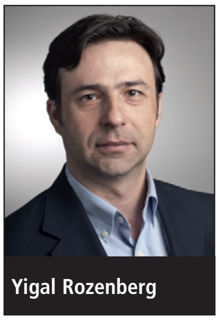
Challenges in PII data protection


Yigal Rozenberg

Yigal Rozenberg, Protegrity

Personal data has become the prime target of hackers and cyber-criminals. Dubbed 'PII' (most commonly taken to mean Personally Identifiable Information, although definitions vary, as we'll see) it can be exploited in many ways, from identity theft, spamming and phishing right through to cyber-espionage. Yet something as simple as a customer database has proven surprisingly difficult to protect. Why is this? And what can be done about it?

What is PII?

PII encompasses any information that can be used to uniquely identify, contact or locate a single person or can be used with other sources to uniquely identify a single individual. By implication, it can be used or sold to facilitate identity theft. The following items generally fall under the core legal definition of PII, although precise details and descriptions will vary by country and legislative infrastructure or requirements:

- Full name, as registered at birth and/or changed by legal deed.
- Date and place of birth.
- Social Security, National Insurance, or other national identification number.
- Passport or residency permit number.
- Driver's licence number.
- Vehicle licence plate number.
- Credit card account information.
- Bank account information.
- Biometric identifiers such as fingerprints and retinal scans.

It is important not to overlook the following secondary items, as sometimes multiple pieces of information – none of which by themselves can be used to uniquely identify an individual – may uniquely identify a person when combined:

- Country, state or city of residence.
- Gender.
- Race.
- Employment details.
- Education details.
- Criminal record.

As long ago as 1990, when datamining technologies were a great deal less sophisticated than they are today, fully 87% of the population of the US could be uniquely identified by gender, ZIP code, and date of birth.[1]

PII protection considerations

The continuing escalation of security breaches involving PII has contributed to the loss of millions of records over the past few years, as well as billions of dollars in lost revenues, falling shareholder values, fines and compensation payments to affected individuals and institutions.

> ***"All PII is not created equal and protection requirements should be evaluated according to its value or impact level"***

To provide appropriate protection for different types of PII, organisations could do a lot worse than follow the advice of McGeorge Bundy, the National Security Advisor to US Presidents Kennedy and Johnson: "If we guard our toothbrushes and diamonds with equal zeal, we will lose fewer toothbrushes and more diamonds." All PII is not created equal and protection requirements should be evaluated according to its value or impact level – the potential harm to individuals and/or the organisation if that PII were to be accessed, used, or disclosed in an inappropriate manner. In assessing these risks, we need to consider the following:

- **Identifiability:** how easily the PII can be used to identify specific individuals.
- **Quantity:** how many individuals can be identified from the PII.
- **Sensitivity:** how sensitive the content of each individual PII data field may be, either alone or in combination.
- **Context:** why the PII is being collected, stored, used, processed, disclosed or disseminated.
- **Obligations:** what laws, regulations or other mandates govern the protection of the PII.
- **Access and location:** who needs to access the PII, where, and how often, in transit and/or at rest.

Any individual organisation may be subject to a different combination of laws, regulations and other mandates related to protecting PII, both in the 'home' country and in other countries in which the organisation does business. This can lead to considerable confusion over which regulations apply to which data in which location, so it's important to include legal counsel and privacy officers in determining current obligations for PII protection.

Note that, while the abbreviation PII is widely accepted, the phrase it abbreviates varies. The 'P' may represent *personal* or *personally* and the first 'I' *identifiable* or *identifying*. The terms are not interchangeable and may affect the legal definition in different jurisdictions and contexts.

