Applying Differential Privacy to Venmo Transaction Data

Ram Ben-David, Amanjot Samra, Jordan Meyer, Kristy Edwards
UC Berkeley MIDS and MICS, W233, Spring 2022

1. Introduction: Why is the problem relevant and important

Growth in US consumers' usage of web-based, mobile, and social media applications for daily activities, has produced enormous increases in personal data collected and maintained by corporations. The simultaneous increase in the number of applications available, for all aspects of one's life, contribute to the growth of these datasets – some of which are publicly available and may contain sensitive information that the users are not aware of.

The payment app Venmo is one such application. Venmo was conceived because of inefficiency of traditional money exchange systems and gained national popularity essentially via word of mouth and peer-to-peer networking. It has evolved to become a social networking platform in addition to its intended use of money exchange. Users can see each other's transactions by default, and it is here that privacy threats can be identified.

To investigate the nature of user information available on the platform, we accessed a Venmo dataset that was compiled by GitHub user **sa7mon** through Venmo's public API's [1]. The dataset contains more than seven million transactions executed by Venmo users from three date ranges in 2018 and 2019. The researcher released the dataset to highlight that anyone can access information related to your public transactions on the platform.

The goal of this project is to implement an interface layer between the data and Venmo analysts querying the data. This interface will apply an ε-differentially private mechanism to the results of user queries so that analysts will not violate the privacy of the constituents of the database, while preserving the utility of the data. Once the methodology has been vetted, it can be rolled out through the Venmo API for external users.

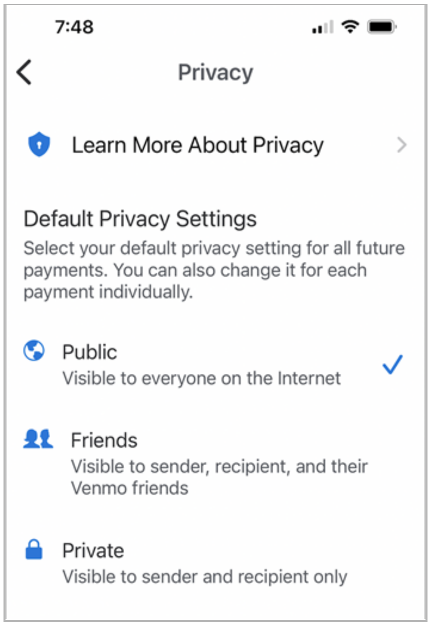
2. Background and Survey of Work

Approximately 70% of Americans use social media of some sort, according to a recent Pew Research Center study [2], despite privacy concerns that cause Americans to have "complicated feelings about social media" [3]. Most people have an expectation of privacy in their electronic communications and social media posts on a private account with a limited audience. However, consumer protections are eroding as the size of datasets increases, when companies collecting the data expose it in unexpected ways, and because of privacy breaches that expose large amounts of personal information.

One scenario with significant privacy engineering implications was the Netflix Prize competition in 2008 in which purportedly anonymized datasets exposed personal information to the public. Researchers Narayanan and Shmatikov [4] were able to de-anonymize the Netflix Prize dataset of 500,000 "anonymous" subscribers' movie ratings. The researchers demonstrated how an *adversary with very limited information* about an individual in the dataset could identify a particular subscriber's record, along with sensitive information such as the subscriber's political beliefs.

Another, more recent research finding was payment app Venmo, which has been scrutinized over privacy concerns resulting from the company's decision to make payment transactions between users' public. The app's *default* privacy setting is for all future payments to be "Visible to everyone on the Internet" as shown in Figure 1. Research by a Mozilla Fellow found that Venmo's resulting public feed reveals troves of personal data and messages shared in Venmo [5, 6]. The data includes payment details that reveal personal and intimate details, as well as possible evidence of drug dealing.

Previous research has examined the dataset of Venmo transaction details collected data amassed via Venmo's public API's. Researcher Dan Gorelick published the results of his analysis of Venmo transactions from their public API [7]. Later, data released by Dan Salmon, the above-mentioned researcher named sa7mon on GitHub, included 7 million actual transaction records [8]. Several articles and security blogs covered Salmon's work as well as Mozilla's research that illustrated Venmo's exposure of private communications [9, 10, 11].



Default Privacy Settings

Select your default privacy setting for all future payments. You can also change it for each payment individually.

Public
Visible to everyone on the Internet ✓

Friends
Visible to sender, recipient, and their Venmo friends

Private
Visible to sender and recipient only

Figure 1: Venmo default privacy settings on Venmo for iOS.

