Critique of "*Crowd-Blending Privacy*"

Callum Mann cm13558

January 2017

1 Introduction

Crowd-Blending Privacy [6] is a paper by Johannes Gehrke, Michael Hay, Edward Lui and Rafael Pass that was published to CRYPTO 2012. The paper introduces a new notion of privacy known as *k-crowd blending* privacy, which extends the existing notion of *differential privacy* in two new ways. Differential privacy has become a framework for introducing a formal level of privacy in primarily statistical databases (although other datasets are possible), in an attempt to prevent deanonymisation techniques given that the dataset is released into the public domain under the method of the framework. Simply stated, an adversary should not be able deduce more information about an individual than they can from all individuals in the dataset or from an independent dataset. This type of privacy is increasingly important in the current technological climate as current trends in machine learning require large amount of data/sets, so it should be guaranteed that individuals in said data are assured of their privacy. For example, previous attempts at anonymising data failed when in 2006 graduate students from University of Austin, Texas deanonymised individuals from a dataset that Netflix released by using previously released IMDB data as a source of background knowledge [1]. Netflix had released their data as a challenge to anyone who could improve their recommendation system. In contexts where data between individuals may be highly correlated such as in social network data, the authors also build upon a definition from their previous work, known as *zero-knowledge privacy*.

2 Definitions of privacy

To effectively distinguish the privacy definitions presented by Gehrke et al. to other, existing definitions, it is first useful to discuss a few of them

before evaluating the papers main result.

Differential privacy was first defined by Dwork et al. in 2006, in the paper "Calibrating Noise to Sensitivity in Private Data Analysis" [4]. Differential privacy is roughly defined as follows: A mechanism *San* is said to be differentially private if for every pair of databases differing in only one value (individual), the output distribution of *San* is indistinguishable to some negligible factor ϵ. Here *San* is shorthand for *sanitizer*, i.e the mechanism, or more precisely the algorithm, that is applied to the database before it is released publicly (alternatively, the mechanism may be a queryable interface).

Differential privacy however, as aforementioned, may not provide sufficient privacy for highly correlated data, so the authors introduce zero-knowledge privacy, which is a stronger definition of privacy[7]. Essentially, there are adversaries and simulators. Adversaries are able to query the database mechanism with auxiliary information, to deduce information about from individual i. Simulators use aggregate functions (sums, averages) that describe the remaining individuals. If the adversary can learn no more about i than the aggregate information (no more than the simulator), then the mechanism has zero-knowledge privacy.

Finally, the definition of crowd-blending privacy can be stated as follows: An individual i blends with another individual with respect to a mechanism *San*, if the two individuals are indistiguishable by *San*. This proposition is extended toward a group of k individuals; if there exist $k-1$ other individuals in the database that blend with i with respect to *San*, then *San* is *crowd-blending private*. In this sense, blending is meant to describe the act of replacing one individuals data with another, and being private, as before, means the output distribution of the mechanism remains the same. The authors emphasize that they attempt to study the "weakest possible" definition of privacy, that is still useful, so that when combined with other methods they are able to construct stronger privacy notions. Indeed, given that the adversary knows some property of the $k-1$ individuals in i's crowd, then certainly the adversary can say something about i. This is not compatible with differential privacy, although the authors argue that since the property is shared by the crowd, it is therefore "non sensitive", and furthermore this looser definition of privacy provides more utility. However, it is known that the crowd-blending sanitizers are not closed under composition, meaning that if an individuals crowds are disjoint under composed sanitisers, then there is little privacy atall. Although, the authors claim this fault is acceptable,

as long as a robust sampling mechanism is used before applying the crowd-blending algorithm. Then one can achieve zero-knowledge and differential privacy.

Throughout the paper, the hierarchy of privacy is proved

$$\textbf{zero-knowledge} \Rightarrow \textbf{differential} \Rightarrow \textbf{crowd-blending}$$

3 The Main Result

The main theorem states that

$$\textbf{sampling} + \textbf{crowd-blending} \Rightarrow \textbf{zero-knowledge}$$

In the sampling stage, each individual is selected from the population with some probability. The mechanism is simply the crowd-blending mechanism applied on the sampled data. Then the mechanism is zero-knowledge private with respect to the aggregate function. The aggregation function is simply the sampling with the aggregate computation applied to the sampled data.

4 Evaluation of the Concepts

The crowd-blending element of the paper itself is largely only an ingredient to the wider usefulness of zero-knowledge privacy, as alone it does not satisfy even differential privacy. The benefit of differential privacy is mainly that, it is quite simple to state, and therefore to easier to design a system that is differentially private than one that is crowd-blending. This may be why differential privacy has become the standard in database sanitation, and why crowd-blending might not be adopted so well. Despite this, it is intuitive that the crowd-blending mechanism may be better at securing the privacy of social network data for instance, as k individuals sharing a common property can blend together.

