Whitepaper

ContactEngine

# Peek-a-boo! They still see you. The deficiencies of de-identified data.

# Introduction

De-identification and re-identification of consumer data has come to the fore recently to further emphasize the mountain of work that remains in truly protecting our privacy as individuals. A July 2019 research paper claims that 99.98% of Americans could be re-identified from an anonymized dataset, if certain demographic information was available[1]. This means that even without massive data breaches like the Capital One hack of 100 million customers revealed in late July 2019[2], customers' data (financial or otherwise) is far from secure. While initiatives like PSD2 and Open Banking may be important to a customer's financial education to learn about spending patterns and budgets, this needs to be complemented with general awareness on how to protect themselves in the digital arena as well.

1. https://www.nature.com/articles/s41467-019-10933-3
2. https://edition.cnn.com/2019/07/29/business/capital-one-data-breach/index.html

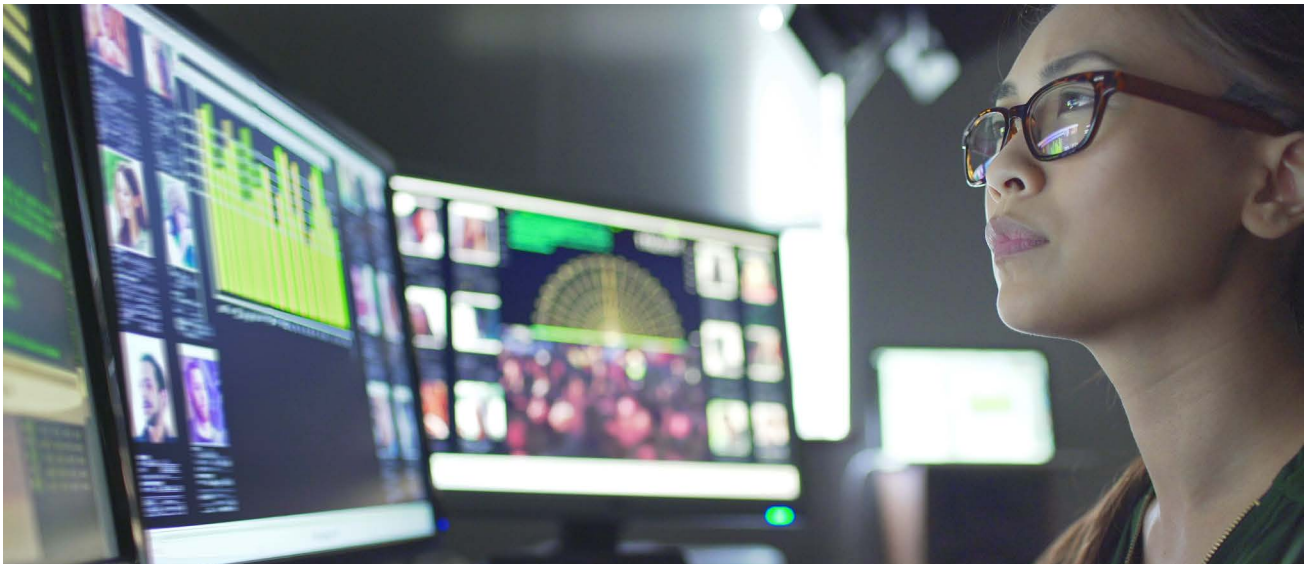
## What is de-identification and re-identification?

Data de-identification is "the process of anonymizing datasets before sharing them[3]," i.e. removing any sort of personally identifiable information (PII) and stripping data down to the bare minimum. Having done this, companies are able to process, as well as share, this information given that it is theoretically not traceable back to a particular individual, thus protecting their privacy.

As a corollary then, re-identification is the process of using certain demographic characteristics, the combination of which may be fairly unique, and matching those to characteristics in anonymized data. According to Rocher, Hendrickx and de Montjoye (2019), even the presence of as few as 15 demographic attributes can help re-identify individuals. Data brokers purchase and sell access to thousands of such attributes. Take for example Oracle, which supposedly has access to 30,000 attributes per individual for over 300 million people across the globe[4].

And this practice of re-identification is not a new phenomenon. As early as the mid-90's, researchers have proved that ostensibly anonymized data can be used to re-engineer a person's identity. This was particularly relevant from a healthcare perspective and the introduction of the Health Insurance Portability and Accountability Act (HIPAA) introduced in 1996, geared partly towards ensuring the privacy of an individual's medical records. In 1997, Dr. Latanya Sweeney managed to use William Weld's (then Governor of Massachusetts) demographic information to identify his anonymous medical records[5].

The context may be different 22 years later, but the roots remain the same. The quantum of information that consumers are compiling against their names/identities has risen sharply and is ever growing. This increases the odds of re-identification of anonymous records that much further.