A previous W233 project has researched and analyzed this Venmo dataset, however our study expands on all above-mentioned work. Our study delves deeper into an application of differential privacy to the dataset, applies another dataset (census data that may reveal sensitive information), and takes an approach that sets Venmo's business and Data Science teams as the agents. Finally, we provide recommendations that both protect user privacy and enable the Venmo business with meaningful utility.

Privacy Risks

With the intention of gaining user insights from the data available in the Venmo dataset while also protecting user privacy, we conducted a risk assessment to prioritize the three primary types of risks along with mitigation measures. At its core, the dataset is a sparse transactional log detailing key interaction, such as the user sending money, the user receiving money and the user-defined comment explaining the purposes of the transaction. We chose to use a scenario in which we can enable Venmo's Data Science team and business units to gain valuable insights from the dataset while simultaneously protecting the privacy of all individuals transacting within Venmo and appearing in this dataset.

Attribute Disclosure Threat – The first threat we assessed is attribute disclosure. This refers to the ability to learn about a specific individual's sensitive attribute either alone with the data in the dataset or along with other publicly available sources. While this dataset contains personal identifiable information such as last names and usernames that are often reused across multiple platforms, the data generated from this analysis for Venmo's internal Data Science team would be cleared of these fields. Additional attributes we assessed were race and ethnicity, but as they were added from an external source we did not deem this a risk. We, therefore, did not find k-anonymity to be suitable in this instance as there were no groupings to generalize, and we would like to retain the data for users that had single transactions as we are seeking ways for the Venmo business units to increase user engagement.

Membership Disclosure Risk – The second threat we proposed was membership disclosure, which refers to the ability to determine if an individual has or does not have a specific sensitive attribute based on an adversary's ability to determine if they are within or without a dataset. As our dataset did not contain any confirmed sensitive attributes beyond direct identifiers, we did not find this risk to be significant. All elements of identification are removed and the only remaining potential attributes are based on the user provided comments themselves.

Confidentiality/Identity Disclosure Risk – The final risk we analyzed was identity disclosure risk. As all the data is composed of individual transactions containing a sender, receiver, and a commentary of the transaction, almost all equivalence classes (Sender & Receiver) are unique. The only ways to generalize and build larger equivalence classes are to remove users and only keep comments or vice versa, which removes the bulk of the utility. For this reason, we have opted with using a differential privacy implementation to gain insight on user data while anonymizing all transactions. Through this approach we can retain the utility within the data and still gain a meaningful deal of knowledge about our user population. The implementation is outlined in the following paragraph.

Differential Privacy

Data privacy researchers, data scientists and statisticians have studied the mathematical basis of re-identification in large datasets, notably in oft-cited publications by Dwork [12, 13, 14] and Sweeney [15, 16]. Differential privacy is a methodology that addresses the confidentiality problem relating to individual persons' membership in static databases, such as the Venmo dataset scraped by researchers. The goal of differential privacy is to allow us to learn things about population while protecting individuals, when using an interactive, queryable database.

We have assessed the implementation of both a local differential privacy and global differential privacy implementation as described in a blog post by Shaistha [17]. As all data was previously public and the attributes analyzed were inferred from the text-based user specified notes, on a transactional basis, there is no guarantee of any truthiness or completeness of actual transaction details and user habits. Additionally, for the data that the Venmo team has access to beyond this dataset, the practice of local differential privacy based on randomized response with respect to banking and identifier details was not immediately applicable and appeared better suited for attribute perturbation.

It is for these reasons that we have determined the best implementation for our purposes is the global approach. We further this belief that a global implementation is also suitable as we consider ourselves to be a trustworthy curator and desire the data to be protected from the Data Science group.

This methodology of our implementation defines a database query as a function *f* (.), where *f* introduces no extra randomness and provides the true answer from the database as output. The response to the query is K, where K is not necessarily the true answer to the query. K is how the privacy mechanism is implemented; this could be through perturbation, by adding noise, suppressing some of the rows, or other mechanisms. The query response is controlled by us, the Controller [12, 13, 16].

