WESTERN SYDNEY
UNIVERSITY

The New Privacy

Emerging Standards for Cloud-Based Security

JANUARY 2019

LUKE MUNN[1]
TSVETELINA HRISTOVA[1]
LIAM MAGEE[1]
DEBRA BOURDIGNON[2]
TIM HORAN[1]
TAMIR LEVIN[2]
TAL NATHAN[2]
LAURENCE PARK[1]

A COLLABORATION BETWEEN
WESTERN SYDNEY UNIVERSITY[1] AND DIMENSION DATA AUSTRALIA[2]

Luke Munn, Tsvetelina Hristova, Liam Magee, Debra Bourdignon, Tim Horan, Tamir Levin, Tal Nathan, Laurence Park
The New Privacy: Emerging Standards for Cloud-based Security

Published by Western Sydney University. This report is commissioned jointly by Western Sydney University and Dimension Data Australia.

Please cite this report as: Munn, L, T, Hristova, Magee, L, Bourdignon, D, Horan, T, Levin, T, Nathan, T & Park, L, 2019, The New Privacy: Emerging Standards for Cloud-based Security, Penrith, Australia: Western Sydney University.

Table of Contents

Summary

From consumer hard drives and enterprise servers, data is migrating to the cloud. Driven by lower costs of ownership, elastic on-demand services, improved interoperability and the insights produced through machine learning, cloud-based computing synthesises the best of previous mainframe and personal computing paradigms.

However the cloud—and the valuable data it houses—is also vulnerable. Breaches, data leaks and linkage attacks are widespread, often bypassing existing security safeguards. Data about individuals has a recognised economic and political value, and the Internet has become a new terrain for cyber attacks on corporations and nation states. In this contested environment, privacy attains a new primacy—a critical issue for customers and a currency of trust for business.

New technologies are emerging to address privacy in the cloud. This whitepaper surveys four approaches: blockchains, differential privacy, multiparty computation (MPC) and fully homomorphic encryption (FHE). While blockchains and differential privacy are relatively mature and well understood, MPC and FHE have been, until recently, obscure topics of academic research.

Today, systems like NTT Secure Platform Laboratories' *San-Shi*, IBM's *HELib* and Microsoft's *SEAL* begin to realise the promise of these new approaches. In this whitepaper we report on two experiments using San-Shi and SEAL, to examine how they manage real-world scenarios. Our purpose is not to evaluate either technology; rather it is to explore how they address privacy challenges in different ways and to consider implications for and limits to their adoption.

For the experiments, we have chosen two fields where privacy is paramount: health and tertiary education. In the health scenario, we consider a patient submitting encrypted blood pressure results to a cloud-based diagnostic service; in the education scenario, we describe how student survey data can be joined with sensitive enrolment records and analysed, without revealing individual identities. Together, these scenarios illustrate how cloud-based encryption can enable institutions to deliver services and derive insights from data while complying with statutory and ethical obligations.

No technology is a silver bullet. If cloud-based encryption is now feasible, it does so while still facing challenges of accessibility, usability and computational cost. Moreover, as debates on Australia's recent Assistance and Access Bill illustrate, privacy remains a heavily contested issue. As more personal data migrates to the cloud, and as its economic and political values rise, many fundamental questions will need answering: who owns it, who can use it, and who willbear the mounting costs of keeping it secure?

Challenges for Cloud Security

Key notes:

- **Increased volume of personal data stored in the cloud**
- **Increased potential of cloud-based data analytics**
- **The number and severity of data breaches rising dramatically**
- **Privacy acknowledged as a 'key pillar' for major cloud providers**

The cloud has become pervasive. Rather than the fixed infrastructure costs associated with custom hardware, advocates argue cloud computation provides a flexible utility, delivered on-demand.[1] Compared with the complexity and cost of traditional enterprise systems, cloud computing presents clear advantages: lower barriers to entry for smaller players, instant access to hardware resources, easy scalability of services, and support for new types of applications and services.[2] Emerging services like machine learning or big data analytics have only foregrounded the intensity of computation required to train models and glean new insights.

Such services exceed the capacities of any individual machine, and underscore the merits of massive parallelism available on the cloud. As the volume, variety and velocity of cloud-based data grows—and adapts into fields of health, education, labor, logistics and smart cities—so too does its potential to deliver new insights, new knowledge and new science.

Simultaneously, media alerts of data breaches, rising in number and in severity, testify to the vulnerability of data stored in the cloud. According to the Breach Level Index (BLI), over 7 million records are compromised every day.[3] According to the BLI, only 4% of breaches are 'secure breaches', where encryption was used and the stolen data was thus rendered useless.[4]

High profile cases further underscore the commercial and legal risks. In 2013, major US retailer Target disclosed that hackers had exploited vulnerabilities in its information systems in order to steal 41 million records related to the company's customer payment card accounts.[5]

In September 2017, consumer credit reporting agency Equifax announced one of the largest breaches to date, revealing that "the names, Social Security numbers, and dates of birth of 143 million US consumers had been exposed."[6] Moreover, congressional statements made later by Equifax management revealed that much of this information was stored in plain text, without being obfuscated, encrypted or anonymized.[7]

Cloud-hosting environments can be threatened by operating system vulnerabilities, poorly configured firewalls, lack of monitoring, weak access structures, and loose authentication.[8] As more individuals, organisations and devices connect to the Internet and depend upon the cloud for data services, these attack points constantly grow in number. According to *Statistica*, Internet of Things (IoT) devices exceed 25 billion in 2018, and are expected to reach three times that figure—75 billion—by 2025.[9] It is unsurprising then that the rate of data breaches appears to be accelerating.

Obfuscation and anonymization of data provide only partial protection. Researchers have shown that, even when names are hashed out or removed, individuals can be re-identified through various techniques. In 2008, for example, streaming giant Netflix published a massive dataset of thoroughly anonymised viewer information for a developer competition. In their paper 'Robust De-anonymisation of Large Sparse Datasets', two researchers demonstrated how this dataset could be cross-referenced against IMDB data in order to identify specific individuals.[10]

In another example, Latanya Sweeney, head of Harvard's Data Privacy lab, identified over 40% of ostensibly anonymous participants in a DNA study using only three key pieces of information: zip code, birthdate and gender.[11] As more information moves online, both purchasable and in the public domain, the frequency of these so-called linkage attacks can only be expected to increase. Data's combinatory possibilities mean that the 'anonymous' dataset of today might well become identifiable tomorrow.