3. https://www.nature.com/articles/s41467-019-10933-3
4. https://www.ft.com/content/f1590694-fe68-11e8-aebf-99e208d3e521
5. https://techscience.org/a/2015092903/

## Current Regulation

Specific to the banking industry, as the world continues to shift more towards digital transactions, it does mean that people are trading off a certain level of privacy for ease of access and convenience. For instance, "[b]ranch transactions account for only 12% and 21% of total monthly transactions in Developed Asia and Emerging Asia respectively[6]". While consumers may still want the physical interaction in a branch for more complex transactions, this does not take away from the fact that simple account-checking and payments are being done online and contributing to one's digital presence. And even branch transactions are stored digitally by banks as part of their portfolio on the customer.

The implementation of regulation such as Europe's second Payment Services Directive (PSD2) further exacerbates the issue of easier access to customer financial data with its focus on "[e]nablement of third-party access to account information[7]" (among other things). For the UK specifically, this has led to the establishment of the Open Banking standard by the Competition and Markets Authority (CMA) which is essentially mandating that the nine largest banks allow access to customer data via open APIs[8]. Though it stems from a desire to increase competition, innovation and customer control of their overall finances, PSD2 does make it easier for third-party providers (TPPs) to gain access to a person's financial history. It is true that these entities are regulated and tested in a virtual sandbox prior to gaining access to the personal information, but the danger of a data breach is perennial.

There is much debate about whether this legislation conflicts with the General Data Protection Regulation (GDPR), also introduced across the EU in 2018. GDPR is aimed at putting consumers in charge of controlling who has access to their data, requiring consent before data can be collected. Most of us have experienced this in some form by now when we readily click on cookie acceptances on the various websites we visit. As such, while companies may be meeting minimum compliance requirements in order to avoid fines and litigation, the goal of actually informing people about where their data is being used and to whom it's being further distributed seems lost in the fray. Furthermore, if consumer data is anonymized, it is no longer protected under the remit of GDPR[9]. Therefore, once anonymized, data can be repackaged and distributed/sold without repercussion. As stated above, this data can conceivably then be married with other known attributes about individuals to re-identify people from it.

6. https://www.mckinsey.com/~/media/McKinsey/Industries/Financial%20Services/Our%20Insights/Reaching%20Asias%20digital%20banking%20customers/Asias-digital-banking-race-WEB-FINAL.ashx
7. https://www.barclaycard.co.uk/business/news-and-insights/what-is-psd2
8. https://www2.deloitte.com/content/dam/Deloitte/cz/Documents/financial-services/cz-open-banking-and-psd2.pdf
9. https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/anonymisation-and#Anonymisation


**Regulations with conflicting goals**

| PSD2 | GDPR |
|---|---|
| **Banks are required to open customer accounts and transaction data to third party providers (TPP)** | *Requires banks to protect customer data and imposes significant penalties for failure to comply (up to 4% of global annual revenue)* |
| **Encourages an open banking environment** | *With the stringent requirements listed in the regulation it might make open-banking implementation less attractive* |
| **Does not require informing customers of their right to withdraw consent** | *Necessitates that customers be informed of their right to withdraw consent* |
| **Data processing and sharing needs to be explicitly requested/approved by the customer** | *Does not require consent of the customer to further share their data (including anonymized data)* |
| **Consent of the customer expires automatically** | *Under GDPR, consent does not expire automatically, and requires customer permission to be removed* |

**Sources: (1) How banks can balance GDPR and PSD2[10] and (2) Harmonization needed among conflicting key regulatory and industry initiatives to enable new payments' ecosystem functionality[11]**

Even blockchain technology, which was touted for its ability to provide increased security because of its decentralised transactions, is susceptible to re-identification. Research has shown that data brokers "typically possess enough information about the purchase to uniquely identify the transaction on the blockchain, link it to the er's cookie, and further to the user's real identity."[10] There are additional technologies layering privacy tools on top of blockchain technology (such as Monero or Zcash), but they do not currently have the network effects to be claimed as viable alternatives.

10. https://www.ey.com/en_gl/banking-capital-markets/how-banks-can-balance-gdpr-and-psd2

## Why does this matter?

An increasing number of high-profile data breaches have been coming to light over the last decade. The impact of events like the aforementioned Capital One data breach or the Equifax breach from 2017 is often lost on individuals unless immediately and directly affected. We hear of the effect on the companies such as costs to mitigate (~$150MM in costs related to the hack for Capital One[11]) or fines to compensate ($700MM for Equifax[12]), as well as the deterioration of the reputation of the company (Capital One stock down went 5% immediately following the hack[10]) and the loss of trust in the company (and hence loss of business). But rarely is the trail followed all the way through to tangible impact on the individual customer.

