Is Privacy Dead?

Reclaiming Data Privacy in the Digital World

TATA CONSULTANCY SERVICES

Sachin Lodha

IN BRIEF

Consumers, and even security professionals, confuse privacy and security. For most people, tight security is in itself a panacea to all privacy ills. Nothing could be further from the truth, as use cases requiring privacy are quite different from those for security.

We have three methods of handling the privacy challenge: query restriction, output perturbation, and input perturbation. TCS has invested in privacy research for close to two decades building tools for various scenarios using a combination of methods.

Organizations today collect a lot data; analytics of this data provides insights. But actions that follow must respect the fine line between extreme personalization and privacy breach. In a privacy-eroding world, GDPR appears as privacy's north star. It is the yardstick against which all past, present, and future privacy regulations will be measured. GDPR and its clones will nudge compute to go to the data rather than data coming to compute. Privacy has to be a part of the design and not a last minute add-on.

I know that we know nothing

A few years ago I had an opportunity to present our data privacy work to the Chief Information Officer (CIO) of a major finance corporation in North America in his tri-state area office. After listening to me very patiently for about half an hour, he asked me if I would be interested in knowing how they were managing their customers' data privacy. Naturally, I was all ears. He called one of his managers into his office, and asked him, "Hey, John! Do we know what all data we have about our customers?" "Nope," pat came the reply from John. "Are we at least on top of where such data resides in our environment?" "Nope, not entirely," said John again. After few more similar back and forth salvos, the CIO turned to me with a twinkle in his eye and said, "You see now how we manage our customers' data privacy. We simply don't know what data we have got on them; it is all hidden from us as well!" Indeed,

Fact File

TCS Research: Data Privacy

Outcomes: Novel Privacy Enhancing Technologies (PETs) that are part of TCS MasterCraft DataPlus offering

Principal Investigators: Sachin Lodha

Techniques Used: Data Masking, Anonymization, Signal Processing

Industries Benefited: BFS, Insurance, Retail, Telecom, Healthcare and Life Sciences

Patents: 34 filed, 11 granted

Papers: 6

a nice Socratic perspective to one of the most intricate problems, I thought.

Houston, we have a problem

Fast forward to now. The new crude is data! Naturally, its security and privacy have attracted universal attention. Yet, rarely consumers or even security professionals can properly differentiate between the two. For most people, while they may not exactly have the above mentioned Socratic perspective, tight security is in itself a panacea to all privacy ills. However, nothing could be further from the truth.

And here is why. Users and attackers are mutually disjoint groups in security. In privacy, on the other hand, the user himself is considered as an attacker. This makes privacy trickier to tackle than security. It also explains why the standard security controls based on encryption, authentication, and so on, do not address privacy problems even if they allow only legitimate users in. Ultimately, whether we have a privacy breach depends on the legitimacy of purpose(s) for which the accessed data is utilized by the user. Therefore, the grand challenge of privacy is to make such data available to the user so that it has enough utility only for the said purposes and nothing else—ensuring privacy to the data subjects, i.e., those individuals whom the particular personal data is on.

There is no silver bullet yet

Contrary to what a layperson may believe, privacy by itself is a very well-researched topic. Both research and innovation in privacy have gathered a steady steam over the past decade. As one of its outcomes, we now broadly know three classes of methods for tackling the privacy challenge described above.

The first among them is called *query restriction*. Here queries to a database are answered exactly, but not all queries are permitted to maintain privacy. Second class of methods is called *output perturbation*. Here responses are suitably perturbed for all the queries. Prime example of this class of methods is differentially private algorithms that aim to provide means to maximize the accuracy of queries output (utility) from statistical databases while minimizing the chances of

Users and attackers are mutually disjoint groups in security. In privacy, on the other hand, the user himself is considered as an attacker. This makes privacy trickier to tackle than security.

identifying its records (privacy). Finally, we have a class of methods called *input perturbation*. Here queries are answered correctly, however, according to a perturbed database. Several popular data masking methods that help create such a perturbed database could fall under this class. Further *k*-anonymity offers another perfect example of this class. A dataset is said to have the *k*-anonymity property if the information for each person contained in the dataset cannot be distinguished from at least *k*-1 other individuals whose information also appear in the same dataset. Such blending in the group of size *k* offers protection to a data subject from being singled out by the data user.

While all these different classes of methods have their strengths, they do also have some serious limitations. For example, query restriction methods typically involve establishing equivalence between two queries, i.e., is the given query A, same as previously seen query B, and this happens to be a complex computational problem with no efficient algorithm yet known. In case of differential privacy, there are well-known, theoretically proven bounds on the number of queries a differentially private algorithm can answer privately, and therefore, all privacy bets are off once the number of queries crosses a particular threshold. The *k*-anonymity model makes strong assumptions about what an adversary can do. Moreover, its most variants turn out to be computationally hard to solve optimally. In addition, theoretical results suggest that the dimensionality curse is a fundamental barrier to privacy preservation, and an increasing dimensionality makes the data resistant to effective privacy across different classes of methods.

