Analytics without tears

or is there a way for data to be anonymised and yet still useful?

In this article, we discuss the new requirements for standards for policy and mechanism to retain privacy when analyzing users' data. More and more information is gathered about all of us, and used for a variety of reasonable commercial goals -- recommendations, targetted advertising, optimising product reliability or service delivery: the list goes on and on. However, the risks of leakage or misuse also grow. Recent years have seen the development of a number of tools and techniques limit these risks, ranging from improved security for processing systems, through to control over what is disclosed in the results. Most of these tools and techniques will require agreements on when and how they are used and how they inter-operate.

HISTORICAL AND TECHNICAL CONTEXT

We have always kept data about ourselves – maybe household accounts to divvy up the food bill amongst students in shared flat each month, or maybe our baby’s weight and height. Its often easier to let someone else look after that data, a bank or our doctor for example, since they can use it for our benefit, and keep it safe. They then store many peoples’ records, and need a way to find our particular one, via some primary key, an identifier that uniquely fishes out our information - perhaps a mix of our name, birthday and postcode.

The two-sided market of Cloud Analytics emerged almost accidentally, initially from click-through associated with users’ response to search results, and then adopted by many other services, whether web mail or social media. The perception of the user is a free service (storage and tools for photos, video, social media, etc.) with a high-level of personalisation. The value to the provider is untrammelled access to the users’ data over space and time, allowing upfront income from the ability to run recommenders and targeted adverts, to background market research about who is interested in what information, goods and services, when and where. User’s data might be valuable in many contexts, specially when aggregated from several sources. This created a market for data, making suitable for the Internet a point made in the 70s regarding television [1]: “if you are not paying for the product, you are the product”.

In this context, we've experienced a shift in how decision-making processes are approached by enterprises and governments: crucial decisions in areas as diverse as policy making, medicine, law enforcement, banking, and the work place are informed to a great extent by data analyses. For example, decisions related to which advertisements and promotions we see online, which of our incoming emails are discarded, what type of TV series are produced, what conditions are attached to our insurance, or what new drugs are developed, are informed to a great extent by the analysis of electronic data.

This trend is far from stabilising. As digital surveillance grows apace, the advent of personal data gleaned not just from social media and online services but from sensors in smart homes, cars, cities, and health devices, becomes more and more intrusive. Something has to give.

There are both technical and socio-economic reasons why this has to change, and this has been recognised amongst regulators and in industry. Privacy failures risk personal and corporate wealth and safety. Theft of credit card data, identity and trade secrets is a real and present danger, increasing with the extended "attack surface" presented in the surveillance society. The volume of detail available admits of inference about people and institutes in ways not recognised or necessarily intended. In some cases, this even can lead to threats to personal safety. As a result, new law, and new technology has been proposed and enacted, which may go some way towards alleviating this. However, the road-map is quite complex and involves choices, as well as agreements between parties. Some of the technical choices will have implications for standards, which may in turn reflect back on how regulation and legislation will evolve. One thing appears clear, that organisations are keen to retain the value of all that big data now being gathered and analysed, whether for entertainment, health, security, or profit. Hence there is a need for careful harmonisation between the new regulation, and the new technology that can support the old value chains.

The General Data Protection Regulation (GDPR) and the ongoing e-privacy regulation effort are significant steps in regulating the protection of sensitive information by placing obligations on data controllers and data processors, as well as specifying user's rights. However, no specific algorithms are mentioned, and hence we are far from effective standardisation guidelines. This is justified, as the technology is not quite there, and more research is needed both from a theoretical and applied perspective. However, its potential has been recognised both in industry and government. The recent report by the US Commission on Evidence-based Policymaking [2] describes differential privacy and multi-party computation as emerging technologies and states that "New privacy-protective techniques [...] may allow individuals in the Federal evidence-building community to combine data and conduct analyses without directly accessing or storing information".

In general, similarly to general security, privacy has proven to be a slippery concept, that requires a robust, mathematically rigorous approach. Nevertheless, very promising advances have been made in the last decade, both from a practical and a theoretical perspective. In this scenario, privacy-preserving analytics has emerged as a very active research topic simultaneously in several fields such as machine learning, databases, cryptography, hardware systems, and statistics. The main challenges include several applications currently being simultaneously pursued within research communities in the above fields, such as finding secure ways of *providing public access to private datasets*, securely *decentralising services that rely on private data* from individuals, *enabling joint analyses* on private data held by several organisations, and *securely outsourcing computations* on private data.

There are several alternative directions that this will evolve in the future.

ASPECTS OF PRIVACY-PRESERVING ANALYTICS: COMPUTATION AND DISCLOSURE.