Attacks on cloud-based vulnerabilities and the adversarial capabilities of techniques like de-anonymization exert increased pressure on privacy. But as efforts to undermine privacy grow, so does its perceived importance. Microsoft has recently made privacy one of it's 3 'core

"The days of single systems are irrelevant, are over... the cloud provides a dial you can turn"[15]

Iain Thomson

"2017 was a monumental year for leaks... the number of data records compromised in publicly disclosed data breaches surpassed 2.5 billion, up 88% from 2016"[16]

Gemalto

pillars'.[12] Facebook plans on hiring 10,000 new employees to address security and privacy in the wake of the Cambridge Analytica scandal.[13] And the European Union's General Data Protection Regulation (GDPR) puts individual privacy at the heart of its legislation.[14]

As more (and more personal) information moves online, so too will new techniques for exploiting this information emerge. Privacy in the cloud is neither a guaranteeable feature nor a facet of computing that can afford to be abandoned. For organisations keen to exploit the commercial potential of data, breaches of privacy—like carbon emissions—are a negative externality that they increasingly must account for. Every institution must now ask: how can it maintain the trust of customers, staff, users and the public, while it works with powerful new tools to produce new insights, knowledge and innovation?

"Promises of privacy are often broken promises"[17]

Privacy scholar Paul Ohm

Secure Cloud Computing: Four Approaches

Key points:

- **Blockchains maintain pseudonymous data in a shared, distributed database**
- **Differential encryption adds "noise" to data sets to obscure individual records, without comprising aggregate results**
- **Multiparty computation splits data into meaningless pieces, shared among many devices**
- **Fully homomorphic encryption encrypts data using a variant of public key cryptocraphy**
- **Though different in approach, both multiparty computation and fully homomorphic encryption enable computation on encrypted data**
- **Privacy acknowledged as a 'key pillar' for major cloud providers**

The challenge of keeping data private constitutes a major challenge that technologies have attempted to address in different ways. While encryption techniques date back at least to the Roman Empire, the development of asymmetric encryption in the 1970s marked the beginning of modern computer security. Schemes like RSA (named after its developers Rivest, Shamir, and Adleman) separated the code or "public key" used for encrypting data, from the secret or private key used to decrypt that data. In the decades since, cryptographic work by researchers has dramatically expanded the number and sophistication of these schemes.

Much of this work focuses on secure computation in a networked environment like the Internet. We discuss four examples: blockchains, differential privacy, multiparty computation (MPC) and fully homomorphic encryption (FHE). They examples are not alternatives; they address different aspects of privacy and security, and in some cases can fit together and complement each other. Moreover, they build upon earlier work in public key cryptography, hashing and network security, and to varying degrees, can be retrofitted to existing IT network and database infrastructure. At the same time, they all introduce penalties in either accuracy, efficiency or cost of data operations. Computationally speaking, privacy does not come for free.

Our survey looks to describe the basic mechanisms and use cases of each approach, along with several of their benefits and shortcomings for cloud computing.

BLOCKCHAIN

Best described as "trust-through-transparency", blockchain technology has a counterintuitive approach to the problem of privacy. Three key components distinguish the blockchain from conventional databases: (1) rather than the stored on a centralised server, the blockchain is copied and shared among all users; (2) records in the blockchain are immutable—once added, they should not be deleted or modified; and (3) records are grouped together in a series of "blocks", each of which has a unique identifier that can be referenced by later records. This final property gives the "blockchain" its name.

For many implementations of blockchains, including BitCoin, each record or transaction includes a further signature of the user—a "hash", or digital code, that is unique to that user. This property is central to the privacy aims of blockchains, since such signatures cannot necessarily be tied to specific individuals. In the case of cryptocurrencies, in theory a payment would not reveal the identities of sender and receiver to outside parties.

Many analysts acknowledge blockchain security can be compromised in practice. As Goldfeder et al. note, online payments allow blockchain identifiers to be linked with user cookies, and hence with their identity.[18] The persistence of such identifiers in other transactions means a user's entire life history can potentially be unravelled. Such attacks do not mean that *all* blockchain transactions are necessarily compromised, but do indicate that blockchains offer at best pseudo-anonymity: there are no guarantees that past transactions remain secret if at some point in the future a user's identity is revealed.

The blockchain's peer-to-peer structure also introduces a novel arrangement of relations between participants. As a database that is shared and synchronised across the network, the blockchain attempts to flatten out informational asymmetries. Everyone has access the same amount of information. Any node (i.e. any user) can verify blockchain data, and any node can write back to the chain. Such an egalitarian, 'trustless' network seeks to eradicate centralized control, granting the same visibility and the same functionality to all users.

This architecture also poses challenges. On the one hand, blockchain underpins cryptocurrencies such as BitCoin and Ethereum, and—despite falling prices in 2018—investors in finance see a number of potential uses for it.[19] On the other, its applicability to other sectors remains largely unproven. As a public record of transactions that is never purged nor modified, its ledger would seem to be ripe for re-identification. As Primavera De Filippi argues, "anyone can retrieve the history of all transactions performed on a blockchain and rely on big data analytics in order to retrieve potentially sensitive information."[20]

Beyond these privacy considerations, such storage might also turn out to be limited technically. Esposito et al. note that, while financial data is linear and highly compressed, personal data in healthcare can be both large in size and relational in structure. They warn that just "how well blockchain storage can cope with both requirements is currently unclear."[21]

As a response to these hurdles, some commentators have proposed a hybrid model, where repositories of 'off-chain' personal data are pointed to by small 'on-chain' references.[22] But this seems to merely defer key questions: who owns blockchain, where is it stored, how is it protected, and who can access it?

