Blogpost: When the signal is in the noise: Exploiting Aircloak's Diffix anonymization mechanism

Andrea Gadotti, Florimond Houssiau, Luc Rocher, Yves-Alexandre de Montjoye

Transcript of the blogpost at https://cpg.doc.ic.ac.uk/blog/aircloak-diffix-signal-is-in-the-noise/

Abstract – *Information about us is being constantly collected, through our phones and the services we use online. This data is hugely valuable but also highly personal, and often sensitive. This raises a crucial question: can we use this data without disclosing people's private information? We studied Diffix [1], a system developed and commercialized by Aircloak to anonymise data by adding noise to SQL queries sent by analysts. In a manuscript [2] we just published on arXiv, we show that Diffix is vulnerable to a noise-exploitation attack. In short, our attack uses the noise added by Diffix to infer people's private information with high accuracy. We share Diffix's creators opinion [3] that it is time to take a fresh look at building practical anonymization systems. However, as we increasingly rely on security mechanisms to protect privacy, we need to learn from the security community: secure systems have to be fully open and part of a larger layered security approach. Privacy is hard, it is time to admit that we won't find a silver bullet and start engineering systems.*

April 24, 2018

In the last decade, the amount of personal data being collected and used has exploded. With half of the world population soon [4] having access to the Internet and with the Internet of Things [5] becoming a reality, this is unlikely to stop anytime soon. While data has immense potential for economic development and scientific advancements, its collection and use raises legitimate privacy concerns. More than 80% of U.S. citizens are concerned [6] about sharing personal information online. Data, directly [7] or indirectly [8], contains sensitive information that could be used against individuals.

To prevent this and allow organizations to collected and use data while preserving people's privacy, data is often anonymized. The idea behind this is that if the data is anonymous, if one can't know that user *MjJ17torTC* is you, the data can't be used against you. Data anonymization, also called de-identification, is a two steps process: first direct identifiers such as name, social security numbers or email addresses are removed and then noise is added to the dataset.

However, a large body of research has shown that pseudonymized and even anonymized data can often be easily linked back to you [9-13], **re-identified**. This increasing amount of evidence has led President (Obama)'s Council of Advisors on Science and Technology (PCAST) to conclude in 2013 that data

anonymization *"is not robust against near-term future re-identification methods"* and that they *"do not see it as a useful basis for policy"* [14].

Privacy researchers have therefore been increasingly interested in the potential of **question-and-answer** (or query-based) systems as a way to use data without disclosing sensitive information.

The idea of question-and-answer systems is simple: rather than sharing the raw anonymized data with analysts (a model called release-and-forget), data holders could allow analysts to remotely ask questions from the data and only get aggregated answers back. For instance, a question could be: " *What is the average income of men older than 50?"*

However, avoiding direct access to the data is, alone, not sufficient to ensure that privacy is preserved. Without additional security measures, query-based systems are susceptible to a wide range of attacks. For instance, it does not prevent the analyst from accessing private information by asking the right question, such as *how many users named Edward Snowden were diagnosed with cancer in 2017*, or a combination of right questions. Attacks relying on multiple, seemingly innocuous, queries such as averaging attacks [15] or intersection attacks [16] have been developed over the years.

Diffix

German startup Aircloak, along with researchers from the Max Planck Institute for Software Systems, developed and commercialized a system called **Diffix** [1] to protect SQL databases from rogue analysts. Diffix relies on a novel, patented, and proprietary approach, called *sticky noise*, which adds noise to each answer to a query with the noise being based on the query.

Aircloak says their approach allows analysts to ask an **infinite number of queries**, with a rich query syntax and minimal noise, all the while strictly preserving people's privacy. According to them [17], Diffix (1) falls outside the scope of GDPR regulations, (2) has been guaranteed to deliver GDPR-level anonymity by the *French Data Protection commission (CNIL)* and (3) certified by TÜViT as fulfilling *"all requirements for data collection and anonymized reporting"*.

