with a PII data source and matching other criteria (so-called latent values)."[1]

Further, it is important for every business to abide by the laws where it does business. The expansion of privacy laws and expectation of privacy from some consumers has led to class-action lawsuits seeking $5000 per individual.[2] Data can be incredibly valuable but keeping around the original data for future analysis or innovation can lead to additional liabilities. The average cost of a data breach is just under $4M USD per breach,[3] with some of the worst breaches getting in the realm of billions of USD.[4]

In addition to the legal liabilities, there are also reputational and customer selection effects that can impact the bottom line. Having a strong privacy stance can attract additional customers.[5] Conversely, not having thoughtful privacy practices can lead to undesirable self-selection of customers. Privacy can be a buffer to prevent misalignment of incentives between an organization and its customers. For example, an insurance company that collects data about how its customers drive may learn a lot, especially about how to price the auto insurance for different customer segments and improve profitability. However, if this data can directly tie back to customer behavior, then customers become incentivized in ways that are harmful to the business. Some safe drivers may think they will be unfairly discriminated against and so choose a competitor, whereas other drivers may attempt to game the system and perform undesired behaviors, like unnecessary excessive driving to lower premiums,[6] or even second-guessing their driving decisions when faced with split-second life-or-death situations with pedestrians.[7] Privacy safeguards, such as the use of synthetic data, can help ensure these misalignments of incentives are reduced or avoided, for example by softening the impact of bad luck on customers and reducing the chances of the customer having a bad experience, while still retaining the

[1] The fragility of privacy – can differential privacy help with a probabilistic approach? Neil Raden, 2020, https://diginomica.com/fragility-privacy-can-differential-privacy-help-probabilistic-approach
[2] https://www.forbes.com/sites/daveywinder/2020/06/03/google-chrome-privacy-lawsuit-could-you-get-a-5000-payout-incognito-mode-class-action/
[3] https://www.ibm.com/security/digital-assets/cost-data-breach-report/
[4] https://www.eweek.com/security/epsilon-data-breach-to-cost-billions-in-worst-case-scenario/
[5] https://www.geekwire.com/2019/privacy-becomes-selling-point-tech-companies-apple-microsoft-leading-way/
[6] https://blog.joemanna.com/progressive-snapshot-review/
[7] https://www.reddit.com/r/Insurance/comments/7ck6kh/dont_sign_up_for_progressive_snapshot_if_you_have/

insights, analytics, and general business model from data that can drive the organization to further success.

2(B) THE RISK TO DATA PRIVACY FROM MULTIPLE ATTACK VECTORS WHEN USING COMMON PETS

Given the importance of privacy, it is important to select an effective solution to ensure privacy of the data while maximizing utility, insights, and accuracy. It can at first seem appealing to select a technique that makes intuitive sense regarding privacy, like **k-Anonymity**, where you ensure that the data released has at least some k records that are sufficiently similar and not uniquely identifying. Or it may seem appealing to apply a technique that is well-known, like **differential privacy** and be done. However, using just one of these anonymity or privacy techniques without examining the different ways the data and the privacy can be attacked can yield unintentional, unknown, and potentially vast privacy leaks. For example, the principle of differential privacy is mathematically solid, but when used alone, it is easy to make egregious mistakes without knowing it.

Further, while well-known techniques, like k-anonymity or differential privacy, can measure or ensure a certain aspect of privacy, there are no globally accepted standards that can tell you how good your privacy actually is or that can characterize the risks of reidentification. This leaves key governance stakeholders to assert their own definition of "good" privacy. The risk of poor privacy practices is a long-term existential risk, especially as new attacks on privacy and personal autonomy continue to emerge. Masked or privatized psychographic, activity, location, financial, or preference data points are often considered to be nonidentifying in some contexts and organizations today. These "anonymized" data sets may be considered safe to distribute, and in some cases they are. However, as data-based insights continue to escalate in ubiquity and efficacy, privacy models based on a one-to-one mapping between the original data and the outputs of analysis and data sharing have potential risks when combined with other data sources, known as auxiliary data, regardless of whether they were anonymized. The ability to combine these pieces of data, which may be unforeseen when the analysis is performed or data released, increases the potential for attack vectors to privacy. These privacy leaks can even yield information far beyond what was intended or included, such as mental health patterns suggested by a user's historical patterns.[8]

Here are some discussions from the literature:

> "The rapid decrease in the sequencing technology costs leads to a revolution in medical research and clinical care. Today, researchers have access to large genomic datasets to study associations between variants and complex traits. However, availability of such genomic datasets also results in new privacy concerns about personal information of the participants in genomic studies. Differential privacy (DP) is one of the rigorous privacy concepts, which received widespread interest for sharing summary statistics from genomic datasets while protecting the privacy of participants against inference attacks. However, DP has a known drawback as it does not consider the correlation between dataset tuples. Therefore, privacy guarantees of DP-based

