ANALYST BRIEF – March 2014

Why Your Data Breach Is My Problem

THE RISKS OF RELYING ON "PRIVATE" INFORMATION THAT CANNOT BE KEPT PRIVATE

Author – Stefan Frei, PhD

Overview

Modern commerce is increasingly conducted online, allowing vendors to offer a wide variety of goods and services around the clock and from any location. As a result, hundreds of millions of users are registered with dozens of diverse online services.

For authentication, users typically rely on only a small number of unique personal information attributes. The same information attributes are used in several places and inevitably are lost, in large numbers, through data breaches. Cyber criminals have built comprehensive profiles of millions of users, which they constantly refine with each new data breach. Once lost, breached data cannot be taken back. This rapid erosion of security (and also privacy) presents huge challenges as this same information, which many still consider "private," is used across diverse services, both online and offline, While users can change login and password information after a breach, social security numbers (SSNs) and date of birth (DOB) information cannot be changed after such an event.

Enterprises that conduct any part of their business online should bear full responsibility for the consequences of data breaches. Those consequences are not purely financial, but involve ongoing risks posed to individuals and society as a whole because of the loss of static personal information such as DOB and SSN.

NSS Labs Findings

- Within the past decade, massive data breaches have become more and more frequent, resulting in an increase in the loss of personal information.
- A small set of information attributes are commonly used to identify and authenticate a user, and inevitably, these attributes are shared amongst numerous online services.
- Cyber criminals systematically collect and correlate information from data breaches, thereby creating complete user profiles for millions of users.
- Data that once was considered confidential is no longer so (for example, social security numbers).
- While society relies on numbers that are used as both identifiers and authenticators, the system remains insecure.
- Once leaked, static personal information attributes, such as DOB and SSN, pose a considerable risk because these attributes cannot be changed or revoked by users.

NSS Labs Recommendations

- Enterprises must reduce their reliance on authenticating users with data that is potentially shared amongst various services.
- Online services must be designed proactively, i.e., with data breaches in mind, in order to minimize risk and allow enterprises to act quickly to protect their users when necessary.
- Enterprises must systematically analyze third-party data breaches in order to inform and protect their users in the event that shared data is leaked.
- Online services must provide functionality to reauthenticate those users considered "at risk."
- In light of the massive erosion of privacy, governments must reevaluate the use and validity of static personal information attributes, such as DOB or SSN.
- Congress or the Federal Trade Commission should consider legislation to prohibit companies from using SSNs as a means to verify identity.

Table of Contents

Analysis

The businesses behind online service offerings have an inherent interest in learning as much as possible about their users in order to better serve them (for example, by promoting relevant products or by displaying advertisements for external products and services), to lock them into their services, and to secure their business before competitors do.

As the number of online transactions has increased, so too has the value of these transactions, and in more ways than just monetary. As users participate in online services, they are expected to divulge personally identifiable information (PII) that is subsequently stored in various online services portals. Because of economies of scale, the network effect, and almost zero marginal cost, the Internet industry tends toward dominant firms, with some online services attracting hundreds of millions of users. At this scale, online services must rely on remote and fully automated user management for account creation, login/authentication, and password reset to manage and protect the personal information that is stored. Because there is a trade off between security and usability, the secure and efficient management of user accounts is a challenge.

Furthermore, remote user authentication commonly uses a finite number of information attributes per person. These PII attributes are associated with a specific individual and therefore persist across numerous different online services with which the individual is affiliated. A finite and small set of PII attributes available per individual implies a considerable reuse of the same information across services since, for example, a user cannot change DOB or SSN information. Thus, for a large number of users, the security of increasing amounts of critical and economically valuable information depends on just a small set of personal information attributes. Further, this information is managed by fully automated and globally accessible systems specifically designed for large-scale operations. The result is an attractive and rewarding target for cyber criminals.

Data Breaches

The frequency with which data breaches occur and the number of records that are lost in those breaches has increased considerably within the past decade. Figure 1 depicts the top ten largest data breaches of the past decade as of January 2014. Half of these breaches occurred in 2013 alone, with a total of 512 million records lost that year.

