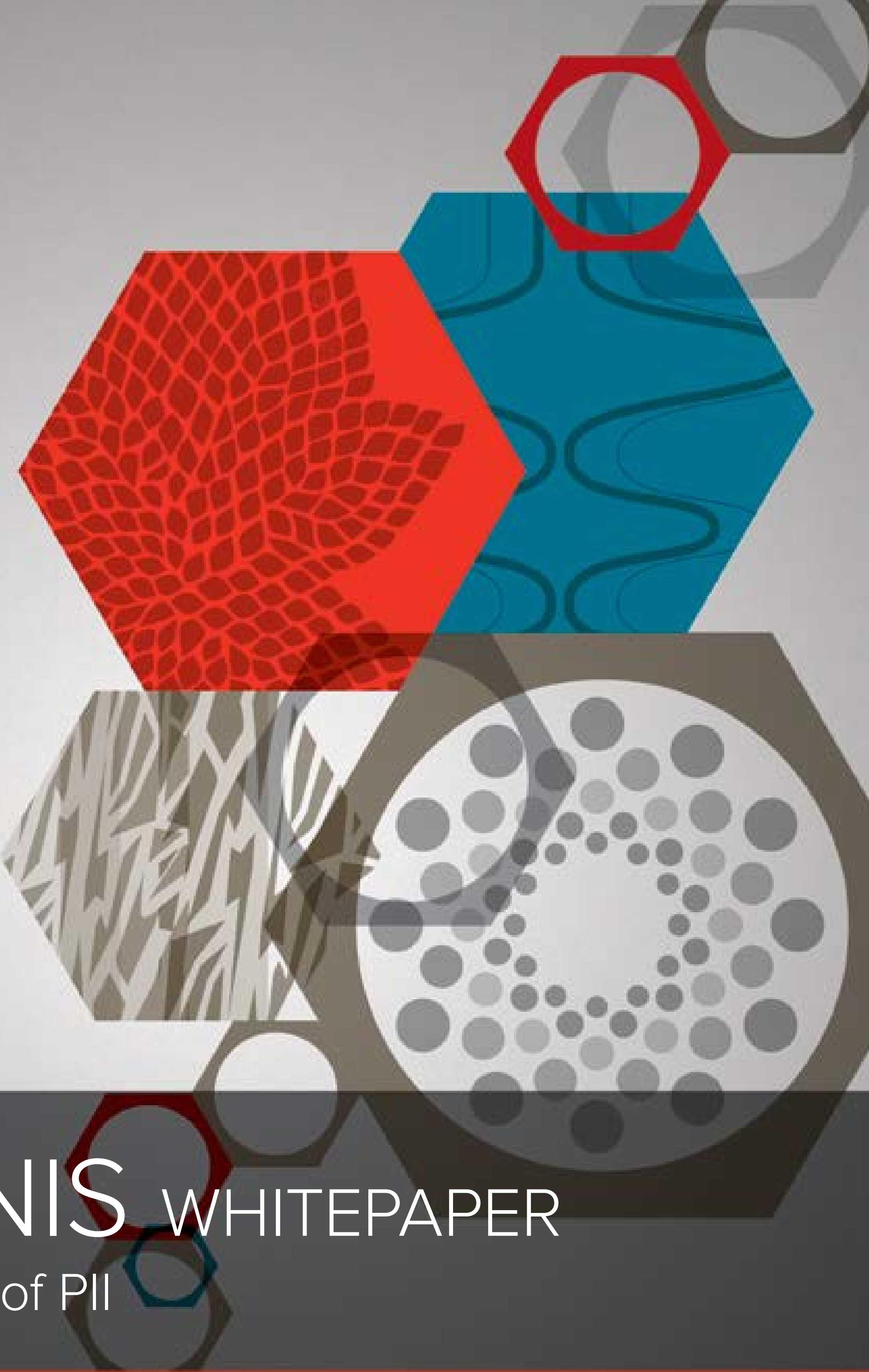


VARONIS WHITEPAPER

The Evolution of PII

CONTENTS

THE EVOLUTION OF PII

Overview

Personally identifiable information, or PII, is at the center of data protection laws and regulations. Why? Under consumer data laws, all data is not treated equally. Organizations are required by law to secure PII and ensure that only authorized users can access it. Anonymous data – think business statistics, aggregated data, or transactional data stripped of identifiers --receives no special treatment under US laws and regulations. In other countries, especially in the EU community, there are similar notions about the types of data that must be protected.