In terms of cryptologic elegance, unfortunately, it can be said that crowd-blending mechanisms do not really possess any. As previously stated, the mechanisms do not compose and therefore are less attractive to a well read cryptologist. This is because often lack of properties such as composition lead to problems down the line in proofs or other reasoning, so it is generally distasteful to allow it. Furthermore, the very concept itself seems to bend some of the rules of privacy, such as properties being "non sensitive" when shared in a group of individuals. This is likely a topic debated intensely by privacy groups and again, limits the concept from

being widely adopted.

Nevertheless, the paper has around 40 citations on Google Scholar at the time of writing this, so it has certainly attracted some attention. For example, in a paper regarding anonymous protocols, the authors suggest that extending crowd-blending to distributed systems would allow k-anonymity in anonymous communication [2]. But, no actual extension is completed, and in general, most of these citations are only part of a wider analysis of differential derivatives within the paper that cites it (additionally, [5]), and it does not appear that any of the work that cites crowd-blending uses it in any further construction.

In a survey of methods in differential privacy [3], existing methods in the field carry exponential time, which Gehrke et al. highlight in their paper and suggest that crowd-blending is a better alternative, although no proof or intuition on this is given. In the survey, most of the mechanisms discussed are targeted toward private learning, that is optimally opening the mechanism toward learning from the dataset. There is no analysis of how well a crowd-blending mechanism would lend itself to private learning, which is a shame since the foremost reason for releasing data is to learn from it, though it would be wrong to suggest this omission detracts from the quality of the paper.

Despite these shortcomings, we cannot ignore the original database mechanisms that are achievable with crowd-blending. For example, the crowd-blending allows one to privately release histogram data with *no noise*. Noise is often added to differential methods to satisfy the privacy requirements, however crowd-blending has achieved better utility as it requires no noise for sufficiently large blending groups. More recent work in the vein of differential methods has targeted graph data[8], which continues to show that there is great diversity in produced research in this area, so crowd-blending is not in itself any more tangential. The field is still young, so the work cannot be expected to be practically useful as of yet, however the incredible demand for data cannot be ignored, and sooner or later efficient methods will have to be invented.

5 Conclusion

Crowd-blending privacy provides some new ideas into the pool of differential privacy derivatives. It has been proved that with good sampling methods it can be zero-knowledge private while still retaining flexibility. It has gained some recognition among further work primarily discussing the concepts explained in the paper, rather than basing new products or theorems on the result. Therefore, it can be said that the paper brought forward some useful ideas, though not as widely accepted as differential privacy, and not as easy to adopt. To conclude, Crowd-Blending Privacy presents itself well within the current scene surrounding differential privacy.

References

[1] Vitaly Shmatikov Arvind Narayanan. How To Break Anonymity of the Netflix Prize Dataset. https://arxiv.org/abs/cs/0610105, 2007.

[2] Michael Backes, Aniket Kate, Praveen Manoharan, Sebastian Meiser, and Esfandiar Mohammadi. Anoa: A framework for analyzing anonymous communication protocols. In *2013 IEEE 26th Computer Security Foundations Symposium*, pages 163–178. IEEE, 2013.

[3] Cynthia Dwork, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li. *Differential Privacy: A Survey of Results*, pages 1–19. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[4] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *TCC 2006: 3rd Theory of Cryptography Conference*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284, New York, NY, USA, March 4–7, 2006. Springer, Heidelberg, Germany.

[5] R Bassily et al. Coupled-Worlds Privacy: Exploiting Adversarial Uncertainty in Statistical Data Privacy. https://arxiv.org/abs/cs/0610105, 2013.

[6] Johannes Gehrke, Michael Hay, Edward Lui, and Rafael Pass. Crowd-blending privacy. In Reihaneh Safavi-Naini and Ran Canetti, editors, *Advances in Cryptology – CRYPTO 2012*, volume 7417 of *Lecture Notes in Computer Science*, pages 479–496, Santa Barbara, CA, USA, August 19–23, 2012. Springer, Heidelberg, Germany.

[7] Johannes Gehrke, Edward Lui, and Rafael Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In Yuval Ishai, editor, *TCC 2011: 8th Theory of Cryptography Conference*, volume 6597 of *Lecture Notes in Computer Science*, pages 432–449, Providence, RI, USA, March 28–30, 2011. Springer, Heidelberg, Germany.

[8] Binh P. Nguyen, Hoa Ngo, Jihun Kim, and Jong Kim. Publishing graph data with subgraph differential privacy. In Howon Kim and Dooho Choi, editors, *WISA 15: 16th International Workshop on Information Security Applications*, volume 9503 of *Lecture Notes in Computer Science*, pages 134–145, Jeju Island, Korea, August 20–22, 2016. Springer, Heidelberg, Germany.