The general goal of research into privacy-preserving data analysis is to develop techniques that get the best utility out of a dataset without violating the privacy of the individuals represented in it. However, there are several interpretations of what we may mean by privacy in this context.

First of all one has to realise that if you belong to a certain population, and an analysis on a that population is disclosed, then your privacy has been breached and there is nothing you can do – or could have done – about it. For example, assume that a predictive model about mobility in London is made public. Let's say that that model is able to accurately predict the location of London tube users, given some of their characteristics. Regardless of whether that model was trained with their data or not, the privacy of all London tube users is breached to some extend. This point might sound obvious, but it is important: technical advances in general do not solve all ethical issues, and privacy is not an exception to that. Every data analysis has some ethical issues regarding privacy associated with it, which must be approached as such.

However, there are many crucial privacy issues in data analysis that technology can help to overcome. Two aspects for which we have in principle satisfactory technical solutions are privacy of stored data, i.e., encryption of data *at rest* (on disk), and privacy of data as it is being transmitted, i.e., encryption of data *in transit*. Current basic research challenges have to do with preserving privacy even during processing, and generally correspond to two orthogonal but tightly-related aspects: privacy-preserving computation and privacy-preserving disclosure.

Privacy-preserving computing: The result and nothing but the result!

Let's say you upload your data encrypted to the cloud, but still allowing for some concrete computations to be performed on it by service providers, such as training machine learning models, or selecting ads tailored for you. This would certainly keep the service providers happy, while protecting your private data from data breaches.

So now how do we execute software on machines owned and maintained by an untrusted party? Or, more generally, how do we compute on private data held by mutually untrusted parties? There are several emerging techniques to do this, which could be combined in principle, and come from the areas of hardware security and cryptography.

Secure enclaves.

The idea behind secure enclaves is based on new technology (not so new on the iPhone but new to servers) called a Trusted Execution Environment [3]. Such trusted hardware provides a secure container into which the secure cloud user can upload encrypted private data, securely decrypt it, and compute on it. Both the decryption and the computation are run in a processor which, in principle, not even its owner can break into. The result is again securely transmitted to the user, together with a proof that it is indeed the result of the intended computation.

It is important to remark that this approach relies on trusted hardware, which is in general hard to patch if vulnerabilities are found. Moreover, there are some limitation to its security guarantees, as it does not protect against cache-timing and physical attacks, as well as limitations in terms of scalability as the amount of available RAM within a container is often limited.

Examples of this technology are Intel's SGX and ARM TrustZone, which are evolving and being adopted quickly. A recent instance of such adoption are the Azure Confidential Computing capabilities.

Homomorphic encryption.

An encryption scheme is said to be homomorphic with respect to a given operation if one can perform that operation on the encrypted data by just manipulating the corresponding ciphertext. For example, if an encryption scheme is homomorphic with respect to addition, two encryptions of arbitrary values, say 23 and 19, can be combined – without prior decryption – to produce the encryption of their sum. Asymmetric key encryption schemes homomorphic with respect to either addition, e.g., Paillier, or multiplication, e.g. ElGamal, have been known for a while, but it wasn't until 2009 that Gentry described the first *fully* homomorphic encryption scheme [4], namely

a scheme homomorphic with respect to both addition and multiplication. Note that, if we operate on a binary domain, i.e., mod 2, addition and multiplication is all one needs to do anything a modern processor can do. This enables secure outsourced computation relying solely on encryption, as opposed to the secure enclave approach, as a user can encrypt all their data and share it with the cloud encrypted, together with the public key of the encryption and a description of the computation. Then the cloud provider can compute on it in encrypted form – as if it was computing blindfolded – and return the encrypted result.

Fully homomorphic encryption is a remarkable breakthrough, as before Gentry's contribution, it was not even clear whether such kind of encryption did even exist. However, although several alternative improved schemes have been proposed since Gentry's, homomorphic encryption is currently far from scaling to the secure cloud computing application, and in particular data analysis tasks involving massive input sizes. Nevertheless, several homomorphic encryption libraries are available, and a limited notion of fully homomorphic encryption supporting a fixed number of nested multiplications called somewhat homomorphic encryption might be enough for some data science applications.

Multi-Party Computation (MPC).

Another alternative is to use Secure Multiparty Computation (MPC), and area of cryptography kicked off by Andrew Yao [5] in the 80s. There are a number of protocols, including the lovely Yao's garbled circuits, that revolve around the idea of sharing secrets without actually giving them away, and then computing on them by transforming their shares, and still keeping them secret. An example is the way to find out who is the richest person in the room, without revealing how much each person actually possesses. These are hard to reason about for the lay person, but can be verified in design, and probably therefore are a promising additional technique. Moreover, MPC techniques are quite efficient and, due to a sequence of theoretical and engineering breakthroughs, have become of practical interest, with many available libraries and applications, and even commercial products.