Despite being such a critical part of US data security laws, PII does not have a single, unambiguous definition. This is partially a result of the patchwork of data laws focused on individual sectors, but also because identity is itself a slippery concept. In the early days of computerization–circa 1974--Congress passed the Privacy Act, which placed restrictions on how federal agencies could use personal data, and for the first time referenced identifiers in a digital context. The Privacy Act mentions a person's name as the sole example of an identifier. The Act, though, left open the possibility of other symbols or identifying numbers.

This is not to say that others at that time hadn't improved on the basic PII definition. In the late 1970s, the Organization for Economic Cooperation and Development (OECD) began developing privacy guidelines for computer records. Their final guidelines issued in 1980 have had a great influence on EU data protection laws, as well as more recent US regulations and policies. The OECD's far more abstract and technologically neutral view of PII as any information that can relate back to a person has become the starting point in current policy discussions[1].

This rise of the social media, consumers' willingness to share personal information, along with more powerful computing resources, were some of the key drivers that began forcing regulators to reconsider their legacy definitions of PII. Beyond classic numeric or code identifiers, new "quasi-identifiers" emerged that had the appearance of anonymous data. However, researchers showed it was possible to use standard computing elements to match quasi-identifiers to public databases and social media profiles, ultimately re-identifying the data subject. Regulators began to recognize that there's simply no longer a strict distinction between PII and anonymous data.

In the US, the OECD guidelines have been embraced by the Federal Trade Commission in their own recently released privacy guidelines, Protecting Consumer Privacy in an Era of Rapid Change. With the OECD as a foundation, FTC regulators are viewing PII in even more general terms. The new question for them to consider is how much computing is needed to connect information to a consumer? Regulators are now proposing as a PII definition any collection of information that, using reasonable means and effort, can be related back to an individual. The addition of the word "reasonable" in the PII definition has also worked its way into the EU Commissions current revisions to its long-standing Data Protection Directive.

Companies in medical and financial areas have long had to protect the PII defined by the relevant laws—HIPAA and Gramm-Leach-Bliley. Newer definitions of PII will impact them first as they will be forced to re-evaluate access and other controls for the human-generated content in their file systems. It is likely that additional laws—for example, the proposed White House Consumer Bill of Rights-- will eventually be passed and thereby bring new regulations based on this expanded view of PII to all companies that collect consumer data.

The potential liabilities for not monitoring and controlling access to PII in corporate digital assets will only grow in the coming years. Organizations can stay ahead of the compliance curve by discovering where their unprotected or poorly controlled PII resides, identifying owners of this high-risk content, and then putting in place proper access controls that are backed up continuous monitoring and auditing.

HOW PII HAS EVOLVED

Classic PII

Most of us have an intuitive idea of codes or identifiers that link back to a person: name, address, phone number, and social security number are all good examples. When regulators first considered the privacy issues with computerized government information, they understood that a key was needed to retrieve the records from the database. So it was quite natural to give this key (and information in which the key was found) special protections. Until recently, this notion of PII as a simple key is found in many US laws, and can be traced to the beginning of data protection legislation.