We use the formal definition of ε-Differential Privacy:

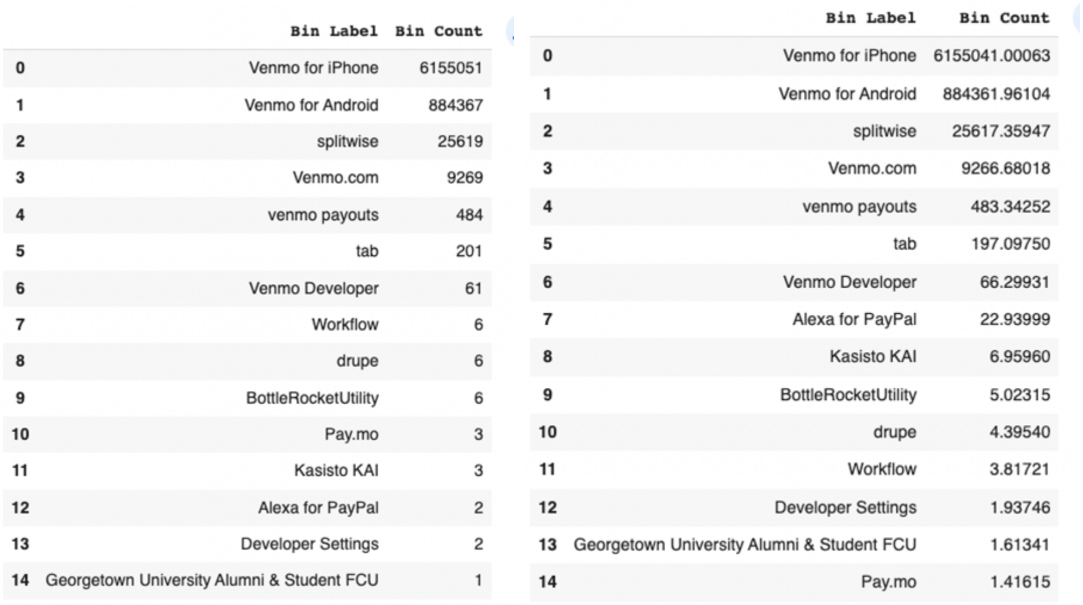
Let K be a privacy mechanism.

> We say that K satisfies **ε-differential privacy** if for all datasets D_1 and D_2 with $|D_1 \Delta D_2| = 1$, and all $S \subset \mathrm{Range}(K)$
>
> $$P[K(D_1) \in S] \leq e^{\varepsilon} \times P[K(D_2) \in S]$$
>
> Where $P[\cdot]$ is taken with respect to the coin tosses of K.

For our purposes we have chosen a privacy guarantee ε of 0.2. We have found that this value offered a good compromise of privacy and utility to our data through protecting the individual transactions and maintaining an approximation of our query results lending full utility. To satisfy 0.2-differential privacy using the global methodology we have generated a mechanism that is used to return summary results (i.e., counts or percentages) of portions of the dataset based on equivalence class groupings, epsilon, and a parameter of interest.

As each query we are demonstrating is considered independent, we can gain the maximum utility by remaining at our upper bound of ε=0.2 rather than needing individual reductions on a query-by-query basis. When the querying function is run, the initial outputs are then passed through the differential privacy mechanism where a random draw from a Laplace distribution centered around 0 is pulled to modify the output. This new result is what is returned to the Data Science team.

An example query from the dataset with and without differential privacy to see the number of users with each type of Venmo application is shown below in Figure 2.



	Bin Label	Bin Count
0	Venmo for iPhone	6155051
1	Venmo for Android	884367
2	splitwise	25619
3	Venmo.com	9269
4	venmo payouts	484
5	tab	201
6	Venmo Developer	61
7	Workflow	6
8	drupe	6
9	BottleRocketUtility	6
10	Pay.mo	3
11	Kasisto KAI	3
12	Alexa for PayPal	2
13	Developer Settings	2
14	Georgetown University Alumni & Student FCU	1

	Bin Label	Bin Count
0	Venmo for iPhone	6155041.00063
1	Venmo for Android	884361.96104
2	splitwise	25617.35947
3	Venmo.com	9266.68018
4	venmo payouts	483.34252
5	tab	197.09750
6	Venmo Developer	66.29931
7	Alexa for PayPal	22.93999
8	Kasisto KAI	6.95960
9	BottleRocketUtility	5.02315
10	drupe	4.39540
11	Workflow	3.81721
12	Developer Settings	1.93746
13	Georgetown University Alumni & Student FCU	1.61341
14	Pay.mo	1.41615

Figure 2: Output of differential privacy mechanism without(left) and with(right) differential privacy applied.