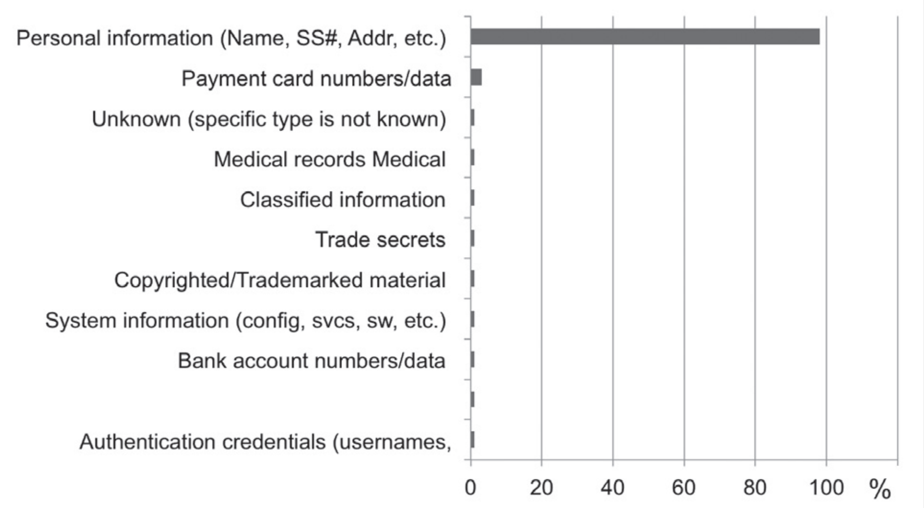


Figure 1: Types of data compromised. Source: 2012 Verizon Data Breach Investigation Report.

When PII protection fails

Each year, Verizon compiles a Data Breach Investigations Report with the US Secret Service (USSS) and the Dutch National High Tech Crime Unit (NHTCU). The recently published 2012 report also took input from the Australian Federal Police (AFP), the Irish Reporting & Information Security Service (IRISSCERT), and the Police Central e-Crime Unit (PCeU) of the London Metropolitan Police.[2]

"It behoves every organisation storing any type of digital information about its customers, employees, shareholders, investors, business partners and other entities to pay close attention to deploying the appropriate technology"

This year's report in particular noted a difference in the types of data targeted by hackers: "Payment card information was again involved in more breaches (48%) than any other data type, but unlike previous years, it was not a runaway winner. Authentication credentials were a close second (42%) in the current dataset. This may allow payment cards to retain the title of 'most stolen' but the title of 'largest hauls' now belongs to the personal information variety, which includes name, email, national IDs, etc. While only 4% of breaches included the loss of personal information, it comprised 95% of the records lost in this report. This is an enormous change from previous years; 4% of record loss in 2009 involved personal information and 2010 showed only 1%."

From email databases (Epsilon) to health records (Blue Cross Blue Shield of Tennessee) to student records (University of Victoria), the vulnerability surface for PII theft is growing exponentially, particularly as data storage facilities shift to the cloud. So it behoves every organisation storing any type of digital information about its customers, employees, shareholders, investors, business partners and other entities to pay close attention to deploying the appropriate technology.

PII protection technologies – what are the options?

Obviously, a number of options exist for the protection of PII, each of which has different strengths and weaknesses and is more or less appropriate for any particular situation. Any and all of the following technologies fall under the general heading of Data Leak (or Loss) Prevention (DLP).

Generically, DLP enables organisations to reduce the corporate risk of the unintentional disclosure of confidential information by identifying, monitoring and acting on unauthorised actions taken on designated confidential data. DLP can be divided into three discrete technologies:

- Endpoint DLP – controlling what the endpoint can do with the data.
- Network DLP. Pattern recognition technology is used to compare data being transmitted out of the organisation against a database of information that's protected from unauthorised transmission.
- Data-at-rest DLP – server-based data protection that may incorporate encryption, tokenisation, and/or obfuscation.

The challenges for PII protection lie primarily in the need for data such as health records or national identity numbers to be available on demand to authorised individuals but kept off-limits to all others. Relying only on access controls, for example, will not prevent the proverbial disgruntled system administrator or DBA from accessing the data.

Hashing

Hash algorithms are one-way functions that turn a message into a fingerprint, which is at least a 20-byte-long binary string to limit the risk of collisions. Hashing can be used to secure data fields in situations in which one does not need to use the original data again, but, unfortunately, a hash will be non-transparent to applications and database schemas since it will require a long binary data-type string. Hashing should be used for passwords – as other solutions are recommended for business data due to transparency and security concerns – and often forms an integral part of DLP systems.