#	Date	Organization	Records	Country
1	Oct 2013	Adobe Systems, Inc.	152 M	US
2	Mar 2012	Shanghai Roadway D&B Marketing Services Co. Ltd	150 M	CN
3	Jun 2013	Multiple South Korean businesses	140 M	KR
4	Jan 2009	Heartland Payment Systems	130 M	US
5	Dec 2013	Target Brands, Inc.	110 M	US
6	Jan 2007	TJX Companies Inc.	94 M	US
7	Apr 2011	Sony Corporation	77 M	JP
8	Mar 2013	IRS agents allegedly seized 60M records of 10M people during raid	60 M	?
9	Aug 2008	Government agencies, state firms, telecom companies	50 M	US
10	Apr 2013	LivingSocial Inc.	50 M	US

Figure 1 – Top 10 Largest Data Breaches with Number of Lost Records in Millions [Source: DataLossDB]

This data demonstrates that many records overlap between the breaches (with a total of 512 million records lost for the United States alone) and that the PII of a considerable share of the population of the United States (319

million) was exposed. Research has further shown the existence of a size effect, such that the largest possible amount of identity losses per event grows faster than linearly relative to an organization's size.[1]

Data breaches affect all categories of businesses and services. Figure 2 shows the distribution of data breaches in 2013 for the United States, by breach category, as reported by the Identity Theft Resource Center (ITRS).[2] These are rough estimates, as the reporting requirements for a data breach differ between categories and states. However, Figure 2 clearly demonstrates that a data breach is not an isolated risk affecting a specific category. Some organizations are targeted regardless of what they do, but most become a target because of what they do.[3]

Category	% of Breaches	% of Records	Country
Banking, Credit, Financial	3.7%	1.4%	US
Business	33.9%	81.7%	US
Educational	9.0%	5.6%	US
Government, Military	10.2%	3.3%	US
Medical, Healthcare	43.1%	8.1%	US

Figure 2 – Date Breach Category Summary for 2013 [Source ITRS]

Personal Data: Aggregation and Correlation

Not surprisingly, the ability to aggregate and correlate data is an inviting proposition for data brokers and cyber criminals. A data breach of a popular online service potentially yields millions of records of PII. Because of the versatility and power of correlation and reidentification algorithms, the absence of specific PII data from one breach does not exclude the remaining data from being used to identify individuals. While some attributes may be uniquely identifying on their own, any attribute can be identifying in combination with other attributes from different data breaches and can lead to a profitable market in collecting and reselling stolen PII and financial information.

It has been documented that cyber criminal-friendly web sites market the ability to look up SSN's, birthdays, and other sensitive information on millions of Americans.[4] Different types of information attributes are collected and correlated when available. This information can be used to commit diverse types of identity theft crimes, for example, financial, criminal, medical, and governmental identity theft.[5]

Data that has been leaked in one breach can readily be correlated with data from different breaches and with publicly available information. The privacy and confidentiality of the information attributes that typically have been used to identify and authenticate users for online services is quickly eroding.

[1] Heavy-Tailed Distribution of Cyber-Risks – http://arxiv.org/abs/0803.2256

[2] Identity Theft Resource Center (ITRS) – http://www.idtheftcenter.org/images/breach/2013/BreachStatsReportSummary2013.pdf

[3] Verizon Data Breach Report (DBIR 2013) – http://www.verizonenterprise.com/DBIR/2013

[4] Krebs on Security – http://krebsonsecurity.com/2011/11/how-much-is-your-identity-worth

[5] The Multiple Faces Of Identity Theft – http://www.idtheftcenter.org/Privacy-Issues/classification-to-mitigation.html

Internet Scale Risk Exposure

In the event of a data breach, it is not only the victim enterprise and its customers/users that are exposed. In today's highly networked environment, a data breach affects any enterprise that hosts user accounts. Each breach reveals information, which, when correlated with previously breached information, enhances the information cyber criminals have about users. Over time, this leads to the creation of accurate profiles and the identification of individual users on an unprecedented scale. Breached data cannot be taken back and will eventually reside in the databases of cyber criminals. To assess the aggregated risk for enterprises and society alike, it is necessary to differentiate between the *static* information and the *dynamic* information that is typically used to identify and authenticate users.

Static Information Attributes

Static information attributes are linked to a specific user (for example, gender) and, with rare exception, cannot be changed by the user without extraordinary effort or the help of the government (for example, changing a SSN). Even in the case of rare exceptions, this information is considered static because it cannot be changed merely for the sake of creating different user profiles for various online services. This information works to the advantage of cyber criminals because new information attributes can be added to existing user profiles as they become available, regardless of the source of the data breach. Further exacerbating the situation is the fact that users frequently share personal information on various social media portals.