MPC technologies allow to move away from the trusted aggregator model for analysis on distributed data. Instead of moving all the data to a single server (where it might be leaked), we can leave data in peoples' devices (smart homes, smart TVs, cars, IoT devices, tablets, etc) and distribute the programmes that do the analytics in a privacy-preserving way. This then moves the results (e.g., market segment statistics) to businesses that wish to exploit them without ever moving the raw personal data anywhere at all. Hence, the parties interested in an aggregated model learn such model, *and nothing but the model*. In principle this permits reversing the business models' direction of value - the subject (user) can now charge for their data! In the distributed approach, since there is no central data center/cloud anymore, there's no need to cover its cost, so the change adds up.

The caveat here is that while MPC techniques allows to keep the data in the parties owning it, such parties must get involved in the computation hence incurring some computation and communication cost. This is in contrast with secure cloud approaches, where the encrypted data only has to be uploaded once. This motivates architectures that include sets of non-colluding untrusted parties that are used to simulate a secure cloud. It is important to remark that MPC techniques provide high-assurance cryptographic guarantees.

Edge Computing.

There are performance advantages to edge computing in some new scenarios, especially in the smart home and Internet-of-Things use cases. At the least, we can remove the burden of sending large amounts of detailed data from very large numbers of edge devices into the cloud. Instead, we retain the data in local hubs (e.g., smart home hubs), and send analytics software to execute there, rather than on the central cloud. This is recognised in the IoT hub work by Microsoft and in several other IoT platforms. We still need to retain all the same approaches to supporting privacy concerning the data, but the very distributedness of the data and computation reduces the risk of a mass-leak of information, as the attack-surface of the whole system is now fragmented. Techniques for decentralised analytics work quite well and can adopt many of the techniques

used for large scale analytics in data centres. Moreover, one can in principle enhance such approaches with the cryptographic techniques mentioned above to yield high-assurance guarantees.

Tailored approaches.

For a concrete problem, for example logistic regression on distributed data, custom "hybrid" protocols that combine several of the techniques above are likely to give the best results, by sacrificing on generality. Research prototypes that follow this hybrid approach have been proposed for private training and classification in models such as neural networks, ridge and logistic regression, nearest neighbours, text classification, random forests, among others.

Privacy-preserving disclosure: How much does the result *actually* disclose?

Although the techniques above can be used to compute a statistical model in a privacy-preserving way, namely not disclosing any unnecessary information, they do not address the problem of quantifying how much is disclosed by such model. This (vague) question regarding "how much is disclosed" has many aspects. For example, one might be interested in quantifying to what extend a sensitive feature of the training dataset is disclosed by the model. Alternatively, one could try to address whether the model would allow to deanonymize a public, in principle unrelated dataset, or whether a given individual can be identified as part of the training dataset. Each of these goals captures something about our intuition regarding privacy, and they may be more or less suitable in different contexts. As with issues such as bias and fairness in statistical models, mathematical definitions of privacy are important even if they only capture part of what we intuitively mean by privacy preservation.

One thing is clear amongst information security experts, that simply removing the primary keys (names, birthday, postcode, etc) of a database and replacing with some pseudo-random numbers, so called "de-identification", won't work in general. There are too many diverse holders of records to prevent trivial re-identification (sometimes called triangulation) by linking data from different sources and inferring who the subject is. Another defence, often referred to as k-anonymisation, consists on "fuzzing" the dataset so that any allowed query includes data from at least k individuals, hence providing some uncertainty that should protect privacy. However this also does not account for the above mentioned linkage attacks. In summary, what makes privacy difficult is dimensionality: a sometimes surprisingly small number of features is enough to make a database record essentially unique. Hence, an attacker with a bit of background knowledge about a given individual can use it to obtain the additional information about that person present in a database.

A particularly successful mathematical definition of privacy, as it is receiving lots of attention from both academia and industry, is Differential Privacy (DP). Intuitively, DP allows us to put an envelope around a database and comprehend just what is revealed by queries regarding whether a record was part of the database or not. As put by Dwork and Roth [6]: "Differential privacy describes a promise, made by a data holder, or curator, to a data subject: *You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.*"

How it is done is a detail varies across applications, and can involve several approaches to filtering data collected, or fuzzing features of the data, or analysing and blocking overly intrusive or too-frequent questions. Given most market research style analytics is concerned with identifying groups (segments or bins) in the data, this may not lose any value at all. Users with really obscure or rare features don't represent significant market opportunities.