[8] E.g., https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2483426

mechanisms may degrade if the dataset includes dependent tuples, which is a common situation for genomic datasets due to the inherent correlations between genomes of family members."[9]

"Four spatio-temporal points are enough to uniquely identify 95% of individuals."[10]

"Our study shows that sharing anonymized location data will likely lead to privacy risks and that, at a minimum, the data needs to be coarse in either the time domain (meaning the data is collected over short periods of time, in which case inferring the top N locations reliably is difficult) or the space domain (meaning the data granularity is strictly higher than the cell level)."[11]

THEORETICAL EXAMPLE 1: A restaurant chain collects customer data to be analyzed by a consumer analytics firm as part of a preference-based marketing campaign and applies differential privacy. Unrelatedly, the consumer analytics firm also collects data from a separate group that collects data about the retail area around the restaurant, and this group only uses data masking to remove apparently identifying features. Individually, these data sets may pose little risk to privacy leaks. However, because the differential privacy was applied per transaction and not per customer, the customers who frequently visited the restaurant had a vastly eroded privacy. Further, the retail analytics firm didn't mask out certain fields for one of the boutique retailers, which sometimes contained names and addresses. When the restaurant data is coupled with the retail data by the consumer analytics firm, it is possible to not only match up and reidentify most of the frequent customers who also visited this boutique retailer, but to obtain the person's name and address for some of those records. A customer of the third party had poor data security and the analytics are leaked to the internet. One of the reidentified restaurant and boutique store's patrons was a high-profile individual who had a picture taken at the restaurant. A savvy internet follower combines this picture with the leaked data and posts private information about the high-profile individual, who then files suit against all three firms. Even though this specific combination of events sounds rare, when enough data is collected and combined, such rare events become common and can have lasting horrible effects for individuals and firms.

Each individual privacy model bears several weaknesses when pitted against various known attack vectors, so no single privacy model or single privacy measurement is sufficient to defend and identify vulnerabilities in the presumed privacy of data. Regulatory standards like CCPA, GDPR, and COPPA are all, generally speaking, liability regimes and do not outline safe harbor practices to either shield you from liability or to protect your data. As such, compliance with any particular privacy regime provides incomplete defense against experienced actors and increasing volumes and types of data collection. The following table details a number of popular privacy techniques and measures and the types of attacks that may be effective against them.

[9] Almadhoun et al., Bioinformatics. 2020, https://academic.oup.com/bioinformatics/article-abstract/36/6/1696/5614817
[10] de Montjoye et al., Sci Reports, 2013, https://www.nature.com/articles/srep01376
[11] Zang & Bolot, ACM MOBICON, 2011, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.651.44&rep=rep1&type=pdf

Here is a summary of *known* attack vectors against PETs[12]:

Privacy Model	Attack Model			
	Record Linkage	Attribute Linkage	Table Linkage	Probabilistic Attack
k-Anonymity	✓			
MultiR k-Anonymity	✓			
ℓ-Diversity	✓	✓		
Confidence Bounding		✓		
(α, k)-Anonymity	✓	✓		
(X, Y)-Privacy	✓	✓		
(k, e)-Anonymity		✓		
(ϵ, m)-Anonymity		✓		
Personalized Privacy		✓		
t-Closeness		✓		✓
δ-Presence			✓	
(c, t)-Isolation	✓			✓
ϵ-Differential Privacy			✓	✓
(d, γ)-Privacy			✓	✓
Distributional Privacy			✓	✓

2(C) ENHANCING PRIVACY THROUGH A COORDINATED, ROBUST APPROACH, USING A K NEAREST NEIGHBOR PET THAT INCLUDES ROBUST, PROBABILITY-THEORY-DRIVEN KERNELS WITH INFORMATION THEORY

To address the potential problems from using single privacy models it is possible to use multiple privacy models to significantly increase the robustness of the of the synthetic data while maintaining high levels of accuracy. We at Diveplane believe that the best approach is to mathematically unify a variety of privacy enhancing techniques, thus enabling a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory. This technique enables maximum entropy noise, achieves various forms of differential privacy, and simultaneously utilizes the sparse space of multidimensional data to only synthesize new data that is sufficiently "surprising" (by the information theoretic definition of surprisal) relative to any original data. By doing this, one can generate new data that maintains both the underlying distributions of original data and accuracy under analytic techniques, while also allowing privacy to be maintained such that none of the synthetic data is too close or similar to the original data, all while not relying on a one-to-one relationship between the original data and the synthesized data. Not only must these techniques be used at generation, but we also believe the generated data should also be tested, using a data quality tool evaluation suite, after synthesis to make sure it maintains the expected privacy. See Section 2(D) below.