In a manuscript [2] we just published on arXiv, **we show that Diffix is vulnerable to a new attack we developed**, which we call *noise-exploitation* attack. We show in the paper how an attacker can successfully learn the sensitive attribute (e.g. HIV status) of someone in the dataset protected by Diffix knowing a set of attributes that uniquely identify them in the dataset (say, age, ZIP code, and education level). Our main contribution is a novel (to the best of our knowledge) technique to **exploit the noise added by Diffix as a "signal"** to learn the target's private information.

In order to understand how our attack works, let us briefly explain the "sticky noise" used by Diffix. When Diffix processes a new query, it computes the accurate result of the query on the dataset and then adds several layers of noise to the output. There are two types of noise layers: *static* noise layers, that depend on the query expression (i.e. the question asked), and *dynamic* noise layers, that depends on the set of users selected by the query (the *query set*). To prevent simple attacks, Diffix also does not report results whose number of users is below a certain threshold (randomly selected according to a normal distribution $N(4, \frac{1}{2})$), they call this bucket suppression. For a query Q, Diffix's output is $\mathsf{output} = \mathsf{true\_value} + \mathsf{static}_Q + \mathsf{dynamic}_Q$, where static_Q and $\mathsf{dynamic}_Q$ denote the sum of all static and dynamic noise layers respectively.

Our noise-exploitation attack relies on three steps to circumvent Diffix's protection (all the details and numerical simulations are available in the manuscript [2]:

1. We can design queries that are similar enough that they will share part of their static noise. This allows us to cancel out some of Diffix's noise.
2. As Diffix's noise depends on the query set, **the noise itself leaks information about the query set**. This is the core of our attack: analyzing the remaining Diffix's noise we develop a statistical test [18] to learn information about the data Diffix is protecting.
3. We exploit logical equivalence between queries to circumvent some of the "stickiness" of the noise by repeating (almost) the same query. This allows us to obtain fairly independent noise samples which, when added to our statistical method, increase the power of our attack.

Combining these steps, we developed a powerful attack allowing an attacker to potentially infer a user's attribute with high accuracy with very limited auxiliary information. For instance, we showed that knowing **only four attributes** could allow an attacker to learn a user's private information with **99% accuracy** making this an important vulnerability. **While further research is needed, we do not see a direct way for Aircloak to prevent this attack.**

Moving Forward

The privacy of question-and-answer systems has been heavily researched under the theoretical framework of differential privacy [19] introduced in 2006. While differential privacy proposes a solid and theoretically appealing mathematical foundation for the protection of privacy, it has so far mostly failed to offer a practical solution to the protection of modern datasets. While a few recent expections exist [20-22], many (including Diffix's creator [3]) believe that it will not be a viable solution for general use cases such as the one tackled by Diffix.

It is therefore time to start investigating new approaches to the privacy of query-based systems including *privacy-through-security* approaches. While theoretical