There are several aspects that make Differential Privacy an appealing definition. First of all, differential privacy neutralises linkage attacks, as it is defined as a property of the analysis, not the data on which the analysis is run. For the same reason, differential privacy is also immune to post-processing: running subsequent analysis on the result of a differentially private analysis cannot result in a less private result. Differential privacy also has nice composition properties,

that allow to build provably private analysis from simpler building blocks. Finally, one of the main characteristics of DP is that it allows for privacy quantification, as it defines privacy in terms of real-valued parameters ε and δ. Setting this parameters properly is in general an open problem that corresponds to the tension between privacy and utility in data analysis: smaller values of the parameters provide better privacy, but might render the analysis useless.

To illustrate the ideas behind DP to provide privacy by a randomised mechanism it is useful to consider the very related idea of *randomised response*. Randomised response is a technique developed in the 60s to collect statistics about illegal or embarrassing behaviour. Every participant of the study is instructed to (a) (privately) flip a coin before replying to the question, if the coin comes up heads then (b) answer truthfully, and if the toss comes up tails (c) randomly answer "yes" or "no", using a second coin toss. The surveyor can then correct the result using their knowledge about the surveying mechanism to get an approximate count. Note that privacy for the participants here comes in the form of plausible deniability: they can always claim "I did not vote for Brexit!, the coin tosses made me report yes ".

Differential Privacy has been shown to be very rich concept with interesting connections to several aspects of information theory and learning. There have also been some applications of it, by big data controllers such as Google and Apple, but a clear path to standardisation does not exist yet. The main challenges have to do with modelling and parameterisation choices, to which DP is very sensitive not only in terms of privacy but also utility.

TOWARDS STANDARDS FOR LARGE SCALE DATA ANALYTICS

The technologies mentioned above and more importantly their interplays are not mature enough to be completely standardised. Moreover, the complexity of secure data analysis will require several kinds of standards, related not only to the different aspects of privacy-preserving analytics discussed above, but also related issues like personal data management and consent. First, just like there are standard virtualisation APIs, we need standard APIs for trusted execution. Moreover, one has to address choices regarding cryptographic protocols, the architectures of which they are deployed (possibly involving semitrusted parties), and with which security guarantees in terms of key sizes and similar parameters. Even if we agree on which protocols to use, and how to instantiate them, there are always a set of services that are required to deploy such protocols in practice, and a reasonable incentive system for parties to provide such services must be in place. This includes tasks such as key distribution, attestation, and verification, which might potentially involve actors focused on these tasks. A major challenge to overcome is that of "how much privacy is enough". Privacy, unlike secrecy or security, is in some cases not a binary predicate, as it undermines utility in many applications. For example, establishing a "safe" differential privacy modelling and parameters for recurrent analysis on sensitive census data is a major challenge. What one means by "safe" would have to be not only rigorously established, but also effectively communicated by, for example, something like kitemarks for safety, but instead for privacy level.

CONCLUSION

Privacy-preserving data analysis is an emerging discipline within data science, which posts several challenges currently being simultaneously tackled from several areas such as hardware/systems security, cryptography, statistics, and machine learning. Several privacy-enhancing techniques have evolved significantly in the last decade from being mainly of theoretical to resulting into academic prototypes and even commercial products and, as recognised by both governments and industry, have the potential to revolutionise the field. These techniques have different tradeoffs, maturity levels, and privacy guarantees, and in some cases solve slightly different problems. A fully fledged approach to privacy-preserving data analysis would still require

significant interdisciplinary effort, some of which have to do with issues such as effective personal data management and consent, which we did not address in this paper.

The need for robust privacy-preserving data analysis technologies has been recognised by both regulators and industry. This would not only mitigate the growing risks of privacy failures, but also enable opportunities based on computing on private data. This is analogous to how encryption revolutionized secure communications, enabling a huge economic development, mainly through secure payments. While regulation and standardization would apparently accelerate this process, the technology is not quite there, and more research is needed before the field as a whole is mature enough to yield precisely defined good practices and regulation, capable of for example enabling audits to ensure compliance.

REFERENCES

1. R. Serra. Television delivers people. https://www.moma.org/collection/works/118185, 1973. Accessed: 2018-02-27.
2. Report of the Commission on Evidence-based Policymaking. https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf, 2017. Accessed: 2018-02-27.
3. V. Costan and S. Devadas. Intel SGX explained. *IACR Cryptology ePrint Archive*, 2016:86, 2016.
4. C. Gentry. Fully homomorphic encryption using ideal lattices. In *STOC*, pages 169–178. ACM, 2009.
5. A. C. Yao. How to generate and exchange secrets (extended abstract). In *FOCS*, pages 162–167. IEEE Computer Society, 1986.
6. C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.