DIFFERENTIAL PRIVACY

Differential privacy seeks to anonymise data by adding a small amount of randomisation rather than encrypting it. It responds to a common privacy attack involving a process of elimination. Suppose a survey were to ask a controversial question, with either a "yes" or "no" answer. If an attacker wants to know how a given person responded to that question, and they had access to the sum responses *before* as well as *after* she responds, they could easily work out her individual answer.

With differential privacy, that individual's contribution to a dataset would no longer be calculable—it would be zeroed out.[23] Differential privacy asks us to imagine two worlds: in one world, an individual takes a survey and

contributes to a dataset; in the other, she does not. It then formalizes the difference between these 'worlds', ensuring that any statistical query will return the same result. For pioneers Cynthia Dwork and Aaron Roth, this indeterminacy enables a privacy promise: "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."[24]

A core part of this technique is adding statistical noise to individual results. Using the example survey question above, the idea is that the participant's actual response might only be used 50% of the time, with the remaining cases using the result of random 'coin toss'. This undecidability protects individuals at the level of the single record, without disturbing broad aggregate trends very much. Properties of the population are revealed; properties of the person are not.

Differential privacy certainly affords some advantages. Researcher Joe Near explains that the technology has several benefits: it is usable for non-experts, who can run queries without understanding the underlying mechanics; it supports the broad range of queries that analysts are already using; and it integrates with existing data environments, rather than requiring new database architectures.[25]

It has also been deployed in many real-world situations. Apple, for instance, have used differential privacy to analyze the power consumption of websites and the popularity of emojis without comprising individual privacy, and Google has employed it for broad insights into browser malware and traffic analysis in large cities.[26]

Despite its power and simplicity, differential privacy is no magical cure for problems of cloud-based privacy. The addition of noise to query results introduces inaccuracies, and as Dwork and Roth explain, there is a necessary trade-off between precision and anonymisation for all privacy schemes, including differential privacy: "overly accurate answers to too many questions will destroy privacy in a spectacular way."[27] Differential privacy aims at just the right balance between the two.

In addition, in order to know the right amount of noise to add, someone must have access to original data results. In the cloud as well as in other contexts, this assumes "the existence of a trusted and trustworthy curator who holds the data of individuals in a database."[28] While differential privacy disables the ability to obtain individually damaging information *via a query*, it still makes possible obtaining the same information by compromising the data curator through other means.

MULTIPARTY COMPUTATION

Multiparty Computation (MPC) aims to address this particular shortfall. Rather than trusting a benevolent provider, it assumes instead the curators as well as users of data are potentially adversarial.

In this imagined antagonistic world, privacy consists in never trusting any single agent with a meaningful dataset. Instead, using a concept of 'Secret Sharing', valuable information is split into worthless pieces. These are distributed to a large number of curators for storage, computation and analysis, none of whom can attach meaningful value to the data they hold. In response to queries issued by a user, the MPC protocol ensures that correct aggregate results can still be obtained.

The distribution of data across multiple devices means some or all of those devices (depending on the configuration) need to be controlled by an adversary. As Guy Zyskind explains, "an attacker would need to compromise *t* servers at any given point in time to get the data back, which is highly unlikely for a large *t*."[29] This property means that even if a coalition of some curators of the data themselves coordinated to try to access the data, they would not be able to do so without taking control of the other parties involved in the network.

MPC has been discussed in theoretical terms for some time, and was first trialled in a practical setting to manage auctions of Danish sugar beet in 2008.[30] MPC has since seen several other notable real-world deployments, such as the evaluation of gender pay disparities in Boston[31] and tax fraud in Estonia.[32] More recently, engineers from Google have discussed how they use MPC to evaluate advertising views or track Android keyboard use while ensuring a degree of privacy.[33]

With blockchains, at least in their default configuration, all users are also curators of an entire data set. Though data is encrypted, every user has potential access to everything. With differential privacy, a single curator has a completely unobscured view of the data set, and therefore must be trusted. MPC appears to solve both problems: no one user can access any part of the data, other than through queries that report on the data in aggregate.

However these benefits also come at a cost. Queries against the data set involve communication with all curators, involving significant network traffic. Since security is a function of the number of data curators, more security also means more traffic and more processing time. For real-time, mission-critical or time-sensitive operations, these overheads may make MPC too time-consuming or costly.

In a similar way, blockchain-based cryptocurrencies have struggled to keep up with the demands of financial transaction processing. However these constraints mainly affect new transactions broadcast onto the blockchain network—that is, "write"-operations. A large-scale SMC system would appear to suffer the same problems also in reverse, during query or "read"-operations.

Again, a balance between ideal security and practical performance appears key. In the case of the MPC system *San-Shi*, discussed further below, a common configuration involves three servers—with a fourth for redundancy—acting in concert but operated by different curators. This balance aims to raise the cost of a dedicated

"personally-identifiable information should never, ever be stored on a blockchain-based network."[42]

Lucas Mearian, security analyst

security attack without compromising performance too greatly.

Algorithmic optimisations can also greatly reduce overheads of unoptimised or "naive" MPC calculations. San-Shi researchers using have shown that execution of Fisher's exact test—a complex statistical test for crosstabs or contingency tables—over large data sets can be reduced from a hypothetical 20 years to 8 minutes.[34] As with blockchains and homomorphic encryption, extracting high performance from MPC systems is an active area of research.

HOMOMORPHIC ENCRYPTION

With sensitive data and operation, cloud-based service providers usually decrypt data in order to run computations and deliver analysis. Even in otherwise highly securitised environments, this temporary decryption presents an unacceptable vulnerability—privacy is potentially compromised at the moment data is retrieved for computation.

The goal for homomorphic encryption is to operate on encrypted data as if it were decrypted, retaining privacy while enabling data analysis. The term "homomorphism" refers to this "as if" property: operations such as addition and multiplication can be performed on two or more ciphertexts, containing encrypted forms of their original values. Once decrypted, the results of these operations will be the same as those produced by equivalent operations on unencrypted values.

A cloud provider could for example accept an encrypted spreadsheet, perform some kind of statistical analysis upon that data, and return an encrypted result to a user. The user would then decrypt this result, safe in the knowledge the provider knows nothing about either the source data or the result it has calculated.