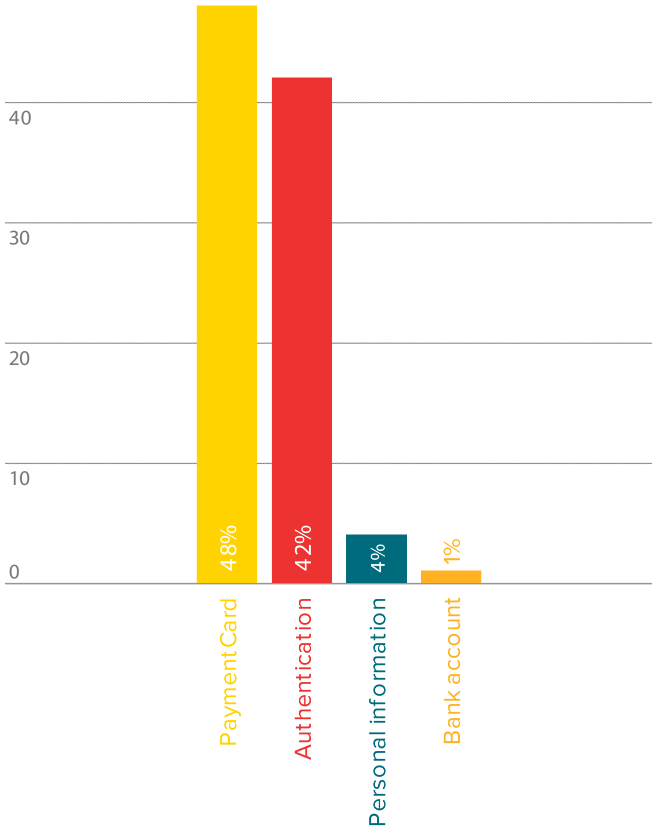
In one of the earliest US law involving PII, The Privacy Act of 1974, Congress responded to the growing use of government databases by placing restrictions on who was allowed to see the personally identifying information. The law required that US agencies receive explicit opt-in from individuals when sharing records with other government departments"[2] . There are some exceptions to this rule, but the intent was clear: data containing identifiers should receive a higher classification off privacy protection than other data. It's also important to note that The Privacy Act broadly excludes from privacy protections any statistical data or summary data from which identifiers have been removed.

Skipping ahead to the 1990s—there was little advancement in PII during the 80s-- the Gramm-Leach-Bliley Act (GLBA) extended consumer data protection and privacy to financial data. It's also one of the earliest examples where PII is itself explicitly mentioned. The law defined nonpublic personally identifiable information or NPI as any information that a consumer provides to obtain a financial product or service—for example, information entered on an application form for an account, loan, mortgage, IRA, etc.—that is not available as publicly viewable data.

Based on the GLBA definition, NPI would normally include, say, name, bank account codes, and social security numbers, along with sensitive financial data, but not, for example, home price values, which are part of the public record. In terms of financial data protection, GBLA was a step forward, but it did little to advance a more robust definition of PII, instead focusing on industry-specific data—i.e., whatever is entered on an application.

Let's refer to the most obvious PII suspects – name, credit card number, social security and bank account--as classic PII. It turns out the classics are still the most sought after identifiers by cyber thieves. Verizon's Data Breach Investigations Report is a good reference for breach incidents. The DBIR team has pointed out that in recent years--even with new Internet-based identifiers-- these traditional PIIs attract the lion's share of attention, with payment card data as their primary target[3]. Hackers and cyber thieves still see credit card data as being the easiest to monetize.

MOST WANTED PII (DBIR 2012)



With the rise of the Internet and online social interactions, new questions about PII began to emerge.

New and Improved Internet-Era PII

In the late 90s, both EU and US regulators realized that digital identifiers—email addresses, chat names, URLs, and other computer-oriented IDs—extended the meaning of PII beyond the classics. The underpinnings for this new view were based on work originally done the by Organization of Economic Cooperation and Development, a policy research organization established by the US and European countries after the war years.