Masking

Masking is another one-way transformation, used to hide or mask information that is presented to users or protected in test databases. Policy-based masking provides the ability to mask selected parts of a sensitive data field. Implemented at the database level rather than at the application level, policy-based data masking provides a consistent level of security across the enterprise without interfering with business operations, and greatly simplifies data security management chores. However, as a data protection strategy, its value is limited to discouraging only the casual thief – a determined attacker will not be deterred by masking.

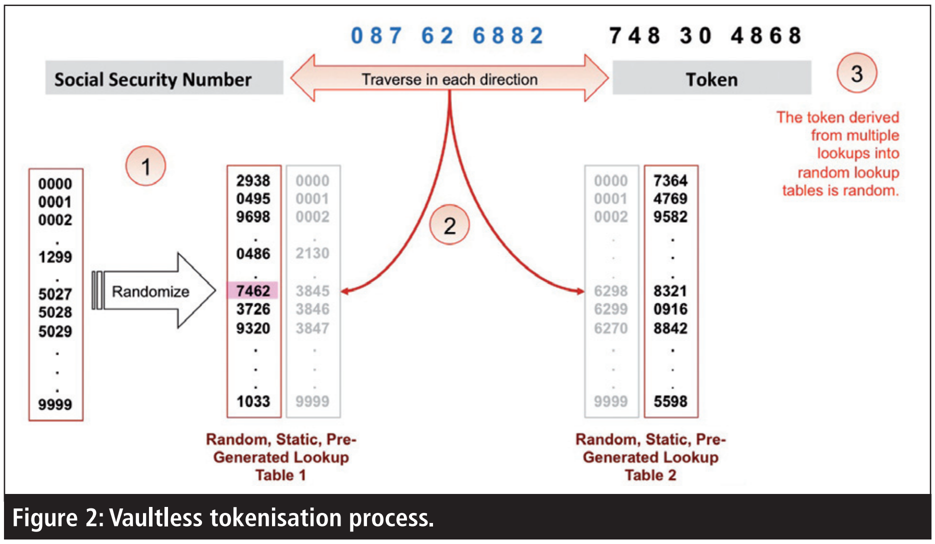


Figure 2: Vaultless tokenisation process.

End-to-end encryption

End-to-end encryption was originally designed to meet the requirements of PCI regulations. It provides strong protection of individual data fields by encrypting sensitive data throughout most of their lifecycle, from capture to disposal, and it is a great way to protect highly sensitive data that need continuous protection in a data flow. It is an emerging data security method today, and many data security vendors carry this type of solution.

As with any other type of encryption, end-to-end uses encryption keys that protect and secure sensitive data based on a mathematical algorithm. While end-to-end encryption provides stronger protection than other forms, the reliance on keys remains a key vulnerability for most types of encryption.

Strong encryption

Strong encryption is cryptography based on industry-tested and validated algorithms, along with strong key lengths and proper key-management practices. This approach is more applicable to high-risk data than formatted encryption. Like formatted encryption, it can also make the process of retrofitting encryption into legacy application environments simpler. However, data fields in transit will be encrypted into a non-transparent binary form, which means strong encryption cannot provide fully transparent protection while the data fields are in use or in transit.

Format-preserving encryption

Also known as data type preserving encryption, format-preserving encryption generates cipher texts of the same length and data type as the input and is typically based on encryption modes that are not standardised. Formatted encryption is known for transparency to applications and databases, and can simplify the process of retrofitting encryption into legacy application environments. It also provides protection while the data fields are in use or in transit, and can be used for lower-risk data and test databases when compliance to industry or government standards is required. However, it also shares with other encryption technologies a high degree of complexity in that different keys are required for different stakeholders and different access needs. This approach is also significantly less secure for short length data, and will require significant overhead for re-encryption of the data to support the required key changes and rotations.