Though discussed since the 1970s, full homomorphic encryption was thought to be practically infeasible. In 2009 Craig Gentry's thesis introduced, through the mathematical notion of ideal lattices and a technique called 'bootstrapping', the first fully homomorphic encryption (FHE) scheme that would allow unbounded or arbitrary computation.[35] Cryptographer Shai Halevi has spelled out the real-world implications of workable FHE: "Files are often encrypted in transit and at rest, but decrypted while in use. This regimen provides hackers repeated opportunities to steal unencrypted files. But FHE plugs those holes by keeping the data encrypted, while still allowing it to be manipulated."[36]

As with the other three approaches discussed here, FHE has its drawbacks. When applied to the volumes of data and processing demands of cloud computing, homomorphic calculations have a significant penalty in performance. The size of the ciphertext grows enormously with each operation, creating vastly slower processing times. Security expert Bruce Schneier responded to Gentry's announcement by stressing the impracticalities of any scheme that increased computation time by a factor of one trillion.[37]

Over the past decade, the optimisations of homomorphic operations has been an active research topic. Hardware improvements, particularly in graphical processor units (GPUs), have also produced order-of-magnitude improvements.[38] One study compared a number of schemes across comparable hardware; with 80 parameters, the total evaluation time for a technique from 2012 took 48 hours to run, while a more recent scheme from 2015 needed just 8 minutes.[39]

By default, homomorphic encryption does not assist applications that need to merge data sets owned by different users. Since each data set is encrypted by each user's unique secret key, there is no guarantee its homomorphic properties are transferable. Recent work suggests this limitation can be overcome.[40] Under such a scheme, merged data can be analysed without compromising information to either data curator or other users.

Cloud-based analysis of always-encrypted data is steadily becoming more feasible. Microsoft has, for example, recently discussed prospects for commercially-available FHE.[41] As cloud computing becomes increasing pervasive, FHE's high computational costs could envisage an user-pays, "Privacy-as-a-Service" business model. As with existing privacy offerings, this in turn will concern those who advocate privacy as a universal right.

A BRIEF COMPARISON OF APPROACHES...

Each of these four approaches has features well suited to different use cases. While a complete technical evaluation is beyond the scope of this paper, we identify key characteristics that distinguish them.

Evident in the success of cryptocurrencies, blockchains provide a shared record of transactions that are resistent to being repudiated or modified. Privacy can be maintained through signing transactions with a public key, the owner of which is not necessarily identifiable. However, as cases of BitCoin and other cryptocurrencies have shown, in many cases transaction owners can be identified trivially, by correlating public keys with other information such as IP addresses.

In the case of differential privacy, privacy is established by modifying the outputs of queries of a data set, such that those outputs are less likely to reveal characteristics of any one record or individual. For example, if a data set records the height of a group of people, a query about the maximum height—which could ordinarily be used to identify whether a particularly tall individual belongs to the data set—might instead return a probabilistic value, most likely a small amount less than the actual correct answer. This value would still be "good enough", without revealing the tall person's existence in the data set. It is relatively easy to implement, but involves a trade-off between the accuracy of results and the level of protection afforded.

Multi-party computation employs what is termed an "information-theoretic" form of security, which means it cannot be compromised through a brute force attack alone. Blockchains and fully homomorphic encryption employ public key cryptography, and any key can theoretically be broken by an adversary with sufficient computational power. In the case of MPC, obtaining a key to a single data store, or "party", is not enough; multiple parties need to be compromised. While MPC does not directly address the problem of identifying individuals in a data set, it can be paired together with differential privacy or other procedures to de-identify records. In the case of San-Shi, frequency table-style queries will not show frequencies below a certain number, to prevent re-identifcation.

However, MPC does require multiple computational servers or devices, network overhead, and an initial intention to participate in a given MPC scheme. For some applications, it may be preferable to simply send an encrypted data set to a service and have some calculations performed on it. Such a scenario has been termed "secure outsourced computation", and homomorphic encryption is ideally suited to its requirements. However, as noted above, computation is currently extremely costly, and encryption keys can be compromised in theory. For long-lived data sets, the need for an adversary to obtain access to multiple, independently secured containers of that data may make MPC a more secure option.

Case Studies

Key notes:

- **We conduct two experiments using implementations of Secure Multiparty Computation (San-Shi) and Fully Homomorphic Encryption (PySEAL)**
- **The experiments examine opportunities in two fields where privacy is critical: education and health**

CASE STUDY 1: TERTIARY EDUCATION AND SECURE MULTIPARTY COMPUTATION

Universities increasingly survey students to monitor course satisfaction, to boost metrics of engagement, or to gather information for research projects. These surveys are often anonymous, but will sometimes include identifying information such as a student ID.

In such cases, data privacy policies and university ethics committees will often constrain the ways such identifiers may be used, prohibiting the merging of research data with other databases containing course results or student enrolment records. Such constraints adhere to the university's duty of student care, but they also limit potential insights.

San-Shi is a secure multiparty computation system developed by NTT Secure Platform Laboratories, a division of the Nippon Telegraph and Telecommunications (NTT) company.[43] San-Shi provides a number of services beyond encryption: user and data table management, fragmentation and distribution of data across multiple servers, and concatenation and computation of data in responses to statistical queries written in the *R* language.

In the words of its authors, San-Shi "achieves a secure environment that collects sensitive and highly confidential data and provides statistical analysis functions and its results to external users and analysts without revealing the data to anyone. NTT develops technologies enabling the combinational use of sensitive data from multiple companies that cannot be shared normally, and contributes to the creation of a new service market."

How might a cloud-based encryption technology like San-Shi expand the possibilities in such a situation? In this speculative scenario, a research team in the School of Business wants to know how well students feel their courses were pre-

paring them for future work. They would like to administer a survey to the university's students, with questions like:

Please state your level of agreement with the following statement: 'I feel confident my current course is preparing me for the future job market'.

In addition, the team would like to know how student responses related to their course of study, their place of residence, and economic factors such as student debt and household income. A motivating research question might be: do students from lower socio-economic backgrounds feel more or less positive about how their course is preparing them for future employment?