One area the OECD looked into was privacy. In a landmark guidelines document, now almost thirty years old, these early privacy researchers defined "personal data"--their name for PII--as "any information relating to an identified or identifiable individual."[4]The key point is that the definition clearly encompassed what anyone would consider classic name-address identifiers, but was abstract enough not to get bogged down in industry-specific terms--like what was done in the US with GLBA and, another early privacy law, the Fair Credit Reporting Act.

This particular way of looking at PII was ultimately adopted by the European Union in their influential Data Protection Directive of 1996, even to the point of including the same OECD terminology—personal data, data controllers, data processors--from the original OECD document.

EU member nations were required to use the DPD as a template for their own data protection laws and to establish individual data protection authorities to enforce the rules. Regulators have used the DPD's flexible definition of personal data to encompass IP addresses (Austria and Italy), digital pictures (Denmark), and even DNA (Estonia, Czech Republic).[5]

In the US, the OECD view of personal data was a little slower to take hold but these ideas eventually made their way into some regulations, most notably in consumer medical data, which we take up next.

Quasi-identifiers and a Broader Definition of PII

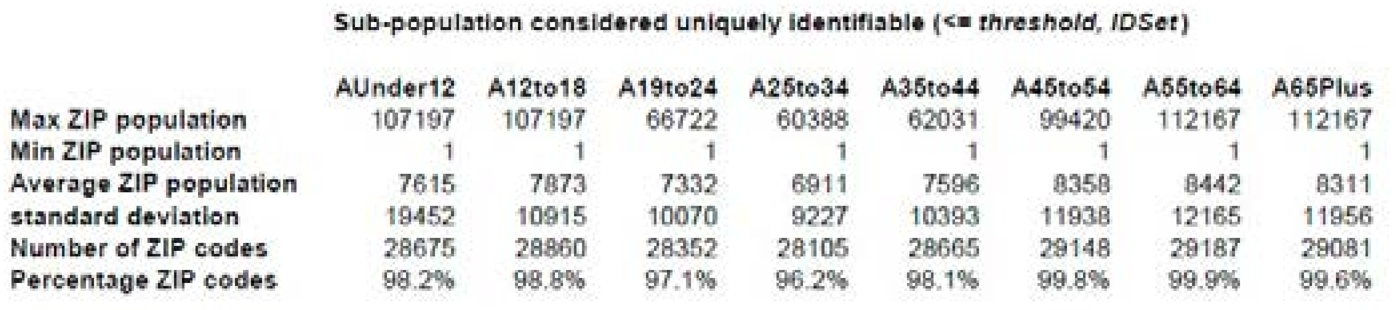
In a well-publicized incident in 1998, MIT graduate student Latanya Sweeney managed to identify the medical condition of the then governor of Massachusetts William Weld from an "anonymous" medical data released by the Veterans' Administration[6]. Matching three quasi-identifiers-- zip code, full birthdate, and gender-- in the medical records to public voting rolls, Sweeney was able to re-identify Weld, along with determining his diagnosis and prescriptions.

The real validation of this type of hacking attack came from Sweeney herself, who statistically analyzed this particular re-identification problem. Using census data (broken down by zip codes and age groups), she was able to prove it was possible to identify 87% of the people in the US working with just those three quasi-PIIs[7].

Sweeney's main discovery was that consumer data can contain other fields, which are not strictly identifiers, but when taken together effectively act as PIIs. While quasi-identifiers are somewhat accommodated by the EU's DPD—although a new update will more clearly take quasi PIIs into account (see below)—US laws lagged behind.

Some of these more sophisticated ideas on PII did make their way to US policy makers at the Department of Health and Human Services. The Health Insurance Portability and Accessibility Act (HIPAA) of 1996 contained a definition of PII familiar to regulators at the EU and other countries with data privacy laws based on the OECD's privacy guidelines (for example Japan and Singapore).

HIPAA defined protected health information or PHI—effectively PII--as any data that “identifies the individual or for which there is a reasonable basis to believe it can be used to identify the individual along with any sensitive health data”. The important words in this definition are “reasonable basis”, which opens the possibility for indirect links to individuals through quasi-identifiers and Internet-style identifiers, as well as identifiers yet to come.