Since all the existing methods work well only under certain favorable conditions and assumptions, a careful analysis is often required even to decide which method(s) to apply for the privacy problem at hand. Even then, we may not simultaneously get the best of both utility and privacy.

There is no silver bullet for the privacy challenge as of now. Crypto researchers hold high hopes for homomorphic encryption schemes that allow the entire data processing to happen over encrypted data. However, their current implementations slow down the data processing by several orders of magnitude, and hence, these schemes are far from being practical.

It is getting dark out here

Today we are collecting data at an unprecedented rate. Big data analytics typically provides insights that help organizations serve their customer better and stay ahead of the curve. Of course, more often than not, such analysis gets very personal and, therefore, leads to privacy breaches.

Organizational privacy policies are designed to be consumer *unfriendly* over personal data collection and its usage. A study conducted by a noted academic institute in 2012 reported that reading and understanding all of the privacy policies an average Internet user encounters in a year would take 76 workdays! It is too big a price to pay.

In fact, in this new David (individual privacy) vs Goliath (organizational utility) war, it looked like Goliath was destined to win from the start. In 2010, Facebook founder Mark Zuckerberg unsurprisingly justified his stance, when he said that people have really gotten comfortable not only sharing more information and different kinds, but more openly and with more people. That the social norm was just something that had evolved over time. This was at a time when Facebook changed its privacy settings for its then-350 million user base, and their approach was rightly criticized by technology experts for manipulating our ideas of privacy for their own benefit.

Every cloud has a silver lining

It took those earth-shaking disclosures by Edward Snowden in 2013 to wake up innocent people from their privacy coma. These disclosures had revealed numerous global surveillance programs run by multiple governments in cooperation with major internet companies that had crucified individual privacy without any remorse. One of the major after-effects of the Snowden episode was that public trust in governments and big enterprises started eroding rapidly. People started demanding more transparency on how their personal data was being managed inside organizations. They also started taking keen interest in proactive and effective utilization of available privacy controls.

A 2014 Oxford survey[1] found that the young adults (age group 18-24 years) from Australia, the United Kingdom (UK), and the United States (US) were foremost among the surveyed population in those countries when it came to managing their privacy on social networking sites. This discovery was complete opposite of what the same study group had discovered in a similar 2006 survey wherein the surveyed young adults were most promiscuous and cared little for their privacy on social sites.

Again, in 2014, there was this famous decision by the Court of Justice of the European Union (CJEU) in favor of a Spanish citizen named Mario Costeja González who wanted Google to forget him and

1 http://oxis.oii.ox.ac.uk/wp-content/uploads/sites/43/2015/01/OxIS-Brochure.pdf

GDPR has twin objectives. One of them is to give EU citizens control back on their personal data. The second objective is to simplify data protection regulations to enable the digital single market in EU.

not show certain undesirable links that came up whenever someone googled him. Among other things, this decision also confirmed a 'Right to be Forgotten' mooted in the proposed General Data Protection Regulation (GDPR) of European Union (EU) that came into force on 25 May 2018.

This brings us to GDPR—probably the greatest thing to have happened to humankind since sliced bread if you were to believe privacy supporters!

GDPR: privacy's shining north star

GDPR has twin objectives. One of them is to give EU citizens control back on their personal data. The second objective is to simplify data protection regulations to enable the digital single market in EU.

It tacitly acknowledges that, in the light of the current state of art, insisting only on confidentiality (privacy) would require organizations to sacrifice data availability (utility). So instead of sacrificing availability (of data to organizations) for maintaining confidentiality (i.e., privacy for data subjects), it brings in data subject control and transparency to replace the confidentiality goal whenever reasonable.

In fact, there are several pertinent changes that GDPR has brought about. First, there is an expanded territorial scope. GDPR also applies to organizations based outside the EU if they collect or process personal data of EU residents. Second, it brings a new set of