The team applies to the university's ethics committee for permission to administer their survey. The committee informs them that the survey can contain basic questions about work preparedness, but cannot contain sensitive questions regarding income, background or place of residence, as these could compromise students' privacy. However students' postcodes are captured by the university's enrolment system, and the team does obtain approval to ask for student ID numbers. The team also explains clearly to all research participants why they are asking for these identifiers, and emphasises they will not be able to use these identifiers to obtain sensitive information from students. After four weeks of running their survey, the team has 1,000 survey responses, including attitudes about work preparedness.

They then upload a spreadsheet of these responses to the San-Shi system, where it is encrypted. The same system also has an encrypted copy of student enrolment records, including postcodes. By matching student ID numbers, the team can cross-index their survey with the enrolment records to generate a more comprehensive set of student data.

Without being able to look at these records, the team can generate statistics about responses by postcode. A clustering of low or high response postcodes might indicate that attitudes vary spatially across Western Sydney. Using measures of socio-economic disadvantage and cartographic data from the Australian Bureau of Statistics,[44] they then generate a series of maps and tables to explore the data.

The figure above shows the distribution of average scores, where 1 = 'Strongly Disagree' and 5 = 'Strongly Agree', across various postcodes in Western Sydney. Using only the aggregate responses extracted from the San-Shi-encrypted data store, the researchers can detect a strong bias in this spatial distribution. Together with socio-economic data, they are able to develop a tentative response to their research question.

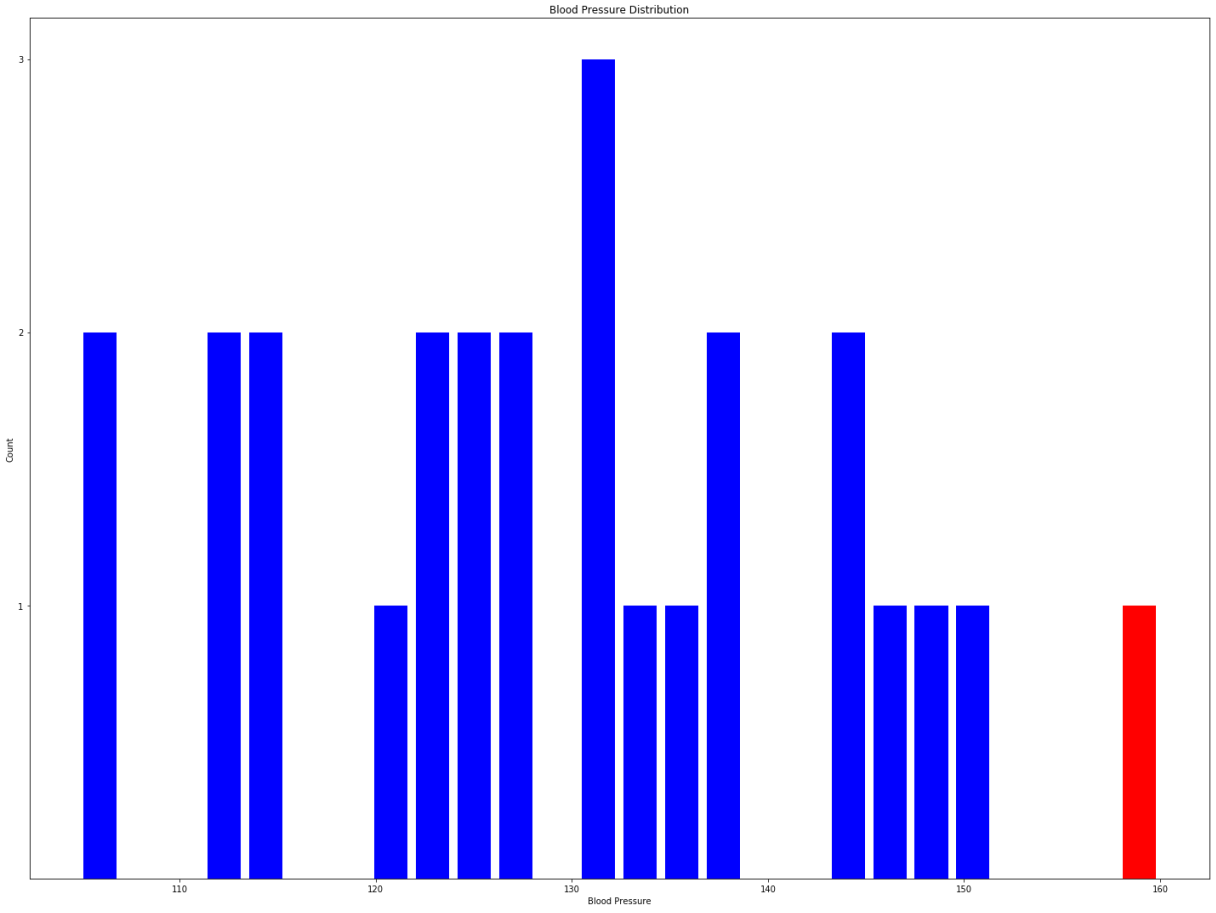
As a cloud-based encryption technology, San-Shi aims to maintain privacy while still enabling computation and analysis. In this speculative scenario, this ability allows the School of Business team to maintain their ethical obligations while also allowing insights into student satisfaction.

The product of several years' intensive research, San-Shi has widespread application in distributed and untrusted computing environments, and its researchers have discussed use cases in healthcare, genome and demographic research.[45] Its core mechanism, the separation of individual data values into "shares" which are encrypted and stored on different machines, has recently been published as an ISO standard, paving the way for interoperability with other systems implementing the same security scheme.[46] Other research has documented how higher order functions such as linear regression can be computed with acceptable accuracy and performance.[47]

CASE STUDY 2: HEALTHCARE AND FULLY HOMOMORPHIC ENCRYPTION

The importance of data security in healthcare has increased with the digitalisation of the healthcare system. Since the introduction of electronic health records in 2012, there have been continuous efforts to improve the infrastructure and security for data transfer between Australian institutions.