3. Methodology – What did we do and how we did it

Previous studies have shown that individuals' identities can be identified using Venmo's publicly facing dataset. We've reviewed previous work in this space and identified an opportunity to expand on it by identifying relationships between users as well as examining trends in their transaction histories. We focused our research on the most potentially revealing information about Venmo users that could have the greatest effect on the privacy of individual users.

We analyzed the published dataset with 7 million records scraped from public facing Venmo records. There are 404 attributes in the dataset, including personally identifiable information about each user. This information includes the user's email address, phone number, first name, last name, display name, profile picture, payment actor identity, and payment details.

We first selected relevant attributes to make our exploration more manageable. We began looking for the most appropriate quasi-identifiers and sensitive attributes to use for our study. After initial exploratory analysis of the dataset, we narrowed down the following attributes:

```
['id', 'app.name', 'payment.target.user.username',
‘payment.target.user.last_name’,
‘payment.actor.last_name’,'payment.actor.username','payment.note',
'payment.action', 'payment.date_created']
```

The `'payment.note'` attribute, colloquially known as a Venmo caption, generally reveals information about the nature of the transaction, either through words or encoded through emojis–or both. Through examination of this attribute, we were able to identify what users were purchasing as well as who they are likely interacting with in their day to day lives. If user A is paying user B recurrently with the caption “utilities” or using emojis that encode rent, for example, then we can infer that these users are roommates or have a landlord-tenant relationship from these transactions alone. Combining this knowledge with an examination of the users’ other interactions with each other on the platform generally provides confirmation that it is one or the other.

To further trim and analyze the dataset, we compiled a list of emojis that are used as code for a few areas of interest that we deemed would also emphasize the gravity of personal risks associated with having public Venmo transactions. Specifically, we narrowed down our search to emojis that signify the purchase of alcohol, marijuana, and cocaine. The relevant emojis representing alcohol purchases were identified from the Drink subgroup of the Food and Drink emoji group. Those for weed and cocaine are based on emoji slang. For this analysis, we focused only on the emojis and not any additional text in the note field. We then parsed the data set to see if we could expose users that were potentially participating in the buying or selling of any of these substances.

```
 alcohol_emojis = ['🍶', '🍾', '🍷', '🍸', '🍹', '🍺', '🍻', '🥂',
'🥃']

weed_emojis = ['🌿', '🍁', '🥦', '🍀', '🍃', '🪴', '💨', '🚬']

cocaine_emojis = ['❄️', '🎱', '👃']
```

While marijuana is largely legal throughout the United States, cocaine is a controlled substance. We categorized all these activities as potentially illicit activities that could be considered sensitive attributes. We were able to identify users who are (allegedly) participating in the selling and purchasing of marijuana and cocaine by looking for a combination of any of the above in the transaction caption.

For cocaine specifically, the pairings of the emojis were critical. Just parsing for the snowflake emoji, for example, did not return results indicative of what we were searching for. We searched for all permutations of pairings and were able to narrow the scope of our results. From these transactions, we examined other interactions between the payer and receiver to understand the nature of their relationship. Some we identified as likely roommates and friends based on the frequency of their interactions, others were one-off purchases that could potentially be transactions with the dealers of the substances themselves.

Joining the Dataset with Census Data

In addition to the privacy vulnerabilities of the Venmo dataset, an adversary can join the data with another dataset to enrich it even further. For example, given the last names present in the Venmo data, the dataset was merged with Census data [18] to infer the user's race and ethnicity. Per last name, the census data lists the frequency of responses by race and ethnicity. Making a simple assumption that the Venmo user is from the ethnic group with the largest frequency for that last name, we can hypothesize the race and ethnicity of the user.

4. Results – How did it all work out?

In Table 1, the frequency of race and ethnicity, as defined by last name, is presented for both the actual census data as well as for the Venmo data. To be clear, the values for the census are not indicative of the percentage of the total population, but rather the frequency of selecting a name as being mostly from a particular ethnicity or race. 13.2% of all the last names were not in the dataset or in the census data. If we remove the unidentified users, we observe that the remaining user base is overrepresented by Asian Pacific Islanders users.

Ethnic and Racial Group	2010 Census (% of total population)	Venmo Transactions (counts)	Venmo Transactions (% of identified last names)
White	78.695	4,652,214.26643	75.721
Black	3.120	120,741.42330	1.965
Asian Pacific Islander	3.773	670,103.07784	10.907
American Indian and Alaskan Native	0.089	2,039.67207	0.033
Two or more races	0.008	656.26847	0.011
Hispanic	14.313	698,152.83236	11.363

Table 1: Frequency of selecting a name from a particular ethnic or racial group

Table 2 shows the counts and percent of transactions for the three sensitive activities we studied, after differential privacy was applied to the full dataset.