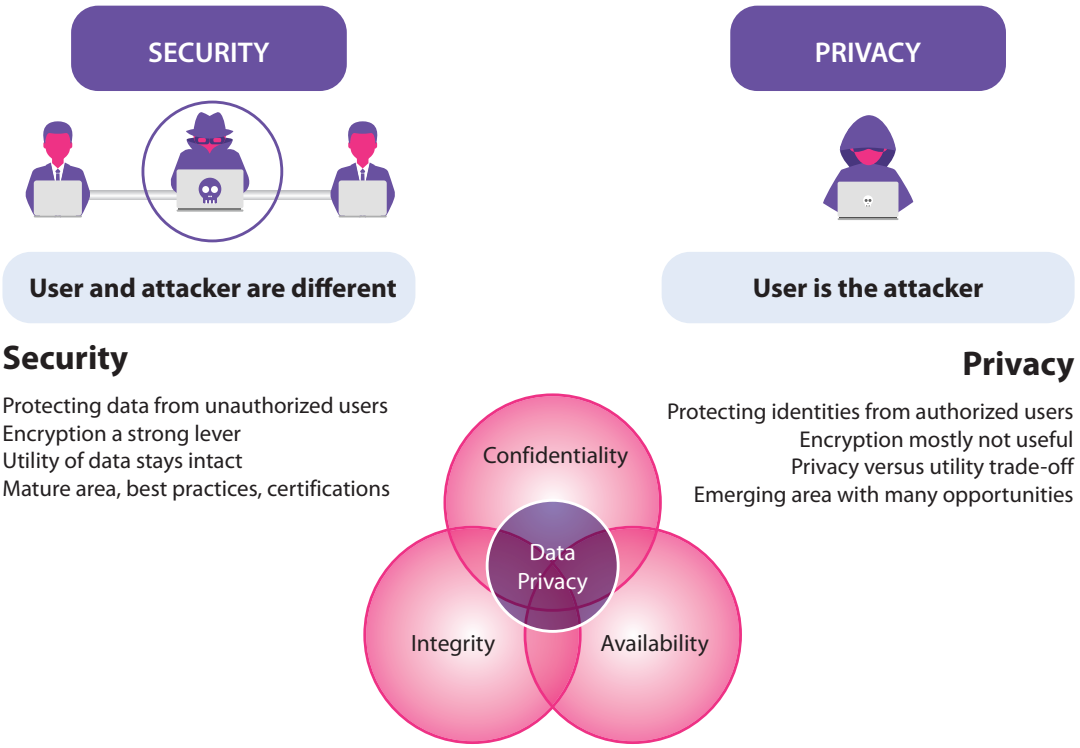

Figure 1: *Security Vs Privacy*

digital rights for EU citizens in an age when the economic value of personal data is increasing in the digital economy. Privacy impact assessment (PIA) and data privacy office (DPO) are a must. There are explicit requirements about breach notification and demonstrable accountability. It also requires privacy by default, i.e., there has to be a clear affirmative and explicit action by individual for letting the organization use her data; so no longer a pre-ticked box for default opt-in. GDPR non-compliance can attract fine of 20 million EUROs, or 4% of the total worldwide annual turnover, whichever is higher.

GDPR is making an individual assume command and control of her own personal data, and shifting the balance of power from callous organizations to simpleton data subjects—a colossal event in the digital world.

Four years before GDPR, we began to focus on the individual-centric privacy paradigm. As one of its major outcomes, we have designed and developed TCS Consent Management Solution that is all set to go to market in near future. It is a novel solution for the entire consent and data rights management life cycle. It facilitates collection, preservation, and enforcement of an individual's privacy preferences over her personal data that is with an organization. Using its granular APIs, organizations can seamlessly integrate the solution into their systems. We are glad to note that the solution has been successfully piloted with TCS' clients. It has also been successfully integrated into TCS BaNCS products as their consent and rights management feature. Going forward as part of our individual-centric privacy research agenda, some of the future goals include (a) building a recommender system to help avoid consent fatigue faced by individuals in the light of GDPR-like regulations; (b) creating aides for improving privacy policy comprehension and privacy related decisions; (c) privacy policy personalization; and so on. We strongly believe that success here would help an individual realize her 'My Data; My Control' dream in the digital era, which is the mission of the new MyData and similar global movements that are springing up in the post-GDPR world.

Looking into the crystal ball

GDPR is going to be the yardstick against which all past, present, and future privacy regulations will be measured. Several countries are currently undergoing GDPR adequacy assessment exercise. Many of them are likely to update their own privacy regulations(s) to be in agreement with the GDPR requirements. For example, the Government of India has set up a Committee of Experts to study various issues relating to data protection in India, make specific suggestions on principles underlying a data protection bill and draft such a bill. Here the objective is same as that of GDPR: to ensure growth of the digital economy while keeping personal data of citizens secure and protected.

Data will become a liability in the light of GDPR and its upcoming clones, as the cost of failure to protect it is going to be high owing to heavy regulatory fines. Expect the data-hungry organizations to be more careful with what personal data they collect, process, or store. 'Lesser the better' will be their new mantra.