**High profile data breaches over the last decade**

| Year | Company | Individuals impacted |
|---|---|---|
| **2019** | Capital One | 100MM |
| **2018** | Marriot | 500MM |
| **2017** | Equifax | 143MM |
| **2016** | Adult Friendfinder | 412.2MM |
| **2015** | Anthem | 78.8MM |
| **2014** | eBay | 145MM |
| **2014** | JP Morgan Chase | 76MM |
| **2014** | Home Depot | 56MM |
| **2013** | Yahoo | 3000MM or 3BN |
| **2013** | Target | 110MM |
| **2013** | Adobe | 38MM |
| **2012** | US Office of Personnel Management (OPM) | 22MM |
| **2011** | Sony's PlayStation Network | 77MM |
| **2011** | RSA Security | 40MM |
| **2008** | Heartland Payment Systems | 134MM |

**Sources: (1) How banks can balance GDPR and PSD210 and (2) Harmonization needed among conflicting key regulatory and industry initiatives Source: The 18 biggest data breaches of the 21st century[15]**

11. https://worldpaymentsreport.com/2019/02/harmonization-needed-among-conflicting-key-regulatory-and-industry-initiatives-to-enable-new-payments-ecosystem-functionality/
12. https://www.bbc.co.uk/news/technology-49070596

The consequences of the loss of information depends largely on the type of data. Personal information such as email addresses, driver license numbers, credit history etc. can lead to phishing attacks and identity theft. Social media data such as likes, browsing history etc. can be used anywhere from targeted marketing campaigns to determining ideological leanings and either reinforce or deter biases, thus affecting the political process. Location history can be used for everything up to and including stalking and physical danger. And credit card information, spending patterns, purchases can cause direct financial loss to customers.

But even if all these breaches are addressed and protected against going forward, and even if the banking industry manages to earn the trust of their consumers, the larger point remains that despite the strictest of regulations, hackers can still get at private information from completely anonymized data.

The only way to try and guard against this is for people to know that it's a possibility and thus potentially limit the number of places where they reveal their data.

## Cybersecurity education for consumers

Most people learn early that touching a hot stove leads to burnt fingers and that the associated pain should be avoided in the future. However, since the consequences of online actions are often delayed or seemingly unrelated to recallable actions, it's not easy to learn the negative cause and effect chain. The most logical solution then would be to educate consumers as much as possible.

Policymakers are starting to find ways to attack this problem, though it does seem like most of the current attempts would be more short-term or quick fixes. For instance, there is ongoing debate about the establishment of a National Data Broker Registry in the United States[13], which would essentially collate a list of the "companies that collect information about consumers who are not their customers." The idea behind this being that consumers can have an objective ranking of who is rightfully or wrongfully using their data. While this may be somewhat of a half measure that addresses one specific aspect of data privacy, it is a helpful interim step while we wait for consumer knowledge and attitude about data privacy to catch up to the extant dangers.

Furthermore, training around social engineering is increasingly mandated at most organizations due largely to the number of high-profile security breaches in the recent past at say Yahoo or Dropbox (Aldawood and Skinner, 2019)[14]. While there is no obligation from a corporate perspective to do so, this training could be amended with aspects of guarding against such attacks in personal situations as well. It stands to reason that looking at awareness outside of the silo of work would make for a more aware and engaged individual, which provides a benefit for the organization in the form of a more rounded employee. This may only help a specific segment of the population, but again is an easy add-on to current infrastructure.

13. https://www.nytimes.com/2019/09/13/opinion/data-broker-registry-privacy.html
14. https://www.mdpi.com/1999-5903/11/3/73/htm

For the long-term, researchers have begun to think about implementing the above either by changing or supplementing the current curriculum at the school and/or college level. Egelman et al. (2015) claim their curriculum has impacted the data privacy attitudes of students in a pilot program[15]. Additionally, certain schools in New York have also tested experimental curriculum with the goal of having children more informed about their online presence and thus growing up more informed[16]. While it is useful to have children exposed to these concepts at a young age, there need to be concurrent efforts aimed towards adults and seniors as the problem is indiscriminate in its reach. With the proliferation of Massive Open Online Courses (MOOCs), this goal is very achievable with even a modicum of impetus.

There may be vehement disagreement over who bears the onus for educating customers; whether it is the responsibility of the policymakers, the companies collecting the data, the third-party data brokers, or the customers themselves. Given that the latter option does not seem to be bearing fruit, other avenues of informing consumers must be investigated to combat the widening breadth of this problem.

The drive towards digitizing our personal and financial information seems inevitable in our current environment. While work can and must be done to safeguard the use of our information by companies when enacting legislation like PSD2 and Open Banking, we are in an era where this regulation can be sidestepped using (i) loopholes in the language and/or (ii) advances in technology that leapfrog the protective laws. Yes, privacy policies need to be more digestible, and we shouldn't be held to ransom by having to click on cookies every time we enter a site, but there is more we can do to be personally informed so that we give it a second thought when we do come across these questions. Thus, having a more informed populace seems the only realistic arrow remaining in the quiver.

15. https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html
16. https://www.cbsnews.com/news/young-students-learn-the-importance-of-protecting-online-privacy/


**Visit** contactengine.com
**Follow** @contactengine
**Email** info@contactengine.com
**Contact** +44 20 33 940 840

## About

The report is published by ContactEngine Limited

Registered Office: The Clergy House, Mark Square, London EC2A 4ER

ContactEngine is the next generation Customer Engagement Hub technology that enables brands to proactively engage customers in AI-driven conversations to fulfil business objectives. ContactEngine automates outbound customer engagement across all channels and generates unique insights into the changing patterns of communication by applying demographic and intent analysis, linguistics and ground-breaking artificial intelligence principles to mans volumes of raw data. ContactEngine transforms the way global brands engage with their customers - saving brands millions and making their customers happier.

For more information, visit www.contactengine.com