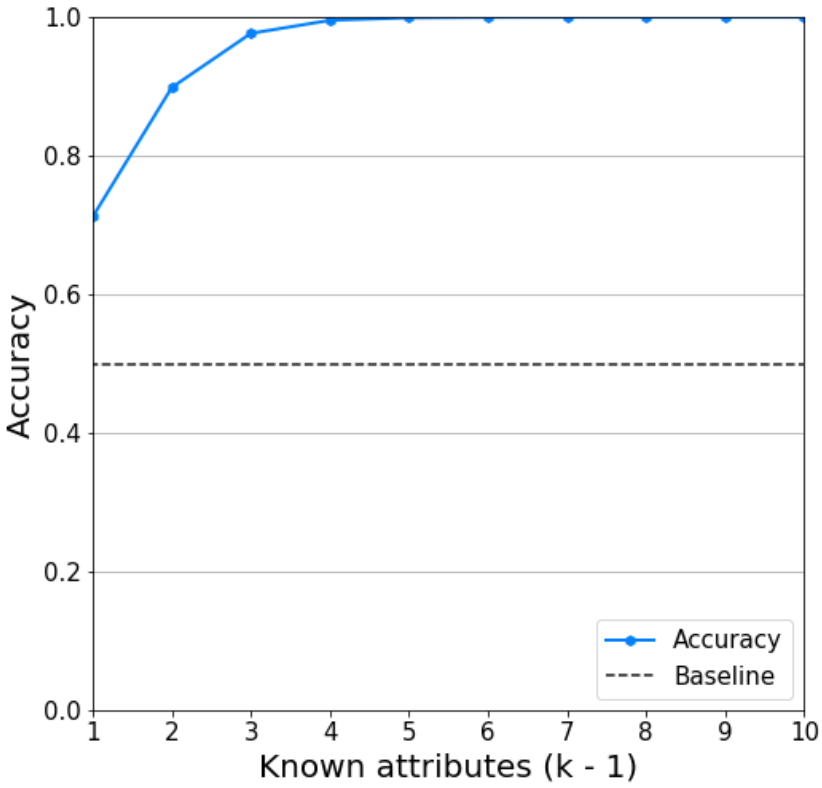

Figure 1: Accuracy of our noise-exploitation attack on Diffix as function of the number of attributes known to the attacker.

frameworks with **provable guarantees** such as differential privacy are likely to continue playing an important role in privacy research, they cannot be the only—or maybe even the main—focus of the community. Moving forward, *penetrate-and-patch* adversarial approaches are likely to become a crucial part of building practical privacy-preserving systems even when differentially private. Even provable guarantees are not enough to make a system safe. Implementation issues, hypotheses, and design choices may always introduce vulnerabilities.

This adversarial engineering approach is powerful but hard as it requires us to change both our expectations of privacy-preserving mechanisms and our way of thinking about privacy. We need to accept that **no system is perfect**. There will be attacks, and some of them will succeed. We need to prepare for this and learn from best practices in security: ensuring that several layers of security exist, to not have all the data in one place (what Jean-Pierre Hubaux calls the Fort Knox approach), etc. We also need standards and systems to be **completely transparent** and **open**. Building secure systems requires anyone to be able to review the code without technical or legal barriers, propose solutions and build upon existing work.

Fresh approaches to protecting privacy are essential moving forward but we need the right environment around them.

UPDATE — May 17, 2018

Paul Francis —director at the MPI-SWS and co-founder of Aircloak— published on April 27 a blogpost [23] and Felix Bauer —CEO of Aircloak— a statement [24] regarding our attack. While both acknowledge the vulnerability we disclosed, they claim that "the conditions under which it could work are so rare as to be practically non-existent". Their claim is based on an empirical analysis of open-data datasets on which the attack would only sometimes succeed, depending on properties of the dataset. We are currently evaluating their analysis and running our own experiments.

Paul Francis also stated on Twitter that this is a "immediate vulnerability disclosure" [25] implying that we did not contact them before publishing the manuscript and blogpost. This is not accurate. We submitted the manuscript to Arxiv (which receives and publishes articles in bulk the day after), the goal being to protect us from potential cease and desist letters or other threats, and e-mailed Paul Francis and Aircloak right after. Paul Francis answered our e-mail a couple of hours later and did not ask or suggest we should delay the public disclosure. We have, as of today, not received a response from Aircloak.

UPDATE — Aug 15, 2019

Our paper *When the Signal is in the Noise: Exploiting Diffix's Sticky Noise* has been accepted to USENIX Security '19! The paper [26] is based on our original

attack published on April 18, 2018 and then extended (after notifying Aircloak) on July 13, 2018 with now two noise-exploitation attacks: a differential attack and a cloning attack.

The cloning attack exploits the same noise addition vulnerability of the differential attack, but instead of using a likelihood-ratio test, it relies on dummy conditions that affect the output of queries conditionally to the value of the private attribute. Importantly, this attack relies on weaker assumptions and automatically validates them with high accuracy.

Using this attack on four real-world datasets, we show that we can infer private attributes of at least 93% of all users with an accuracy ranging from 93% to 97%, issuing only a median of 300 queries per user. We show how to optimize this attack, targeting 55% of the users and achieving 92% accuracy, using a maximum of only 32 queries per user.

Aircloak proposed a fix to our original attack in an article on their blog. Based on the high-level description in the article (no technical document is available yet), it seems that this is a mitigation that may prevent the differential and cloning attacks. However, it potentially also opens up new vulnerabilities, as it does not directly address the risk of data-dependent noise, and instead introduces a new data-dependent measure.