One of the biggest challenges is the need to ensure data interoperability and exchange across private and public health providers, and health insurance schemes, such as Medicare. This need has accelerated adoption of cloud storage in recent governmental documents and policies. The e-Health Strategy for NSW Health 2016-2026 pledges, for instance, to "progressively shift elements of its infrastructure to 'cloud-based'



and 'as-a-service' models."[48]

The advantages of cloud storage and computing are also among the recommendations in a recent white paper on perceptions of interoperability in healthcare published by MedicalDirector based on interviews with over 320 industry professionals. However the paper also shows that, while there is a widespread understanding of the importance of data sharing, concerns about security remain unresolved—only 3% of the interviewees stated they trust data sharing.[49]

The healthcare field now places greater importance on health data analytics for diagnostics, prognostic healthcare, and research. But from the transfer of data between medical institutions to big data research scenarios, privacy remains crucial and hard to ensure. Cloud-based encryption presents strong potential for this domain, able to be applied to real-time analytics and predictive diagnostics.[50]

In this scenario, a patient takes a blood pressure test at a local clinic. She would like to know whether this reading indicates a risk of hypertension. The clinic has recently found out about a new secure cloud-based service able to determine if the patient's blood pressure reading is abnormally high using machine learning techniques. Such techniques have been shown "to provide solid prediction capabilities in various application domains including medicine and healthcare, including in the area of hypertension."[51]

The patient is unwilling to share her unencrypted history with online services, concerned that any discovered risk factors might be shared with health insurers or potential employers. Her doctor informs her that her data will be first encrypted, and that the cloud-based service performs its computations solely on that encrypted information. With her consent, the clinic submits the patient's details, including her blood pressure, alongside a database of other, comparable patients—all encrypted.

In this notional dataset, blood pressure values are generated from the World Health Organization and flagged if in the top 5%.[52] The figure shown above plots the patient's result (in red) with a set of other randomly generated readings. The service determines that the patient's reading are indeed abnormally high, and could be a predictor of hypertension.

Implemented in code, this scenario articulated both the possibilities and pitfalls of today's cloud-based encryption. For example, tool sets can be limited or undocumented. In this particular framework, commonly-used operations, like *mean* and *variance*, had to be coded manually. Parameters also have to carefully calibrated—too low and calculation errors emerge, too high and performance declines significantly.

Yet the scenario also sketches possibilities for cloud-based encryption—protecting the privacy of patient data in the cloud, while allowing the client to make use of the computational capabilities of the storage facility.[53]

This scenario accommodates the need for interoperability and transfer of data, while navigating the changing demands for privacy protection and data ownership. Such an encrypted and highly focused analysis allows the patient to make informed choices about lifestyle, diet, and potential treatment, while retaining control of her private and highly valuable health data.

Final Thoughts

Key notes:

- **New advanced privacy technologies for cloud computing are maturing**
- **Usability—for developers, administrators and end users—remains a key issue**
- **Available and cost of privacy options will be a business opportunity—and a political consideration**
- **Cloud providers face a crisis of trust, and will not be resolved by technology alone**

The collaboration between social scientists and IT professionals in preparing this white paper offers a unique insight into the social significance of encryption technologies. The growth in cloud storage and computing for personal and business use establishes privacy as a key commodity with great social importance. Innovations in cloud security—what we have termed technologies of a "New Privacy"—need to be evaluated and planned in consideration with this importance.

IMPROVED PERFORMANCE MAKES CLOUD-BASED ENCRYPTION FEASIBLE

Hardware and software developments over the last decade have led to order-of-magnitude improvements that reduce the substantial computational costs of computation on encrypted data. Performance is no longer a roadblock for real-world deployment. While encryption "has historically been plagued by computational inefficiencies, the field is rapidly advancing to a point where it is efficient enough for practical use in limited settings."[54] Having sufficient information about the capabilities afforded by different security technologies and their social and practical applicability becomes of central importance for choosing a particular cloud security solution.

Products like San-Shi show how data stored in a distributed environment can be made highly secure, and still calculate statistical results accurately and efficiently. San-Shi shows that Multi-Party Computation schemes are becoming feasible, though they require additional security layers to ensure multiple participating nodes cannot be easily compromised. Homomorphic encryption systems cannot be compromised in a similar manner, but are vulnerable to brute force attacks on decryption keys. Current FHE implementations such as *PySEAL* still appear far too slow for cloud-scale computational requirements, though performance is an active area of research for both MPC and FHE schemes.

USABILITY REMAINS A KEY ISSUE

Integrating secure multiparty computation and FHE into real-world projects remains daunting, and limits the scope of adoption. As one encryption specialist observes, "to transform programs into circuits, carefully configure FHE computations, manage encryption and decryption, and other complexities make programming FHE applications the domain of a small number of expert researchers."[55] Current libraries are technically capable, but hardly intuitive for the non-expert.

In the case of San-Shi, several manuals describe how the system needs to managed and used from the various points of view of the system administrator, the database manager and the data analyst. The analyst manual describes common statistical functions for aggregating and filtering data, while higher order functions such as linear regression either need to be composed out of these basic functions, or require separate approaches that to date have not been documented or released as software. Several papers illustrate how these functions nonetheless can perform, often at performance levels acceptably close to those of unencrypted operations.[56]

As this example illustrates, blockchains, secure multiparty computation and fully homomorphic encryption remain complex technologies that are more difficult to install, configure and program than less secure alternatives. The need to operate multiple servers with complex software and database configurations pose daunting challenges for smaller organisations and individuals. Even when deployed to the cloud, each of these approaches mean data sets require careful consideration as to how they are stored, partitioned and secured. The software industry has demonstrated that a focus on usability is necessary to achieve more widespread adoption, and the degree to which an intuitive, seamless experience can be implemented in cloud-based encryption will help determine the success of different solutions.

ACCESS TO PRIVACY IS AN OPEN QUESTION

Issues of usability extend into accessibility. Encryption technologies require certain expertise and access to computational architectures—and this naturally includes certain constituencies while excluding others. Enryption in the cloud affords new opportunities for market growth and consolidation, but this needs to be balanced with public expectations that privacy be considered a basic right rather than a fee-for-service. Such questions become more complex with the rise of public/private data partnerships. Access to privacy thus foregrounds some basic questions: who gets to have privacy, who provides it, and is it a public right or a personal service?