Sub-population considered uniquely identifiable (<= *threshold, IDSet*)

	AUnder12	A12to18	A19to24	A25to34	A35to44	A45to54	A55to64	A65Plus
Max ZIP population	107197	107197	66722	60388	62031	99420	112167	112167
Min ZIP population	1	1	1	1	1	1	1	1
Average ZIP population	7615	7873	7332	6911	7596	8358	8442	8311
standard deviation	19452	10915	10070	9227	10393	11938	12165	11956
Number of ZIP codes	28675	28860	28352	28105	28665	29148	29187	29081
Percentage ZIP codes	98.2%	98.8%	97.1%	96.2%	98.1%	99.8%	99.9%	99.6%

With the good aspects of this type of definition—abstraction and flexibility—come less than desirable operational challenges: IT departments at hospitals and HMOS would prefer to follow a checklist of specific identifiers that need to be protected. In response to compliance uncertainty, the regulators came up with the Safe Harbor De-identification rules-- essentially a list of 18 identifiers.

The list covers the classics, Internet-era identifiers, biometric, quasi-identifiers, at least as they relate to zip codes, and semantic identifiers (see chart). In recognition of Sweeney’s work, the Safe Harbor rule says that no geographic unit smaller than a state can be included in publicly released patient data, and full dates (e.g., admission, birth, etc.) must also have the year removed.

Are there other quasi-PII’s out there? Of course. Location and timestamp information is an important ingredient in this new breed of quasi-identifiers. For example, GPS coordinates are routinely collected by mobile devices. This raw data could be matched up with public check-ins from social media sites (for example, Foursquare) to re-identify the owner of the device.

Role of Social Media and PII

The larger problem is that consumers are sharing detailed information about themselves on web sites and online social forums. In a possible scenario, an online retailer collects preference data about its customers—interests, hobbies, web clicks, etc.—along with geographic data and then sells the dataset, stripped of classic PII, to a third-party marketing company. This is not an entirely hypothetical monetization strategy since these types of transactions have been making headlines.[8]

If the marketing company has the computing resources, it could potentially re-identify the records by matching the preference information and geo data with personal data collected by crawling websites. While not violating any specific law, the company then targets these customers with highly specific messages.

Is this re-identification approach using preference data only theoretical or can it can be successfully performed in the real-world?

Some may remember that in 2006, Netflix, the movie rental service, announced a public contest to improve on its existing algorithms for suggesting new films to subscribers. To give contestants something to work with, Netflix released an enormous data set of de-identified movie ratings from their database–essentially long rows of numbers indicating a Netflix subscriber's 1- 5 evaluation of titles in the Netflix inventory.

Two University of Texas researchers analyzed the public Netflix data, not to enter the contest but instead to see if they could re-identify Netflix users[9]. Their strategy was to compare the rows of data from Netflix against separate ratings submitted by subscribers to IMDb, the popular movie information site. The researchers succeeded: identifying the full preferences of two users with very high confidence.

In other words, by scanning the social networking component of a site where community members reveal only a small set of their movie ratings–say, for 6 movies–the researchers were able (using a straight-forward algorithm) to identify specific users in the Netflix data set, and their complete movie likes and dislikes.

In another re-identification effort, an NYU-Poly computer science professor was able to exploit the social connections in Facebook. The algorithms developed made successful inferences about Facebook subscribers who had private profiles by examining the public profiles of direct friends, and even friends of friends.[10]

These re-identification techniques are based on standard methodologies from data mining, involving clustering or segmentation of larger populations into much smaller more manageable groups, and then finding matches with the highest likelihood. The success of these real-world examples has had impact on regulators (see below), especially in term of extending the scope of PII to cover the possibility that one can using common computing resources and other "reasonable" methods to link data back to a specific individual.

