Massive Data Collection by Mistake?

Arnold Roosendaal

Massive Data Collection by Mistake?

Arnold Roosendaal LLM MPhil

Tilburg Institute for Law, Technology, and Society (TILT)
Tilburg University, PO Box 90153, 5000 LE Tilburg
The Netherlands

A.P.C.Roosendaal@tilburguniversity.edu

Abstract. Recently, three major ICT companies were confronted with public outrage about the way they collected massive amounts of personal data without informing data subjects, let alone obtaining their consent. Google harvested data concerning Wi-Fi routers while cruising around with their StreetView camera cars, Facebook tracked potentially every internet user with the help of tracking cookies and the 'Like' button, and Apple collected and stored location data from iPhones. In all three cases the companies stated that it was a mistake, sometimes took the blame, fixed the issue, and continued their work. The central question is whether they were really mistakes and why the companies could continue their businesses without major problems. Analysis of the three cases leads to hypotheses on whether they were mistakes or a strategy, and signals a trend towards increasing privacy breaches by powerful companies.

Keywords: Data collection, Privacy, Accountability, Google, Facebook, Apple

1 Introduction

Over the past year, three major ICT companies came in the press with regard to extensive collection of personal data. First, in April 2010, there was Google, of which it became known that their camera cars used to take pictures for StreetView also captured private information sent over unencrypted routers. Then, in November 2010 it was brought to light that Facebook was using tracking cookies, which allowed the company to track and trace members as well as non-members of their social network site every time a website displaying the 'Like' button was visited. Finally, in April 2011, Apple appeared to have collected and stored location information of iPhones and iPads. In all three cases, outsiders found out about the data collection and made the practices public, instead of the companies themselves. And, in all three cases, the companies plead innocent, at least to some extent. Either they admitted to have collected the data, but only as the unfortunate result of a bug, or they denied until being confronted with evidence of the opposite and then claimed it to be the result of a bug or of a mistake being made in the settings of the company's software. Nevertheless, there are different approaches of the companies in dealing with issues. Apple reacts by posting a Question and Answer and by releasing a software update to fix the issue.

adfa, p. 1, 2011.

Google reacts in a blog post and contacts data protection authorities to discuss how to delete the mistakenly collected data as soon as possible. But Facebook never reacted in public, except from a denial of the practice, and only confirmed the findings in an unpublished communication with a German data protection authority (Hamburg) who had started investigations concerning the practice.

The companies stated not to have used the data they inadvertently collected, simply because they did not even know they were collecting the data and because they were not interested in the data anyway. Nevertheless, the software and systems of the companies are designed in such a way that the collection of the data is made possible. The question that comes to mind is how it is possible that ICT companies of this size make these kinds of mistakes and do not even notice the mistakes themselves. Clearly, the companies have ICT developers who are among the most qualified in the world. Are the mistakes made in the development process, or did something go wrong in the management of the companies responsible for launching the products or features? Are they really mistakes or did the companies consciously collect and store the data, hoping that it would not be discovered? And, if it would be discovered, blame it to a bug or mistake and hope to get away with it fairly easily? Or is ICT development so complex and is high speed development required to keep pace, implying that it is impossible to be completely responsible for and aware of the features of a product a company develops? In any case, since this happened to three of the most important players in the field within a year, somewhere there is a problem. Thus, it is important to analyze what went wrong and how to prevent this from happening in the future.

In this paper, subsequently the case of Google (2), Facebook (3), and Apple (4) will be analyzed. In section 5, the three cases will be compared in order to find important similarities and differences. Finally, in section 6, an assessment is made of the likeliness of the non-compliance with data protection laws and an indication is given of a trend to infringe upon, and step by step diminish, privacy. From the discussions later on the indications are that the privacy infringements are not a mistake but a strategy.

2 Google StreetView

In Google Maps it is possible to view panoramic images of city streets, with a service called Google StreetView. In order to compile these images, Google has a fleet of vehicles, equipped with special cameras, which they drive around. Google also intended to record the identity and position of Wi-Fi hotspots in order to power a location service it operates.[1] The position of the vehicle, and thus the image, could be defined accurately by using triangulation within these networks. The idea was to collect network data like SSID information (the name of the network) and MAC addresses (unique numbers given to devices such as routers) in order to identify and locate the networks. These data could be used to improve Google's location services, such as Google Maps. The German data protection authority (DPA) in Hamburg raised some concerns over the Wi-Fi data that were collected, which prompted