At Diveplane, we believe that those hoping to truly advance privacy using PETs should synthesize data using a PET that can create a verifiable synthetic 'twin' dataset with the same statistical properties of the original data, but without including the real-world confidential or personal information. This results in a data set rich in value but with no risks of re-identification. We believe this can be accomplished with a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory, and that:

[12] Fung et al., ACM Computing Surveys, 2010, https://dl.acm.org/doi/abs/10.1145/1749603.1749605

- These synthetic data sets are far more robust, accurate, and safe compared to other masking and privacy techniques that may still be susceptible to reidentification, based on our experience with thousands of data sets.

- These synthetic data sets help companies navigate through national international privacy laws including GDPR, CCPA, CPRA, and HIPAA.

- These PETs can be used to create datasets that look, act, and feel realistic for the purposes of data modeling and analysis, but do not contain information to identify an actual person from the original dataset.

- These PETs can create synthetic data from all structured data, including time series data and relational databases.

- These PETs should include a data quality tool that enables users to identify the levels of privacy and accuracy of the synthetic data vis-à-vis the original data.

Benefits over differential privacy of a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory

Differential privacy is a pillar of modern privacy enhancing technologies. It is a mathematical technique that ensures that, for a given probability as modulated by a "privacy budget", the statistical results of certain types of analysis of a database with an individual's record in it should in expectation be indistinguishable from the statistical results as if the individual's record were not in the data. This creates a strong plausible deniability for analysis and generally prevents an individual from proving that they were in the data. As strong as differential privacy is, it is equally as easy to misuse in practice. If data has multiple records for individuals, or there are some strong correlations in the data such as family records or geographic information data, the actual privacy achieved may be far less than planned. Further, the typical applications of differential privacy either require upfront planning of analytics or use ML systems that train via a differential privacy budget but may extract private information into the model via correlations.[13]

Further, because each feature or combination of features could potentially be statistically identifying when combined with future data, all fields must be considered as potentially sensitive and have some form of differential privacy applied. If differential privacy is used in this manner, it leads to noisy, unusable data sets. An additional concern, though rare, is that certain queries with differential privacy can make one person appear like another, so synthetic data solutions that only rely on differential privacy may have more than one chance to synthesize data that appears almost exactly like the original data.

A k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory applies mechanisms that are differential privacy to all relationships in the data, not just the sensitive fields. It scales the noise based on the predictability in the data, ensuring that privacy and anonymity are maintained in every area, from dense clusters to far outliers. This combination of differential privacy

[13] E.g., Hitaj et al., ACM SIGSAC, 2017 (https://arxiv.org/pdf/1702.07464.pdf) and Papernot et al., ICLR 2017 (https://arxiv.org/pdf/1610.05755.pdf)

techniques acts as a sensitive data cross-shredder, while making sure it still generally follows the underlying distribution of the data.

2(D) PETS SHOULD NOT STOP AT SYNTHETIC DATA GENERATION, BUT SHOULD PROVIDE A WAY TO CHECK THE PRIVACY OF THE GENERATED SYNTHETIC DATA VIS-À-VIS THE ORIGINAL DATA

Benefits of using an anonymity preservation check as a data quality measure after creating synthetic data with PETs

We at Diveplane believe that PETs should include anonymity preservation check, which is an approach that finds for each data point in the synthesized data set, the closest corresponding data point of the original dataset, and measures how easy it would be for an original data point to be recognized in the synthetic data. We believe that a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory can generate data points that are just outside of the certainty manifold of predictability from the original data, ensuring that it doesn't generate anything too predictable within the data set.[14] A k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory can find the nearest data so that a user can audit and determine what is the closest synthesized datapoint to any original datapoint, thereby enabling the user (i.e. data scientist) to check whether the synthesized is too close to the original data. This measurement is a ratio between 1) the distance between a synthetic data point and the closest original data point and 2) the average or minimum distance between a data point and its closest neighboring data point in the original data, among the locally relevant data points and also all global data points. Distance in the can be scaled according to the data's range for each feature. A ratio greater than 0.5 for minimum distance ratio is a good indicator that privacy being preserved even in the worst case, as it is possible that any given data point is sufficiently could potentially be mistaken for another. A ratio of 1.0 or higher means that privacy is being preserved well enough such that in the worst case, an individual record from the synthetic dataset looks at least as different as any two different cases in the original dataset, and any higher value means that privacy is even stronger. The indices of the closest matches are reported so that an auditor could inspect the data to ensure that privacy is maintained.