Cyber criminals have already been collecting and correlating breach information, and eventually they will be able to accurately identify individual users in large numbers. Therefore, in the long term, these static information attributes will no longer be considered "private."

Static Information	• Date of birth • Place of birth • Social security number (SSN) • Gender • Citizenship	• Biometric information • Physical home address • Email address (to some degree) • Security challenge question and answer (if preset by portal) • Mobile number (to some degree)

Figure 3 – Typical Static Information Attributes Used to Identify Users

Once exposed, these static PII attributes are no longer useful for secure user identification and authentication in the online domain, and they considerably expose the user's risk of identity theft. However, since the attributes are static, the user cannot prevent their abuse and accordingly cannot defend against continued identity theft (for example, a user cannot revoke an exposed SSN or DOB). Further, recent research has demonstrated that SSNs can be gleaned from public data, which implies that data breaches are not required to obtain an individual's SSN.[6]

The notion that static information attributes are confidential is flawed; therefore, using these attributes online for security purposes puts a user at considerable risk of identity theft.

[6] http://www.wired.com/wiredscience/2009/07/predictingssn/

It is not possible for users to maintain confidentiality of their SSNs because this information is required when registering for various government and business services in everyday life. SSNs therefore are exposed to data breaches beyond the control of the users, as shown in figure 1.

Enterprises and governments should refrain from using these attributes for online security purposes, although historically they have been considered confidential. While the use of these attributes for security purposes is acceptable in the offline world (requiring physical presence and documents for identification), they are inappropriate online. Physical documents can be forged, but this can be an expensive undertaking, and their use is bound to physical presence, which eliminates the economies of scale provided by the Internet, and places the cyber criminal at serious risk.

Dynamic Information Attributes

Unlike static information attributes, users can change dynamic information attributes with relative ease. Having the ability to make these changes allows users to create variance among their various online profiles. However, managing a large set of different information attributes for various online services is a challenge in its own right.

Although users are aware that passwords should be complex and should not be repeated for different services, analysis of breached data continues to reveal a considerable number of weak passwords and passwords that are shared amongst unrelated online services.[7] Thus, it is safe to assume that users share the majority of the dynamic information attributes listed in Figure 4 across a considerable number of online services. Enterprises must assist users with the daunting task of creating variance among the identifying information that is used, or at least the design of their online services should not inhibit users.

Dynamic Information	• Username/Login • Email address • Password • Security challenge questions and answers (If the questions and answers can be freely chosen by user) • Credit card number • Mobile number (for multi-channel authentication)

Figure 4 – Typical Dynamic Information Attributes Used to Identify Users

Single Sign-On

Single sign-on solutions, where the web portal allows users to register with credentials from existing accounts (Google, Facebook, Twitter, Microsoft), lessen the burden of managing a large number of different login credentials. While the providers of these services typically deploy sophisticated and secure account management functionality, they are also high-value targets for data breaches. Reliance on single sign-on services benefits enterprises that lack the in-house expertise and resources required to provide secure account management functionality.

[7] Trustwave SpiderLabs – http://blog.spiderlabs.com/2013/12/look-what-i-found-moar-pony.html

However, such solutions concentrate the risk on the single sign-on providers, requiring them to be trusted partners (confidentiality, integrity, and availability). Further, these providers are able to track users since they know which external services the users are registered with. Since these providers are largely US-based, there is a potential privacy issue associated with using their services.

Planning For Data Breaches

You Are Not the Only Online Service

As an increasing number of services move online, users manage dozens of online accounts with a variety of service offerings from retailers, social media, applications (apps), governments, and healthcare to customer reward programs. Figure 5 shows the diversity within a sample of frequently used online services. Enterprises typically build their online services by focusing on *their* business needs rather than on the needs of the prospective users and customers. This leads to a specific set of design decisions, which are implemented in their online services.

This isolated view, however, ignores the fact that users are affiliated with a growing number of other online services. With only a limited number of personal information attributes available, substantial reuse of this data across other services is inevitable.