Basic (vaulted) tokenisation

Tokenisation came into being as a result of transaction-driven organisations looking for a more manageable, less intrusive solution than encryption to protect credit card data and still meet PCI requirements. It does this by substituting characters for the credit card numbers, using a look-up table. The tokenised data fits where credit card numbers fit because the token assumes a data type and length identical to the data type and length used to store credit card numbers in databases and to manipulate credit cards in applications. This makes it an ideal solution for structured data – credit card numbers, social security numbers, email addresses or any data that follows a predetermined format. The key concern with basic, or vault-based tokenisation, is its reliance on look-up tables that will quickly become unwieldy, requiring a complex infrastructure and introducing high levels of transaction latency.

In-memory (vaultless) tokenisation

Vaultless tokenisation is much more manageable than its vaulted cousin, particularly in the context of data that needs to be frequently accessed and manipulated, as it uses small, distributed, random, pre-generated look-up tables instead of a single large table. Using multiple small look-up tables reduces or eliminates latency, enabling data to be quickly tokenised and detokenised as needed, and distributing the tokenisation infrastructure provides control over performance and scalability. This approach enables tokens to expose business logic that can be mined and manipulated by business processes without having to return the data to its original form – the 3-2-4 pattern is a popular token pattern – thereby improving accessibility and performance while keeping the data protected. This is the process that enables the security to travel with the data – it's tokenised throughout the workflow, whether at rest, in use or in transit.

Combining technologies for optimum protection

It's clear from even the brief analysis above that hashing and masking are not practical, as the data needs to be accessible in the clear to authorised users. Encryption is not practical as a primary method, although it does form an essential part of the overall data protection strategy as the strategy relates to unstructured data. Key management would quickly become a nightmare when executive A is authorised to access one set of records and executive B is authorised to access a different set, but not the first set. In addition, the overhead in performance and size due to data padding would render encryption impractical as a primary tool. Data type-preserving encryption addresses the size overhead issue but will be significantly slower than the standard encryption algorithms. For a small data warehouse with just a few thousand records, this performance overhead might be negligible right now, but with ever-growing volumes of information, it is not a scalable solution. And there is the security aspect of forever re-keying the data and changing the keys.

Vaulted tokenisation offers better security than encryption, but the growth in data volumes will have a devastating performance impact on the system. Vaultless tokenisation, on the other hand, is a true random tokenisation solution that does not require large, expensive vaults, making it lightweight and fast. As with any tokenisation solution, it will retain the format, size and type of the original data, but will replace it with random generated values. Minimal key management is required, but it is built into the tokenisation system, so organisations do not need to be aware of it. One click will re-encrypt all the static look-up tables with a new key to support key rotation requirements.

The lightweight nature of vaultless tokenisation means all the heavy lifting can be done in memory within the data warehouse or database system – the latency issues of vault-based tokenisation cease to be a factor. Because PII can contain both structured and unstructured data that, in today's 'big data' world, is subject to more or less frequent demands for analysis and manipulation, neither encryption nor tokenisation alone will serve the full spectrum of business needs. So forward-looking organisations are combining the two to apply appropriate value-based protection to its data.

A leading healthcare informatics organisation that has recently adopted vaultless tokenisation for data security is a classic example of how this approach can deliver real business value. The company provides clinical business intelligence solutions, using a Software as a Service (SaaS) model, that connect patient information across multiple medical settings and time periods to generate targeted reports and analyses such as trends in treatment protocols or drug usage. Prior to adopting tokenisation, the company had relied solely on access control and authentication to protect patient data.

The company takes in data – including social security numbers and other PII – in multiple formats, converting everything to a standardised format. With a current database of more than 15 million patients – a number that is expected to triple within five years – any breach would have quickly become unmanageable with this approach. The company would not have been able to identify which data had been affected, nor would it have been able to identify which individual employees had had access to the data.

With vaultless tokenisation, the company is now able to apply protection to the data as part of the format standardisation process, ensuring that it will be able to run analytics on demand while keeping the data fully secured – an approach that would not be possible with other options for data protection.