It will imply that rather than data (from client devices) coming to compute (i.e., an organizational

Cyberattacks, including ransomware attacks, on personal data will rise as everyone in its food chain—the data owner, controller, or processor—will have something to lose. Organizations will have to be absolutely on top of their security and privacy regimen.

server), compute will go to the data. Here is one such possibility. First organizations will sample data *globally*. Second, based on the sampled data, they will build appropriate piece of logic *centrally* on their side. Finally, deploy and run that logic *locally* on client devices for achieving the desired effects. Notice how no personal data needs to be shared for availing organizational services here.

Cyberattacks, including ransomware attacks, on personal data will rise as everyone in its food chain—the data owner, controller, or processor—will have something to lose. Organizations will have to be absolutely on top of their security and privacy regimen.

In the IT world, security and privacy are often afterthoughts and seldom part of the system design. This will have to change as it may not always be possible to do meaningful privacy as an add-on later, and the cost of failure could be very high. Therefore, expect more adoption of the Privacy by Design approach, especially since it is required by GDPR.

Smarter organizations will make an effort to earn customer trust by ensuring that privacy as a competitive differentiator for their businesses and business technology agendas. When that happens, privacy indeed will have come a long way from merely being a lip service or yet another regulatory compliance burden to being part of the strategy.

That said, there would still be several "Frosty" miles to go before we could reach Ayn Rand's lofty goal of "setting man free from men" when she wrote, "Civilization is the progress toward a society of privacy. The savage's whole existence is public, ruled by the laws of his tribe. Civilization is the process of setting man free from men."

From Socrates to Rand, it is going to be some journey!

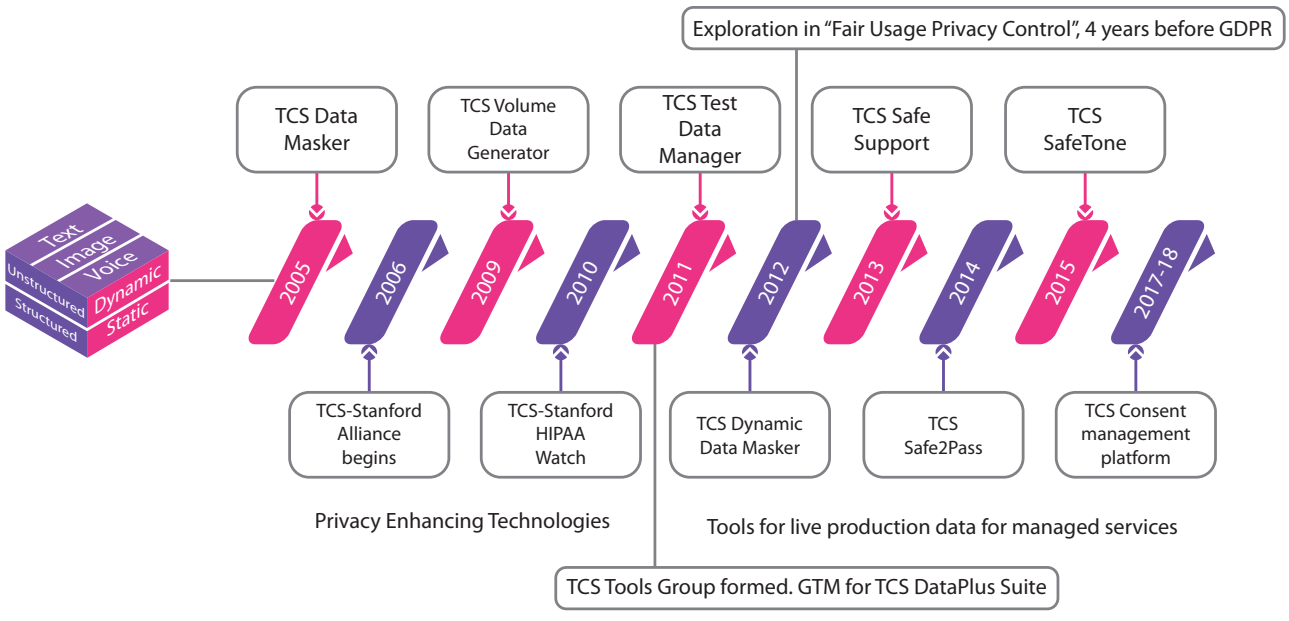

Figure 2: *TCS Privacy research through the years*

Sachin Lodha

Sachin Lodha is a Principal Scientist at TCS Research and Innovation. He leads Cybersecurity and Privacy R&D efforts within the organization, and has a special interest in privacy related topics. His efforts on that front have led to multiple research papers, patent applications, and award-winning innovations that are now available as TCS products. He was also the Principal Investigator for a TCS-Stanford University research collaboration on data privacy that ran successfully between 2006 and 2010. He is also an ACM India Eminent Speaker. Sachin received his Ph.D. in Computer Science from Rutgers University, USA in 2002.