Substance	Count of Transactions	Percent of Transactions
Alcohol	223,414.41893	3.158

Marijuana	32,306.10884	0.457
Cocaine	227.57503	0.003

Table 2: Example of differential privacy applied to sensitive attributes.

This is an example of how the data could be used in a way that will protect the identity of the transactors and receivers with sensitive information such as the use of substances displayed in Table 2.

5. Discussion – What does it all mean?

Through our exploration of data acquired through Venmo public API, we were able to expose not only the likely identities of various users, but also expose sensitive attributes about their personal lives. Upon examination of the API documentation through the lens of differential privacy, we have identified the following changes that we would recommend that Venmo make to preserve the privacy of Venmo users:

1. Sample or perturb the data in the API response so that users remain unidentifiable.
2. When giving anyone access to Venmo's transaction data, do not include any personally identifiable information of the payer or payee such as usernames. Anyone using this data with good intentions will be able to use an anonymized version while protecting the privacy of the transactors.
3. Consider deprecating the API in favor of an interactive and differentially private dataset and provide a list of preset queries that anyone can use to query the data. This preset list of queries will ensure that the output is not small enough to risk re-identification. There is an argument to be made that someone can take the output rows of one of these preset queries and further narrow it down by running their own queries outside of the interactive platform, and thus users should be able to just customize the queries. There is further value in this—Venmo will have some insight into what kinds of data is being queried if the company decides to track these interactions with the platform. Regardless of which approach is chosen to access the data from the interactive data set, if the output of any query is so small that there is risk of re-identification, do not show the results. This shows that Venmo is, in fact, working to protect the privacy of its users to the best of its ability.

Venmo transactions are public by default. Users must go into the settings on the mobile application and manually make the change to limit visibility of their transactions to "Friends only" or "Private". This violates one of the seven principles of the Privacy by Design Framework, "privacy as the default setting". When users sign up for Venmo, there should not be any action needed on their part to protect their privacy. Our recommendation is for Venmo to make the highest level of privacy the default setting, so that new users are not unknowingly making public transactions and exposing themselves to potential privacy threats.

We also recommend that Venmo provide better transparency in its user interface by alerts each time a user is making a public transaction. This alert should require some interaction on the part of the user to dismiss so that they are blocked from completing the transaction until they acknowledge this message This extra layer of friction will ideally defer users from making public transactions.

On the other hand, let's discuss the rights of the consumers, specifically those whose transactions are in this publicly available dataset. We recommend that these individuals change their usernames going forward so that they cannot be traced back to this dataset. Users in California are protected by the California Consumer Privacy Act (CCPA) and have the right to request access to and deletion of their personal data and information from this dataset.

Finally, while we do not believe that the burden of making transactions private should fall on the shoulders of Venmo users, we do feel it is important to inform users that there is a setting in the application that allows users to make all previous and future transactions private with one click.

Further Directions

On Differential Privacy Implementation, future opportunities exist for optimizing epsilon based on the dataset size and final Data Science team use cases such as described in the paper, "How Much Is Enough? Choosing ε for Differential Privacy" [19]. Furthering this, the development of a differentially private interactive summary dataset would provide the same utility but limit the ability to re-query and reconstruct the dataset through a similarity type attack with repetitive but slightly different queries.

There is also opportunity within Venmo to incorporate queries for other aggregate measures such as averaging and median requests, especially with transaction amounts, which are not currently available for us.

A final optimization that may be made is the type of mechanism applied. We have opted to use a Laplace mechanism but there exist additional distributions that may be preferred such as a Gaussian, Poisson or others laid out by Google's Differential Privacy Team in their Differential Privacy Accounting Paper [20].

In conclusion, we believe that the differential privacy approach we applied improves upon Venmo's current approaches and provides better protections when linking other datasets such as ethnicity from census data. Additionally, further opportunities exist, which could incorporate these additional privacy engineering mechanisms that would further improve the privacy/utility tradeoff for this dataset. These improvements would benefit Venmo's business, their Data Science team, and their customers.

6. Members' contributions

Team members cooperatively came up with a research question, considered avenues of analysis, and analyzed and drew conclusions from the dataset.

Ram Ben-David: problem statement, data exploration, analysis, report, and presentation.