TRUST IS SOCIAL AS WELL AS TECHNICAL

With the feasibility of cloud computing and encryption, trust may appear a feature of the technological environment. Innovations are creating new expectations, affordances, and limitations. If cloud-based data can stay encrypted and secure throughout its lifetime, privacy concerns may abate, and become less effective arguments for restrictions on data capture. At the same time, breaches of privacy and misuse of data highlight the ongoing importance of robust data governance and renewed social contracts for public and private institutions. Technology's impact on institutional trust is powerful and fast-changing, and warrants further study.

Given these new conditions, questions shift from 'having or not having' data to 'what is to be done' with it. What kinds of analyses can be run on encrypted data, and to what ends? The prevalence of linkage attacks should caution institutions about the combinatory potentials of datasets to reveal the intimate and the unexpected. What types of partnerships will cloud-providers enable between public and private sectors, or civil and intelligence communities? While technological advances promise new horizons of capability for the cloud, the wider implications for an increasingly connected society remain—to continue the metaphor—very much up in the air.

Endnotes

1 Rajkumar Buyya et al., "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," Future Generation Computer Systems 25, no. 6 (June 1, 2009): 599, https://doi.org/10.1016/j.future.2008.12.001.
2 M.G. Avram, "Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective," Procedia Technology 12 (2014): 531, https://doi.org/10.1016/j.protcy.2013.12.525.
3 Gemalto, "Data Breach Statistics by Year, Industry, More," Breach Level Index, retrieved from https://breachlevelindex.com.
4 Gemalto, "Data Breach Statistics by Year, Industry, More"
5 Kevin McCoy, "Target to Pay $18.5M for 2013 Data Breach That Affected 41 Million Consumers," USA TODAY, May 23, 2017, https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/.
6 Sean Gallagher, "Equifax Breach Exposed Millions of Driver's Licenses, Phone Numbers, Emails," Ars Technica, May 8, 2018, https://arstechnica.com/information-technology/2018/05/equifax-breach-exposed-millions-of-drivers-licenses-phone-numbers-emails/.
7 Lily Newman, "6 Fresh Horrors From the Equifax CEO's Congressional Hearing," WIRED, October 3, 2017, https://www.wired.com/story/equifax-ceo-congress-testimony/.
8 Farzad Sabahi, "Cloud Computing Security Threats and Responses" (IEEE, 2011), 245–49, https://doi.org/10.1109/ICCSN.2011.6014715.
9 Statistica, "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)", https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/.
10 Arvind Narayanan and Vitaly Shmatikov, "Robust De-Anonymization of Large Sparse Datasets," in Security and Privacy, 2008. SP 2008. IEEE Symposium On (IEEE, 2008), 111–125.
11 Adam Tanner, "Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study," Forbes, April 25, 2013, https://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/.
12 Satya Nadella, "Microsoft CEO: We're Focused on 3 Core Pillars" (Microsoft Build Conference, Seattle, May 7, 2018), http://fortune.com/video/2018/05/07/microsoft-ceo-were-focused-on-3-core-pillars/.
13 Laura Hautala, "Can Facebook's New Hires Take on Troll Farms and Data Privacy?," CNET, April 11, 2018, https://www.cnet.com/news/can-facebook-mark-zuckerberg-new-hires-take-on-troll-farms-and-data-privacy-after-cambridge-analytica/.
14 European Union, "Article 5 – Principles Relating to Processing of Personal Data," § EU General Data Protection Regulation (2018), https://gdpr-info.eu/art-5-gdpr/.
15 Iain Thomson, "Meet TPU 3.0: Google Teases World with Latest Math Coprocessor for AI," The Register, May 9, 2018, https://www.theregister.co.uk/2018/05/09/google_tpu_3/.
16 Gemalto, "2017 Data Breach Level Index: Full Year Results Are In," Gemalto blog, April 13, 2018, https://blog.gemalto.com/security/2018/04/13/data-breach-stats-for-2017-full-year-results-are-in/.
17 Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," UCLA Law Review 57 (2010): 77.
18 Steven Goldfeder, Harry Kalodner, Dillon Reisman and Arvind Narayanan, "When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies", (arXiv 2017), https://arxiv.org/abs/1708.04748.
19 Finextra Research and IBM, "Banking on Blockchain: Charting the Progress of Distributed Ledger Technology In Financial Services" (London: Finextra Research, January 2016), https://www.finextra.com/finextra-downloads/surveys/documents/32e19ab4-2d9c-4862-8416-d3be94161c6d/banking%20on%20blockchain.pdf.
20 Primavera De Filippi, "The Interplay between Decentralization and Privacy: The Case of Blockchain Technologies", Journal of Peer Production (2016), 7.
21 Christian Esposito et al., "Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy?," IEEE Cloud Computing 5, no. 1 (January 2018): 31–37, https://doi.org/10.1109/MCC.2018.011791712.
22 Guy Zyskind, Oz Nathan, and Alex "Sandy" Pentland, "Decentralizing Privacy: Using Blockchain to Protect Personal Data" (IEEE, 2015), 180–84, https://doi.org/10.1109/SPW.2015.27.
23 Cynthia Dwork, "Differential Privacy," in Automata, Languages and Programming, ed. Michele Bugliesi et al. (Springer Berlin Heidelberg, 2006), 1–12.
24 Cynthia Dwork and Aaron Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends® in Theoretical Computer Science 9, no. 3–4 (2013): 216, https://doi.org/10.1561/0400000042.
25 Joe Near, "Differential Privacy at Scale" (USENIX Enigma 2018, Santa Clara, California, January 16, 2018), https://www.youtube.com/watch?v=pk_DCSUayDA.
26 Joe Near, "USENIX Enigma 2018 - Differential Privacy at Scale."
27 Dwork and Roth, "The Algorithmic Foundations of Differential Privacy," 5.
28 Dwork and Roth, "The Algorithmic Foundations of Differential Privacy," 11.