THEORETICAL EXAMPLE 2 – Using anonymity preservation to check privacy for synthetic data. Consider a theoretical company that uses differential privacy with a GAN (generative adversarial network) to produce synthetic data, and distributes it to several parties. Even though the data is synthetic and used differential privacy properly, somebody believes they are recognizable in the data and files a lawsuit against the company. Because a dozen other people recognize themselves in the millions of synthetic data points, this lawsuit turns into a class action lawsuit. Had the company used a model that included anonymity preservation checking, the system would have inherent guardrails that would deny any synthetic data from being created that did not meet the minimum distance criteria.

Benefits of using entropy comparison as a data quality measure after creating synthetic data with PETs

[14] This can be easily parameterized for workflow needs. For example, if there are default data point that is repeated frequently and identically in the data but does not match an individual and thus won't leak privacy, these can be optionally included.

Another technique that is often used is entropy comparison. Entropy comparison compares the entropy and KL-divergence to provide a measure of disorder/surprisal within a data set. Entropy of a dataset is a measure of how compressible or predictable a given feature is, which is generally correlated with the skewedness of the class distribution. Similar values for original and generated entropy are an indication of similar class-wise distribution. KL-divergence values measure how similar the distributions of the generated and original dataset are.

A dataset which fails this measure is an indicator that the original data is very noisy and unpredictable or that the privacy model/algorithm tools being applied is underperforming or not working. If the privacy system adds significant noise without degrading accuracy, the original data set has a strong signal and with predictable values.

Benefits of using equivalence class measurements (k-Anonymity, l-Diversity, and t-Closeness) as data quality measures after creating synthetic data with PETs

The three equivalence class measures track equivalence classes, records in which many people share one or more characteristics (zip code, ethnicity, eating preference, etc.). For privacy, one person should not be an island. Any outlier, the only or one of very few people in an equivalence class, may be discoverable.

These measures are designed more specifically for anonymized or masked data rather than synthetic data. They are still a useful heuristic for exploring characteristics of the data set. These concepts are closely related to how a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory adds noise; in addition to the other methods mentioned, a k nearest neighbor PET that includes robust, probability-theory-driven kernels with information theory adds noise from the global distribution into every area of the data to ensure that these metrics are met in expectation, even if the metrics were not built to apply to synthetic data. For example, if a synthetic data point has insufficient anonymity under k-Anonymity, but it does not remotely correspond to any original data point, then the only privacy loss is that of a fictitious data point.

k-Anonymity measures the number of values within an equivalence class. If an equivalence class contains fewer than k values, that is considered a violation. k-anonymity is included to identify cases where there are low number of observations in an equivalent class (or outliers). It is useful for identifying cases where PII may be leaked if the equivalent class is unique. This protects the user from sharing data where a real person is potentially identified from auxiliary information.

l-Diversity measures the entropy within an equivalence class. If an equivalence class contains fewer than l unique values, that is considered a violation. This metric is not designed for determining privacy for synthetic data but is still a useful heuristic for certain characteristics of the data set. This metric is included to provide a measure of the diversity within a subpopulation (equivalence class) and ensure each type / outcome is represented in the class. It is useful for detecting the prevalence of an outcome throughout the dataset (e.g., is noise injected everywhere?) and protects against the scenario where a group is identified through a unique set of features criteria.

t-Closeness measures the distance t between distributions of an equivalence class and the global distribution. As t shrinks, a data set can be considered more private. This metric is not designed for determining privacy for synthetic data but is still a useful heuristic for certain characteristics of the data

set. This metric is a measure of the distribution of a subpopulation vs. the overall population. It is useful for detecting cases where a group may be identified from a set of unique feature values (equivalence classes).

Responses to additional questions from the RFI

Question 1. ***Specific research opportunities to advance PETs.***

We at Diveplane believe the government has the opportunity to advance PETs by funding development of k nearest neighbor PETs that include robust, probability-theory-driven kernels with information theory as well as post-synthesis data quality tools that help data handlers assess the privacy of the synthetic data that they have created.

Question 2*. Specific technical aspects or limitations of PETs*: See extensive response above.

Question 4. ***Specific regulations or authorities that could be used, modified, or introduced to advance PETs.***

As we have previously expressed, https://www.nextgov.com/ideas/2020/09/need-unified-data-protection-us/168643/, we believe that it would increase the certainty, understandability, and compliance by data-handling companies if the US had federal privacy regulations instead of a scattershot of laws and enforcements at the state and local levels. Further, and while we understand that creating safe harbors is difficult, the inclusion of safe harbors in the federal privacy regulation that outline approaches that companies could take to comply with the law would allow the vast majority of companies to comply by simply using these approaches.

Question 7. ***Risks related to PETs adoption:*** Our response to #2 also includes extensive discussion of the risks related to common PETs, such as differential privacy.

Question 8. ***Existing best practices that are helpful for PETs adoption:*** As we note in our response to #2, we believe PETs should have accompanying data quality tools to help assess the privacy of the generated synthetic data *after* it has been generated.