Diffix's approach—the use of sticky noise and a security-based solutions to protect personal data—is very interesting and promising. We agree with their pragmatic approach to enable practical data analysis while minimizing the risk of individual data leakage. However, we disagree with its characterization as a "silver bullet" that, alone, is sufficient to rule out practical attacks. Privacy in query-based systems will require a layered approach that combines anonymization mechanisms with defense-in-depth measures such as access control, intrusion detection, and above all auditability (which we discuss in the paper).

We have updated our manuscript to ArXiv [2] to reflect the one accepted to USENIX. The version on ArXiv contains some additional discussion and experiments. We have also published the code for our attacks and experiments, available here: https://cpg.doc.ic.ac.uk/signal-in-the-noise.

Bibliography

[1] Francis, P., Probst-Eide, S., Obrok, P., Berneanu, C., Juric, S. and Munz, R., 2018. Extended Diffix. *arXiv preprint arXiv:1806.02075.*

[2] Gadotti, A., Houssiau, F., Rocher, L. and de Montjoye, Y.A., 2018. When the signal is in the noise: The limits of Di x's sticky noise. *arXiv preprint arXiv:1804.06752.*

[3] P. Francis (2017), 'Differential privacy at the end of the rainbow', IAPP, September 26

[4] Statista, 'Worldwide internet user penetration from 2014 to 2021'

[5] Wikipedia, 'Internet of Things'

[6] Rose, J., Barton, C., Souza, R. and Platt, J., 2014. Data privacy by the numbers. *Boston Consulting Group.*

[7] J. Hoeksma (2014), ['The NHS's care.data scheme: what are the risks to privacy?'] (https://doi.org/10.1136/bmj.g1547), *the bmj*, February 17

[8] P. Tucker (2014), 'How the NSA Can Use Metadata to Predict Your Personality', *Defense One*, March 28

[9] Sweeney, L., 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics, 25*(2-3), pp.98-110.

[10] Narayanan, A. and Shmatikov, V., 2008, May. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111-125). IEEE.

[11] De Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: The privacy bounds of human mobility. Scientific reports. 2013 Mar 25;3:1376.

[12] De Montjoye, Y.A., Radaelli, L. and Singh, V.K., 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221), pp.536-539.

[13] P. Cowan (2016), 'Health pulls Medicare dataset after breach of doctor details', *itnews*, September 29

[14] Holdren, J.P. and Lander, E.S., 2014. Big data and privacy: a technological perspective. *President's Council of Advisors on Science and Technology.*

[15] Denning, D.E., 1980. Secure statistical databases with random sample queries. *ACM Transactions on Database Systems (TODS)*, 5(3), pp.291-315.

[16] Denning, D.E. and Denning, P.J., 1979. The tracker: A threat to statistical database security. *ACM Transactions on Database Systems (TODS)*, 4(1), pp.76-96.

[17] F. Bauer (2017), 'Building Trust', *Aircloak blog*, July 5

[18] Wikipedia, 'Likelihood Ratio Test'

[19] Dwork, C., 2008, April. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation* (pp. 1-19). Springer, Berlin, Heidelberg.

[20] Erlingsson, Ú., Pihur, V. and Korolova, A., 2014, November. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of*

the 2014 ACM SIGSAC conference on computer and communications security (pp. 1054-1067).

[21] Apple, D., 2017. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8).

[22] Uber Security (2017), 'Uber Releases Open Source Project for Differential Privacy', Medium, July 13

[23] P. Francis (2018), 'Report on the vulnerability in Diffix-Birch announced by Imperial College London and CU Louvain', Aircloak blog, April 27

[24] F. Bauer (2018), 'Statement regarding the attack on Diffix-Birch by Imperial College Scientists', Aircloak blog, May 4

[25] Twitter, https://twitter.com/anonymity_R_Us/status/989415748597673984

[26] Gadotti, A., Houssiau, F., Rocher, L., Livshits, B. and De Montjoye, Y.A., 2019. When the signal is in the noise: exploiting diffix's sticky noise. In *28th {USENIX} Security Symposium ({USENIX} Security 19)* (pp. 1081-1098).