• Social media • Online/offline stores • Online content (books, music, movies, newspapers) • Online games	• Cloud services • Booking services (airline, train, hotel) • Financial (banking, insurance, credit cards) • Healthcare	• Dating • Shipping and transport • Membership/loyalty reward programs • Government services

Figure 5 – Sample of Frequently Used Online Services

When building and assessing the risk of its online services, an enterprise must take into consideration the growing number of data breaches and the effect of third-party data losses, in particular that:

- There is a small and finite set of personal information attributes available to identify and authenticate users.
- Static information attributes can no longer be considered private.
- Users inevitably have to share a large set of information attributes amongst different online services. While users should not share passwords amongst services, they have no option but to share static information attributes.
- Data breaches (internal or third-party) will continue to occur; enterprises must establish processes to address any compromises and keep the risk to users low.

Enterprises should plan and design their online offerings to minimize risks and to assist users in creating and maintaining variance in the identifying information they use for different services.

Protect Your Customers – Anticipate Internal and External Data Breaches

Do not Store Excessive Data

- Store only the minimum amount of personal information required for a service to function; data attributes that are not collected or stored cannot be leaked.

- Do not store personal data if it is used only for one transaction; delete the data after the transaction has completed or after a grace period.

Anonymize, Encrypt

- Wherever possible, do not store sensitive information as plain text; data should be stored in a cryptographically protected format (see *Password Hashing and Salting* in the appendix section).
- Wherever possible, anonymize data to disassociate it from a specific user (for example, data used for analytics and statistics).

Account Termination

- Allow users to terminate an account and thereby have all personal data deleted securely.
- Elimination of data should include personal data that is retained in backups.

Help Users Remain Secure

The design of the online service must enable users to stay secure – and should not prevent them from doing so. Further, in the case of a suspected breach, enable users to become part of the solution (for example, once users learn their endpoint has been compromised).

- Allow long passwords and do not restrict password characters.
- Prevent users from using passwords leaked from known breaches or that are included in cracking dictionaries.
- Allow users to generate their own challenge questions.
- Allow users to monitor, control, and constrain activity on their account (based on geo-location, maximum amount of purchases, notifications for specific actions).
- Show the user recent account activity (for example, login history, transactions).
- Give users the option to temporarily lock their accounts in the event of suspected identity theft or compromise of their endpoints. This helps prevent further abuse while the matter is under investigation (for example, users are able to lock their accounts from a trusted PC, and they receive an unlock code to remove the lock once the matter is resolved).

Prepare for a Data Breach

Enterprises must systematically analyze data from third-party data breaches in order to identify any threats to their user base. For example, organizations can inform users at risk, flag accounts at risk, and re-authenticate users at risk upon next login.

- Upon account creation, establish a secondary communication channel with the user in the event that the primary channel is compromised (for example, a secondary e-mail address [but not the same address as that used as the login name] or a verified mobile number).
- Create a call center with an emergency number to reauthenticate users (personal interaction with call center agents is less scalable to cyber criminals than is automated account takeover).
- Notify users about suspicious account activity.
- Temporarily disable the critical functionality of users at risk.
- Prepare to reauthenticate users at risk.

Reauthentication

Challenge questions for reauthenticating users at risk must not be derived from information that is likely shared with other online services. Such challenge questions can be created automatically in case of need, based on the users unique profile data and the usage history with the service. Data-mining algorithms that have proved powerful in social media and online ads (for example, to suggest new content relevant to a specific user) can be used to create such challenge questions. The advantage being that such questions and answers need not be established and stored (risk of breach) upon account creation. Examples of such questions include:

- Users can be asked questions regarding their transaction history (What was the user's first/last/largest/smallest transaction/purchase (for example, for online stores)? Enterprises can draw on psychological research when composing questions; for example, research shows that within a series of events, people best recall the first and last event, as well as any outliers.
- Users can be asked how frequently, or from what locations, they have used the online service within a specific period of time (correct and incorrect options can be derived from usage data)?
- Questions can be asked based on social network or on contacts within the user's profile, for example, users can be asked whether they know/are affiliated with/communicate with person X.

Industry Collaboration

The continuing large-scale erosion of privacy related to data once considered confidential poses a challenge not only to the industry but to society as well. Governments and the industry should reevaluate their reliance on identifying information attributes that users cannot keep confidential (such as SSN or DOB). Large-scale fraud through the misuse of such data is already occurring: The Internal Revenue Service (IRS) may have delivered more than USD $5 billion in refund checks to identity thieves who filed fraudulent tax returns for 2011.[8]

Governments and the industry should consider setting up a trusted clearing-house that systematically collects and analyzes breached data in order to notify and consult the operators of services at risk and to help users assess their risk.