Google to publish a blog discussing information collected by the StreetView cars and the purposes of this collection.[1] Google stated to collect only SSIDs and MAC addresses. However, the DPA in Hamburg asked Google to audit the Wi-Fi data which led to a discovery by Google that the earlier statement was incorrect and that payload data (information sent over the network) of open Wi-Fi networks was collected as well.[2] Google claimed this to be a mistake due to a piece of code that was included in the software, "although the project leaders did not want, and had no intention of using, payload data."[3] Google stated it was profoundly sorry for the error and took steps to delete the data immediately in cooperation with regulators. It also had an independent third party perform a check on the software and the data it collected.[2] In the meanwhile, Google is faced with a number of legal procedures and fines from DPAs. During the investigations by the *College Bescherming Persoonsgegevens* (the Dutch DPA) it appeared that Google had collected almost 30 GB of payload data in about two years time in the Netherlands.[3]

Given the continuous stream of privacy issues Google is involved in, it is not surprising that many people were suspicious about Google's explanation. However, there are also claims that it was "almost certainly an accident", because the data packages are so small and fragmented that they are relatively useless; there is no evidence of the data being used by Google, and; there is no explanation of what Google would want with the payload data.[4] Still, it is admitted that, even when it was a mistake, Google should have realized it much earlier and never allowed such data to be captured. In the words of the US Federal Trade Commission: "the company did not discover that it had been collecting payload data until it responded to a request for information from a data protection authority. This indicates that Google's internal review processes - both prior to the initiation of the project to collect data about wireless access points and after its launch - were not adequate to discover that the software would be collecting payload data, which was not necessary to fulfill the project's business purpose."[5]

3 The Facebook 'Like' Button

A second case that received considerable attention concerns the 'Like' button as exploited by Facebook. This button, a thumbs-up symbol which can be clicked to let Facebook members share things they like with their friends, is displayed on more than 2,5 million websites. In November 2010, research revealed that the button facilitated Facebook with the opportunity to track and trace potentially every internet user via this button, combined with Facebook Connect, regardless of whether someone clicked the button or not and regardless of whether someone was a member of Facebook or not.[6] On the basis of this, Facebook could create individual profiles of browsing behavior and interests.

---

[1] See: <http://googlepolicyeurope.blogspot.com/2010/04/data-collected-by-google-cars.html>.
[2] See: <http://googleblog.blogspot.com/2010/05/wifi-data-collection-update.html>.
[3] See: <http://googleblog.blogspot.com/2010/05/wifi-data-collection-update.html>.

There was no official public reaction by Facebook, but in personal communications the practice was denied. Nevertheless, the research triggered a number of authorities to start investigations. After being confronted directly with the findings of the research, Facebook admitted the extensive data collection in a communication to the German DPA in Hamburg.[4] Facebook confirmed the findings but claimed the tracking activities to be the result of a bug in a software development kit (SDK). In addition, Facebook stated that it changed the software as soon as they became aware of it.[5] Nevertheless, even after this admission, Facebook denies the tracking possibilities in a reaction to class action complaint based on the research.[6]

In Germany, even stronger objection towards Facebook's social plugins came from the DPA in Schleswig-Holstein. This DPA also performed an investigation, specifically aimed at checking the validity of the arguments made by Facebook for using tracking cookies. As main interests, Facebook mentioned the prevention of fraudulent access to accounts, protecting accounts that have been accessed via public computers, and preventing minors (under age 13) from signing-up to the service. It appeared that the claims were not valid, because the technical support for protection and prevention was not provided by the use of the cookies. In the end, the DPA concluded that the use of Facebook Fan pages and social plugins was infringing the state's data protection laws.[7]

Another interesting issue is that half a year after Facebook fixed the bug a similar tracking practice was highlighted by a researcher. Again, the tracking cookie was issued via third party websites, enabling Facebook to track logged-out users over the web.[8] In this case, Facebook reacted to the findings, which were written down in a blog post, by leaving a comment from an engineer. However, Facebook did not give an official reaction, but a spokesperson just referred to the comment under the blogpost without making an official policy statement.[9]

It is striking that Facebook admitted the practice to the German DPA while denying it in other jurisdictions. The reason for this difference is unclear. Besides, it can be questioned whether these diverging reactions are ethically acceptable. It might be the case that there are legal considerations behind this, which connect to the legal culture in the different jurisdictions. However, the difference also occurred between the Netherlands and Germany, which are comparable countries concerning the legal culture. Therefore, it might also be the case that Facebook's representatives in the different countries were not completely aware of each other's responses and the entire data collection practice in general, which lead to individual reactions. However, if that is the case, this seems to be another organizational shortcoming within the company.