Semantic and Other Bleeding Edge PII

We've looked at several varieties of PII that all share a key similarity: they contain at their core something that looks like a keyword, either a unique code or, as discussed in the previous section, a preference rating or category. Is it possible for a PII to be a string of words—a sentence or paragraph?

Let's call this new type of PII, a semantic identifier. The idea is that simple textual descriptions about a person can contain enough meaning so that one can, with the help of a search engine, effectively re-identity the data.

The easiest such semantic identifier to see is a sentence containing a job title, "The patient is a vice president of sales at Varonis, a software governance company." This is not a classic PII or quasi-identifier, but it contains enough information to connect it to a unique individual. If you'd like to see this for yourself, try entering this sentence into Google.

There are other examples of semantic identifiers: along with job titles, descriptions involving locations and other personal details may be enough to re-identify or at least greatly narrow down the possible subject. This leads to a broader discussion of what's called the "semantic web". In brief, Google and a few others are already doing leading edge work on extracting meaning and knowledge from web content and have started introducing semantic searches. Very soon search engines will be used to make non-obvious inferences from the crawled data and then provide answers.

Will Google and other social web sites be able to one day answer a question, "Who was the man in the white shirt that I saw at Starbuck's yesterday at 10am?" With a combination of geo-data, image recognition, and, of course, a rich source of social profiles, it may be possible.

Besides semantic searches, there are still other bleeding edge technologies that push the boundaries of PII.

It's now possible to use non-proprietary software and hardware to pull facial information out of digital images. It's already being done on a commercial basis. Retailers have installed digital signage in mall kiosks to serve up ads based on the gender, age, and other demographic information of the consumers viewing the informational screens.

Even more impressive is that current facial recognition technology has reached a high-level of accuracy in comparing photos to find matches. The National Institute of Standards and Technology (NIST) reports that the false reject rate–percentage of comparisons incorrectly rejected–has been cut in half every two years. As of 2010, this important photo mis-match metric now stands below 1%.[11]

In 2011, Alessandro Acquisti, a researcher at Carnegie-Mellon University, took advantage of these rapid improvements in facial image matching technology. He was able to re-identify anonymous photos from a dating site as well as random students walking around campus by using off-the-shelf facial identifiers, cloud computing, and social media sites.[12] As with the quasi-identifiers mentioned earlier, location information was important in narrowing down the searches and a rich source of public data-- facial images in Facebook profiles-- was the final link in re-identifying the anonymous data.

EVOLVING REGULATIONS ON PII

In general, US laws and regulations have not kept up with the new breed of identifiers. One exception, mentioned previously, is HIPAA, whose more robust definition of PII gave regulators the flexibility to include a broad range of identifiers. However, policy makers at the Federal Trade Commission, an important privacy and data protection regulator, have been studying the problems of privacy, anonymity, and re-identification.

In Protecting Consumer Privacy in an Era of Rapid Change, published in 2012, the FTC acknowledged the limited scope of PII in protecting consumer privacy[13] . Their new policy recommendations are informed by recent academic work—the Netflix de-identification and Acquisti's facial imaging studies—as well as the OECD's overall framework on privacy. The report also recognizes that the difference between classic PII and anonymous data has been "blurred" in recent years.

The FTC is proposing a policy to secure any "consumer data that can be reasonably linked to a specific consumer, computer, or other device." As with the definition of HIPAA's PHI , the important wording, again, is "reasonably linked" , which now places all the new PII forms discussed in this paper, regardless of the industry sector, under more equal footing with classic PII.

The Protecting Consumer Privacy report does not have the power of law and the FTC is asking for voluntary best privacy practices—"privacy by design—as well as more careful analysis by companies of their data anonymizing techniques—i.e., they should assure a "reasonable level of confidence" that published data about consumers can't be linked back to an individual.