During policy pricing, the cyber insurance industry should take into account an enterprise's measures to protect against internal and external data breaches.

[8] IRS May Have Lost Billions to Identity Theft, Treasury Says - http://www.huffingtonpost.com/2012/08/02/irs-identity-theft_n_1733905.html

Appendix

Definitions

The standard definitions used in legal context vary by country and state; an overview for the United States is found here.[9]

Personally Identifiable Information (PII) Personal Information	There is no universally accepted definition. In the US context, the abbreviation PII is widely used, while in countries with privacy protection laws derived from the Organization for Economic Cooperation and Development (OECD) privacy principles, the term "personal information" is more common. "Personally identifiable information" (PII) or "personal information" is information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context.
Data Breach	A data breach is the intentional or unintentional release of secure information to an untrusted environment.
Identity Theft	The improper or illegal use of an individual's personal identifying Information (PII)

Password Hashing and Salting

Hash algorithms are one-way functions that convert any amount of data into a fixed-length "fingerprint" that cannot be reversed. Further, if the input changes by even a miniscule amount, the resulting hash is completely different. The general workflow for account registration and authentication in a hash-based account system is as follows:

- The user creates an account.
- The password is hashed, and only the hash is stored in the database. The plain-text password is never stored.
- Upon login, the hash of the password entered is checked against the hash stored in the user database.
- If the hashes match, the user is granted access. If not, the user is informed that invalid login credentials have been entered.

The online service must never alert the user if the username or password is entered incorrectly. Instead, a generic message such as "invalid username or password" should be displayed. This prevents attackers from identifying valid usernames without knowledge of their corresponding passwords. If the user database is breached, the attacker will not find clear-text passwords. To render password cracking and the use of lookup tables (for example, from previously breached data) ineffective, the password must be further randomized. This is achieved by *salting* the password before applying the hash.

A random string, termed a *salt*, is appended to the password before hashing. This converts the same password hash into different strings every time. To verify if a password is correct, the salt is required; the salt is usually

[9] http://www.bakerlaw.com/files/Uploads/Documents/Data%20Breach%20documents/Data_Breach_Charts.pdf

stored in the user account database along with the hash. The salt does not have to be secret. Just by randomizing the hashes, lookup tables become ineffective as each user has a different salt. For example, an attacker can no longer simply search for the hash of a password "123456" in the database. Every user's salt would have to be tested against "123456," which renders such attacks ineffective. Thus, attackers cannot compute in advance a lookup table for all users. Salting has been highly effective in "mass crack" activities, for example, simultaneously cracking millions of LinkedIn passwords; however, if the hacker is targeting a single password, salting makes little difference.

Password Cracking and Dictionary Attacks

To extract passwords from hashes, attackers automatically generate plain-text passwords, apply the hash algorithm, and then test the resulting hash for matches within the breached user database. The plain-text password is revealed when a match is found. To do so attackers automatically generate random passwords and use dictionaries of common words (or previously leaked plain-text passwords) to seed their cracking engine. With ordinary hardware, it is possible to test tens of millions of passwords per second; a determined attacker with specialized hardware can test billions of passwords per second. Hashing and salting considerably slow down such attacks and prevent the use of lookup tables that have been computed in advance. Still, the recent massive data breaches have taught cyber criminals what patterns to look for, and this has made their guesswork far more successful.

Reading List

Top 20 Best Practices to Help Reduce the Threat of the Targeted Persistent Attack. NSS Labs, October 2012
https://www.nsslabs.com/reports/top-20-best-practices-help-reduce-threat-targeted-persistent-attack

Multiple Drivers for Cyber Security Insurance. NSS Labs, November 2013
https://www.nsslabs.com/reports/multiple-drivers-cyber-security-insurance

Online Banking Fraud 2 – The Shifting Legal Burden on Banks. NSS Labs, July 2013
https://www.nsslabs.com/reports/online-banking-fraud-2-shifting-legal-burden-banks

Contact Information

NSS Labs, Inc.
206 Wild Basin Rd
Building A, Suite 200
Austin, TX 78746 USA
+1 (512) 961-5300
info@nsslabs.com
www.nsslabs.com

This analyst brief was produced as part of NSS Labs' independent testing information services. Leading products were tested at no cost to the vendor, and NSS Labs received no vendor funding to produce this analyst brief.