---

[4] See:<http://reporter.kro.nl/uitzendingenreporter/_2011/facebook-vrienden-voor-het-leven 2.aspx >.

[5] This appears to be the case, indeed.

[6] See: Reuters Press release by T. Baynes: <http://newsandinsight.thomsonreuters.com/Legal/News/2011/05_-_May/Facebook_sued_for_using__Like__button_to_track_online_activity/>.

4 Apple Location Data

In April 2011, some technology researchers drew attention to a file on Apple's iPhones and iPads that recorded the GPS coordinates of nearby Wi-Fi access points and cellphone towers.[10] Apple came with a public reaction in a blog post with questions and answers.[7] Apple stated that it did not track the location of individual iPhones, but said that the data were only sent to Apple in an anonymized and encrypted form and could not be connected to the source by Apple. Furthermore, the data were stored on the iPhones and iPads in order to have the device more accurately calculate its current location. Storing the data on the device itself meant that Apple did not have individualized access to the data.

Strangely enough, it appeared that the location data were also updated when location services were turned off. This was the result of a bug, according to Apple. This bug would be fixed shortly, which was actually done with a new iOS version, together with a number of other software updates to reduce the amount of data stored, cease backups of the cache, and delete the cache when location services is turned off. In any case, the stored data were specific and frequent enough to give a detailed view of the iPhone user's whereabouts over the past months. However, to view or analyze this, physical access to the iPhone is necessary.

5 Comparison of the cases

The three cases concern the three major companies in ICT. It is striking that, within a year, all three of them were at the center of massive data collection that led to public outrage. What is even more striking is the fact that in all three cases the data collection was revealed by external parties, or came out during an externally instituted audit. The latter was the case for Google, which was the only case where the data were stored internally and the collection of the data could not be discovered from outside. Does this mean that the companies were lacking adequate organizational structures to prevent improper data collection or to check software by means of regular audits?

Another important issue is that all three companies point at a 'mistake' or a 'bug' as the reason for at least part of the 'inadvertent' data collection. It seems as if they want to say that it was not exactly their fault, but that the technology is just too complex to prevent any mistake from happening. But isn't it the case that if companies develop such complex technologies and distribute them all over the world, their responsibilities and control mechanisms should be in line with this level of complexity as well?

A third point of attention is the way the companies dealt with the issue. Apple and Google both came up with an official public reaction in which they explained what had happened and how they wanted to solve the issue. Nevertheless, there is an important difference between the two. On the one hand, Google indicated that the data

[7] Available at: < http://www.apple.com/pr/library/2011/04/27location_qa.html>.

were not intended to be collected and tried to convince people that the data were never used by Google and even are completely useless. On the other hand, Apple indicated to have the data collected and stored on purpose, albeit that there were unnecessary caches, too extensive data sets, and location data being updated while Location Services were turned off (this was the bug). The data is only sent to Apple in an anonymous and encrypted form and Apple cannot identify the source. This might imply that the data no longer qualify as personal data. However, the data stored on the iPhone are. What does this mean with regard to responsibility if an iPhone gets stolen or lost? The data on the phone are personal data which are processed in a manner decided on by Apple. So, Apple can be the controller, but loses control when selling the iPhone to a user. This user cannot make changes to the software and can, thus, not be held responsible, in particular because the data processing was not communicated properly and, thus, unknown to users.

Facebook did not come up with a public reaction. On the contrary, the data collection was only admitted in a communication with the German DPA in Hamburg and in all other cases there was no reaction at all or the practice was simply denied, although the software was updated to stop the tracking activities of non-members. Now, only members are still tracked until they explicitly log out of their Facebook account.[11] Besides, Facebook does use the data for advertising purposes. This makes the explanation that the extensive data collection was the result of a 'bug' less plausible. Nevertheless, it seems as if Facebook is trying to reduce all attention for the unlawful tracking and monitoring they performed and, to some extent, still perform.

Case	**Public reaction**	**Admit/ Deny practice**	**Defense**
Google StreetView payload data	Yes, blogpost	Admit	Programming mistake
Facebook Like Button tracking	No	Deny	Bug
Apple Location data	Yes, Q & A	Admit	Bug and partly intentional

Table 1. Comparison of the three cases.