The primary focus, for now, of the FTC regulators are "data brokers", who have proved adept at linking data scattered across the Internet to individuals in order to create large unregulated databases of consumer profiles. The FTC has begun conducting investigations of these brokers, and has suggested that Congress pass new laws to police this new industry--similar to what the FCRA has done for credit reporting agencies.[14]

However, the FTC still has considerable room to interpret current laws under their new privacy framework. For example, they've issued a series of complaints and extracted heavy fines against several social media companies for not living up to their advertised privacy terms[15] . If a company doesn't honor the privacy promises in its terms of service or other advertising, they're violating the FTC's deceptive business practice regulation. Armed with a more expansive view of PII, the FTC has additional leeway in enforcing this law.

It's not only the US that has included the reasonableness wording into the definition of PII. The EU is currently in the process of changing its Data Protection Directive. The goal is to create a uniform set of regulations across all EU nations along with a centralized authority to enforce the rules. EU regulators have also noted there is no longer a strict separation between PII[16] and other data. In their new definition for personal data, EU regulators have revised the language so it allows for any data that can "directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person" to be considered a link to an individual.

Finally, recent legislation proposed by the White House, called The Consumer Privacy Bill of Rights, is based heavily on the findings of the FTC.[17] It incorporate the FTC's "reasonably linked" wording in its definition of PII and would extend data privacy regulations to any company—not just in medical or financial areas—that collects consumer data.

CONCLUSION: PROTECTING PII

With a broader view of PII now taking hold, privacy and security risks for organizations are likely to increase in the coming years. One of the largest sources of unprotected PII are the human generated files found in corporate files systems. Think of all the identifiers--especially quasi and semantic, along with the classics—that are likely to be found in loosely permissioned documents, presentations, and spreadsheets scattered across a large corporate file system.

To identify these PIIs, IT departments will need to effectively deploy a file system crawler that searches clear-text content for patterns by using powerful regular expression engines and perhaps techniques from the world of machine learning.

There nothing's necessarily wrong in finding PIIs in unstructured files—after all, corporate file systems are where workers do most of the collaboration and content sharing. The key challenge is to make sure the files are properly permissioned, that only authorized users are accessing the PII, and their use is monitored for potential abuse.

To implement an effective approach to managing PII and reducing breach risks, it's important that the data owner or subject expert—the ones in the company who have the deepest knowledge of the data—to be part of any process.

Forrester Research suggests a methodology known as PICWIC--protecting information consistently with identity context .The central idea of PICWIC is that you should assign file data to business owners at all times. That way you can properly permission new user accounts with correct levels of access, recertify access entitlements regularly, and take the appropriate actions when an employee changes roles or is terminated. By following these best practices, you will reduce the chances of accidental data leakage of PII.

As a final measure, IT departments also need to continuously monitor their file systems, especially those directories known to contain PII. No data protection system is perfect, so putting in place detection controls that spot unusual activities or other patterns consistent with a breach will ultimately reduce company liabilities.

HOW VARONIS CAN HELP

DISCOVER AND PROTECT WHAT'S SENSITIVE

The Varonis IDU Classification Framework is the only solution that identifies the highest concentrations of sensitive data that are most at risk and provides a clear methodology to safely remediate that risk without manual effort.

A built-in report shows you a prioritized list of folders that contain the most sensitive data and the most exposed—through global access groups (Everyone, Authenticated Users, etc.) and/or normal groups that contain too many members. Other metrics can be used to prioritize remediation, including activity, size of files, and density of files.

Ensure only the right people have access to data

DatAdvantage gives you a consistent view of permissions across Windows, NAS, UNIX/Linux, Exchange, and SharePoint. With Varonis you can:

- Find and remediate data is that sensitive and overexposed
- Model permissions changes in a sandbox before executing
- Provide data owners intelligent recommendations on where to reduce access to their data
- Clean up unused users and groups
- And much more!

BIG DATA SECURITY ANALYTICS

Varonis logs every user's activity across your entire environment, and uses bidirectional cluster analysis and machine learning to predict which permissions they really need. This helps you eliminate risk and remain compliant.