29 Guy Zyskind, "Computing Over Encrypted Data," Enigma, May 30, 2017, https://blog.enigma.co/computing-over-encrypted-data-d36621458447.
30 Peter Bogetoft et al., "Multiparty Computation Goes Live," 2008, http://eprint.iacr.org/2008/068.
31 Andrei Lapets et al., "Web-Based Multi-Party Computation with Application to Anonymous Aggregate Compensation Analytics" (Boston: Computer Science Department, Boston University, 2015), http://www.cs.bu.edu/techreports/pdf/2015-009-mpc-compensation.pdf.
32 Dan Bogdanov et al., "How the Estonian Tax and Customs Board Evaluated a Tax Fraud Detection System Based on Secure Multi-Party Computation," in International Conference on Financial Cryptography and Data Security (Springer, 2015), 227–234, http://fc15.ifca.ai/preproceedings/paper_47.pdf.
33 Tim Wood, "Secure MPC at Google," Bristol Cryptography Blog (blog), January 13, 2017, http://bristolcrypto.blogspot.com/2017/01/rwc-2017-secure-mpc-at-google.html.
34 Koki Hamada et al., "Privacy-Preserving Fisher's Exact Test for Genome-Wide Association Study" Genopri (2017),
35 Craig Gentry, "A Fully Homomorphic Encryption Scheme" (Stanford University, 2009), https://crypto.stanford.edu/craig/craig-thesis.pdf.
36 IBM Research, "Elegant, Disgusting Cryptography: Fully Homomorphic Encryption," IBM Blog Research, March 21, 2018, http://www.ibm.comhttps://www.ibm.com/blogs/research/2018/03/elegant-disgusting-cryptography/.
37 Bruce Schneier, "Homomorphic Encryption Breakthrough," Schneier on Security, July 9, 2009, https://www.schneier.com/blog/archives/2009/07/homomorphic_enc.html.
38 W. Wang et al., "Accelerating Fully Homomorphic Encryption Using GPU," in 2012 IEEE Conference on High Performance Extreme Computing (2012), 1–5, https://doi.org/10.1109/HPEC.2012.6408660.
39 Abbas Acar et al., "A Survey on Homomorphic Encryption Schemes: Theory and Implementation," ArXiv:1704.03578 [Cs], April 11, 2017, http://arxiv.org/abs/1704.03578.
40 López-Alt, A., Tromer, E., & Vaikuntanathan, V. "Multikey Fully Homomorphic Encryption and Applications," SIAM Journal on Computing, https://epubs.siam.org/doi/10.1137/14100124X.
41 Satya Nadella, "Microsoft CEO: We're Focused on 3 Core Pillars" (Microsoft Build Conference, Seattle, May 7, 2018), http://fortune.com/video/2018/05/07/microsoft-ceo-were-focused-on-3-core-pillars/.
42 Lucas Mearian, "Will Blockchain Run Afoul of GDPR? (Yes and No)," Computerworld, May 7, 2018, https://www.computerworld.com/article/3269750/blockchain/will-blockchain-run-afoul-of-gdpr-yes-and-no.html.
43 NTT Secure Platform Laboratories, "Performing a Statistical Analysis of Multiple Companies' Sensitive Data" (NTT R&D Forum, Musashino Research and Development Center, Tokyo, February 13, 2017), http://www.ntt.co.jp/RD/active/201702/en/pdf_eng/03/C-28_e.pdf.
44 Australian Bureau of Statistics, "Socio-economic indexes for Australia (SEIFA) 2016", (2018), retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001 ; Australian Bureau of Statistics, "3218.0 - Regional Population Growth, Australia, 2015-16", (2018), retrieved from http://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/3218.0Main%20Features702015-16?opendocument&tabname=Summary&prodno=3218.0&issue=2015-16&num=&view= .
45 Satoshi Tanaka et al., "Secure statistical computation system on encrypted data: An empirical study of secure regression analysis for official statistics," UNECE Work session on Statistical Data Confidentiality (SDC), (2017), https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/6_secure_computation_system.pdf; Eizen Kimura et al., "Evaluation of Secure Computation in a Distributed Healthcare Setting," Medical Informatics Europe (MIE), (2016): 152-156.
46 ISO/IEC, "SO/IEC 19592-2:2017. Information technology -- Security techniques -- Secret sharing -- Part 2: Fundamental mechanisms", (2018), retrieved from https://www.iso.org/standard/65425.html. A useful visual explainer on how secret sharing works to enable secure addition, multiplication and sorting can be found online: http://www.ntt.co.jp/sc/project_e/data-security/NTT-secure-computation.pdf.
47 Tanaka et al., "Secure statistical computation system on encrypted data: An empirical study of secure regression analysis for official statistics".
48 NSW Ministry of Health, "EHealth Strategy for NSW Health 2016–2026" (Sydney, 2016), retrieved from http://www.health.nsw.gov.au/eHealth/Documents/eHealth-Strategy-for-NSW-Health-2016-2026.pdf.
49 MedicalDirector, "Interoperability in Healthcare" (Sydney, 2018), retrieved from https://www.medicaldirector.com/wp-content/uploads/2018/05/Interoperability_White_Paper_2018.pdf.
50 Micahel Naehrig, Kristin Lauter and Vinod Vaikuntanathan, "Can homomorphic encryption be practical?" in Proceedings of the 3rd ACM workshop on Cloud computing security workshop, 2011, 113-124. ACM.
51 Sherif Sakr et al., "Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project," PLoS One (2018), 13(4), 10.1371/journal.pone.0195344.
52 World Health Organisation, "Mean and standard deviation (SD) of blood pressure", retrieved from https://thl.fi/publications/monica/bp/table8.htm
53 Aderonke Ikuomola and Oluremi Arowolo, "Securing patient privacy in e-health cloud using homomorphic encryption and access control," International Journal of Computer Networks and Communications Security (2014), 2(1): 15-21.
54 Roger A. Hallman et al., "Homomorphic Encryption for Secure Computation on Big Data," SCITEPRESS - Science and Technology Publications (2018): 340–47, https://doi.org/10.5220/0006823203400347.
55 David Archer, "Revolution and Evolution: Fully Homomorphic Encryption," United States Cybersecurity Magazine, Summer 2016, https://www.uscybersecurity.net/csmag/revolution-and-evolution-fully-homomorphic-encryption/.
56 Tanaka et al., "Secure statistical computation system on encrypted data: An empirical study of secure regression analysis for official statistics"