In conclusion

Data leak prevention technology is a valid approach for the protection of complex data stored only on-premise and where the threat comes solely from disaffected insiders, but could be considered overkill for typical PII storage. Encryption can fulfil many of the requirements for PII protection, especially where unstructured data is involved, but for any business that needs to be able to access and analyse that data in the course of its everyday processes, it has limitations; it simply is not feasible to encrypt and decrypt information on-demand.

In many ways, vaultless tokenisation is ideal for PII – as long as the information is structured. The risk of data loss is to all intents and purposes eliminated, because data that's not readable has no value. This alone makes tokenisation superior to encryption for the protection of structured data: if encryption is deployed insecurely, if administrative accounts are hijacked, or if encryption keys are compromised, the data is exposed and the game is over.

The ideal approach combines encryption for data that's unstructured and/or rarely needed for deep-dive analysis with vaultless tokenisation for structured data that needs to be easily accessible for analysis.

About the author

Yigal Rozenberg joined Protegrity (www.protegrity.com) in 2005. He has more than 20 years' experience in the software *development industry, and over 10 years of developing security solutions. Before joining Protegrity, he was one of the innovators of web application firewall technologies, working for Kavado as vice-president of research and development. Yigal is a co-author on four patents, including patents for web application protection, web application vulnerability detection, database security methods, and co-operative processing and escalation in multi-node application-layer security systems. He holds a degree in computer science from Tel-Aviv University in Israel.*

References

1 Standards of Privacy of Individually Identifiable Health Information, Carnegie Mellon University.

2. '2012 Verizon Data Breach Investigations Report'. Verizon, 2012. Accessed May 2012. www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2012_en_xg.pdf.

Disguising the dangers: hiding attacks behind modern masks

Steven Furnell, Centre for Security, Communications and Network Research, Plymouth University, UK

There's a well-known proverb that says, "the more things change, the more they stay the same". Unfortunately, this can often prove to be the case with our attempts to provide security. Just when we start to get a grip on the nature of the threats and what to look out for, the changing face of the technology we use can serve to confound our efforts. As the technologies change, they can enable old threats to persist in a new guise. Such a pattern has been seen many times with malware, which has managed to hijack a succession of new technologies (from email through to social networks, from desktops through to mobile devices), thus ensuring that as soon as users have been educated and trained to be aware of the threat in one context, it pops up in another that they are not expecting.

In addition to simply representing new threat vectors, the technologies themselves can sometimes provide an ideal mask to hide behind. In order to demonstrate the point, this article considers how a variety of today's technologies can be used to disguise threats and conceal them from users. In all cases, the technologies concerned are things that are typically being used to simplify and improve the user experience, and the aim is to show that this can unfortunately bring additional risks to the unwary.

The first part of the discussion considers the potential risks introduced by shortened URLs and Quick Response (QR) codes, both of which can offer users a more convenient means of linking to online resources by reducing or removing the amount of information they need to remember or enter. However, an implicit effect of both technologies is to obfuscate the details of where the link is leading. In this sense, they can effectively increase risk by giving users a shortcut to a threat. This means that users can't recognise the threat so readily, but can get to it more quickly! The discussion of QR codes has a direct relevance to mobile devices in particular, and so we'll then move on to consider how mobile devices can end up making their own inadvertent contribution to threat masking. In particular, the discussion examines how the simplified experience offered by some mobile devices may lull some users into a false sense of security by convincing them that threats that they are used to on the desktop are not applicable in the mobile environment.

"There are now several ways in which users can be given an address to visit without any means of checking its legitimacy. Two prominent examples are QR codes and shortened URLs"

Masking the threats

To begin the discussion proper, let's turn our attention to the potential for threat masking in the context of links to online resources. Clicking on links represents one of the fundamental means of navigating the web, and so any mechanism for compromising this can clearly represent a threat to a large population of users. Of course, the risk of deception and misdirection is not a new phenomenon in the online world, with a standard ploy in phishing messages being to dress up malicious links to look like they are going somewhere different. However, the savvy or suspicious user could always take steps to check the real destination in advance of actually clicking the link. There are now several ways in which users can now be given an address to visit without any means of checking its legitimacy. Two prominent examples of technologies that