Amanjot Samra: data exploration, analysis, research into recommendations, report, and presentation

Jordan Meyer: initial data exploration, differential privacy and epsilon research and implementation, report, and presentation.

Kristy Edwards: differential privacy and epsilon research, survey of work, report, and presentation.

References

[1] Salmon, Dan. Venmo Transaction Dataset. Retrieved from https://github.com/sa7mon/venmo-data. 2019.

[2] Brooke Auxier, Monica Anderson. Social Media Use in 2021, Pew Research Center, 2021.

[3] Lee Rainie. Americans' complicated feelings about social media in an era of privacy concerns, 2018.

[4] Narayanan and Shmatikov. Robust De-anonymization of Large Sparse Datasets, 2008. [Netflix]

[5] Hang Do Thi Duc (Mozilla Fellow). Public By Default - What Venmo (and the Whole World) Knows About You blog and Public by Default website. 2018.

[6] Kaya Yurieff. A researcher studied a year of public Venmo transactions. Here's what she learned, CNN Business. 2018.

[7] Daniel Gorelick. "Hacking" the public Venmo API. 2016.

[8] Salmon, Dan. Venmo Transaction Dataset. Retrieved from https://github.com/sa7mon/venmo-data. 2019.

[9] Zack Whittaker. Millions of Venmo transactions scraped in warning over privacy settings, TechCrunch. 2019.

[10] Venmo scraping - data thread, Slap Magazine. 2019.

[11] Lisa Vaas. Millions of Venmo transactions scraped (again), Sophos Naked Security. 2019.

[12] Cynthia Dwork (July 2006). "Differential Privacy". *33rd International Colloquium on Automata, Languages and Programming, part II* (ICALP 2006). Vol. 4052. Venice, Italy: Springer Verlag, pp. 1–12.

[13] Cynthia Dwork (2008). "Differential Privacy: A Survey of Results". In: *Theory and Applications of Models of Computation*. Ed. by Manindra Agrawal et al. Berlin, Hei- delberg: Springer Berlin Heidelberg, pp. 1–19.

[14] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390. ACM, 2009

[15] L. Sweeney *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

[16] L. Sweeney. *k-anonymity: a model for protecting privacy.* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[17] Fathima, S. (2020, October 3). *Global vs local differential privacy*. Medium. Retrieved March 30, 2022, from https://medium.com/@shaistha24/global-vs-local-differential-privacy-56b45eb22168

[18] Bureau, US Census. "Frequently Occurring Surnames from the 2010 Census." *Census.gov*, 8 Oct. 2021, https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

[19] Lee, Jaewoo, and Chris Clifton. “How Much Is Enough? Choosing ε for Differential Privacy.” *Lecture Notes in Computer Science*, 2011, pp. 325–340., https://doi.org/10.1007/978-3-642-24861-0_22.

[20] Google, Differential Privacy Team. “Differential-Privacy/Distributions.cc at Main · Google/Differential-Privacy.” *GitHub*, Google Differential Privacy Team, 23 Feb. 2022, https://github.com/google/differential-privacy/blob/main/cc/algorithms/distributions.cc.

Appendix

Hari, Dhipin. “A Study on Privacy Preserving Approaches in Online Social Network for Data Publishing.” *Academia.edu*, 4 Apr. 2019, https://www.academia.edu/38717519/A_Study_on_Privacy_Preserving_Approaches_in_Online_Social_Network_for_Data_Publishing

Khanna, Aran. “Have You Left a Money Trail for Anyone to Find on Venmo?” *Business Insider*, Business Insider, 30 Oct. 2015, https://www.businessinsider.com/have-you-left-a-money-trail-on-venmo-2015-10

Kraft, Ben, et al. *Security Research of a Social Payment App*, 14 May 2014, https://courses.csail.mit.edu/6.857/2014/files/13-benkraft-jmoldow-mannes-venmo.pdf

Lin, Jacob. “Venmo's Information Disclosure.” *Medium*, BerkeleyISchool, 18 Oct. 2019, https://medium.com/berkeleyischool/venmos-information-disclosure-5d05174982ce

Matthew Joseph, Aaron Roth, Jonathan Ullman, Bo Waggoner. Local Differential Privacy for Evolving Data. *32nd Conference on Neural Information Processing Systems (NeurIPS*), 2018.

Michael Hawes, et al. Differential Privacy and the 2020 Decennial Census. NCSL Webinar, 2020.