MONITOR USE, ALERT ON ABUSE

More than 95% of file access activity is not monitored by IT because native auditing is slow and hard to use. With Varonis' sortable and searchable audit trail, you always know who is touching important business data. What's more, you can setup alerts whenever abnormal access activity or privilege escalations happen.

GET RESPONSIBILITY FOR DATA OUT OF IT – SUSTAINABLE SECURITY

DatAdvantage can see who is actually accessing data, so it can lead IT right to the appropriate business owner and get them timely information about their data. DataPrivilege makes it painless for business users to review and authorize access. The end result is the right people, with the right information, making the right decisions.

[1] *Thirty Years After the OECD Privacy Guidelines (oecd.org)*

[2] *Privacy Act of 1974 (ftc.gov)*

[3] *2013 Data Breach Investigations Report (verizon.com)*

[4] *OECD Guidelines on the Protection of Privacy and Transborder Flows of Information (oecd.org)*

[5] *Article 29 Working Group Annex to Impact Assessment of Data Protection Regulation (europa.org)*

[6] *Exposed :The erosion of privacy in a digital era (harvardmagazine.com)*

[7] *Simple Demographics Often Identify People (dataprivacylab.org)*

[8] *ATT Will Start Selling Consumer data (gigaom.com)*

[9] *Netflix Cancels Contest Plans and Settles Suit (nytimes.com)*

[10] *Interview with NYU-Poly's Professor Ross (poly.edu)*

[11] *Face Recognition Grand Challenge (nist.gov)*

[12] *More than facial recognition (cmu.edu)*

[13] *Protecting Consumer Privacy in an Era of Rapid Change (ftc.gov)*

[14] *FTC to Study Data Broker Industry's Collection and Use of Consumer Data (ftc.gov)*

[15] *Facebook Settles FTC Charges That It Deceives Customers (ftc.gov)EU Data Protection Regulation (europa.eu)*

[16] *EU Data Protection Regulation (europa.eu)*

[17] *Consumer Data Privacy in a Networked World (whitehouse.gov)*

[18] *"Your Data Protection Strategy Will Fail Without Strong Identity Context," Forrester, July 29, 2011*

ABOUT VARONIS

Varonis is the leader in unstructured and semi-structured data governance software. Based on patented technology and a highly accurate analytics engine, Varonis solutions give organizations total visibility and control over their data, ensuring that only the right users have access to the right data at all times from all devices, all use is monitored, and abuse is flagged.

Varonis makes digital collaboration secure, effortless and efficient so that people can create and share content easily with whom they must, and organizations can be confident their content is protected and managed efficiently.

Free 30-day assessment:

WITHIN HOURS OF INSTALLATION

You can instantly conduct a permissions audit: File and folder access permissions and how those map to specific users and groups. You can even generate reports.

WITHIN A DAY OF INSTALLATION

Varonis DatAdvantage will begin to show you which users are accessing the data, and how.

WITHIN 3 WEEKS OF INSTALLATION

Varonis DatAdvantage will actually make highly reliable recommendations about how to limit access to files and folders to just those users who need it for their jobs.

WORLDWIDE HEADQUARTERS

1250 Broadway, 31st Floor, New York, NY 10001 **T** 877-292-8767 **E** sales@varonis.com

UNITED KINGDOM AND IRELAND

Varonis UK Ltd. Warnford Court 29 Throgmorton Street London, UK EC2N 2AT **T** 020 3402 6044 **E** sales-uk@varonis.com

WESTERN EUROPE

Varonis France SAS 4, rue Villaret de Joyeuse 75017 Paris France **T** +33 (0)1.82.88.90.96 **E** sales-france@varonis.com

GERMANY, AUSTRIA AND SWITZERLAND

Varonis Deutschland GmbH Robert Bosch Strasse 7 64293 Darmstadt **T** + 49-0-6257 9639728 **E** sales-germany@varonis.com