6 Analysis and future perspective

The cases are striking and all three concern massive collection of personal data. Although there are numerous speculations about whether they really were mistakes or bugs, conspiracy theories and explanations that also contain irrelevant information to distract the attention from the main issue, this is not essential for drawing an im-

portant conclusion. In any case, the companies were not compliant with the Data Protection Directive (DPD).[8]

Some intriguing questions come up for all three cases. In the StreetView case, Google indicated that they had no interest in the payload data being collected. But why was the software code included in the program then? And; how was it possible that apparently no one at Google knew about the code being enabled? Basically, the amount of data collected was considerable and the collection took place over a long time period. Apparently, there had been no audits or audits had been inadequately performed, while good auditing could have led to quitting the practice earlier.

In the Facebook tracking case, the question arises why there was first no reaction from Facebook, even though the findings about the tracking received quite some attention in media worldwide. As was described, a first official reaction came when forced to respond by Data Protection Authorities. Recently, Facebook explained this to be the result of not having any contact person available for the company for Europe, Middle-East, and Asia, whereas now they have two and actively go into debate with researchers and authorities.[9] More important is the question whether Facebook did not know of the tracking cookies, or whether they did not take action if they knew about it. It seems to be that case that Facebook was aware of the tracking cookies. Only the setting of the cookies via Connect implementations on third party websites was claimed to be a bug, and the data collected via the tracking cookies was stated not to be used for profiling and advertising purposes. However, Facebook now admits the use of the tracking cookie, but claims to use it for security purposes, such as preventing false logins, inadvertently not logging out explicitly on public terminals, for instance in an internet café, and preventing minors from creating a profile page.[10] Strikingly, the Hamburg DPA issued a report in which it was shown that all purposes, indicated by Facebook to defend the use of the tracking cookie, were not performed with the help of this cookie.[12] It is, thus, still unclear what the exact truth in this case is.

Finally, Apple's location tracking gives rise to a few questions. Apple claimed that only aggregated data were received which could not be linked to specific devices. However, the data were communicated from each separate device, so how can this be aggregated beforehand? Another question is interesting with respect to data protection specifically. Once the data are aggregated they are no longer personal data. However, they are as long as they are on the iPhone or iPad. Apple decides that the data are collected and stored and, thus, seems to be the controller. But are they also the processor? And who is responsible if an iPhone gets lost or stolen? These questions are from a slightly different angle, which also indicates that the location tracking by Ap-

---

[8] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281.

[9] Indicated in a personal communication by Luc Delany, European Policy Manager Facebook, on November 10th 2011.

[10] Indicated by Luc Delany (see above) as well as Gregg Stefancik, Facebook engineer, in a personal conversation, on file with author.

ple is probably the least striking case of the three concerning massive data collection.[11]

Supposing that the data collections were inadvertently taking place because a software engineer made a mistake, the organization seems to be incapable of controlling what the employees are doing and lacks control mechanisms to prevent or detect mistakes in software that is exploited. The DPD requires sufficient technological and organizational measures to prevent personal data from being lost, altered, unlawfully disclosed or accessed, but also to have these measures to protect personal data against all other forms of unlawful processing. This requirement is not met and should be more strongly incorporated in the companies' policies. The fact that the discoveries of the unlawful processing were done by external parties (researchers) or by the company itself when forced to an audit of data underscores this. External checks, such as privacy impact assessments (PIA)[13] appear to be essential.

Another opportunity lies with the users of the services. In all three cases, there was public outrage concerning the massive data collection that took place. Broad media attention for the activities can have a huge impact on the companies responsible for the data collection practices. As can be seen, the companies see themselves being forced to come up with reactions in more or less elaborate forms in order to control the damage. That is one step. However, a next step might be much more important, but is also much more difficult to take. That is the step of users quitting the use of the services. If a large number of users object to the practices by deleting their membership accounts and changing their provider, the companies will be seriously affected. In order to achieve this, firm action by the community is needed. Nevertheless, the difficulty lies in the relatively large dependence a lot of people have on these services, also in light of the time they often invested in creating a profile page (Facebook), or the ease of use because the service is simply the biggest and best in its field (Google). Or there has been a serious investment in money to buy a device and a subscription (Apple iPhone).

If the data collections were consciously taking place, the companies seem to lack responsibility for their activities and at least fell short in meeting their information duties as laid down in the DPD. Data subjects have to be informed about their data being processed, for what purposes the data are processed, and how to exercise data subject rights. In all cases, these requirements were clearly not met. This also indicates that existing concepts, such as privacy by design (PbD)[14], will not solve the problems with these powerful companies. On the contrary, it seems that there is a trend to infringe upon privacy rights and each and every time take a new step in eroding privacy.[15] It is not a coincidence that Facebook CEO Mark Zuckerberg stated that users eventually get over privacy.[12]

To conclude, regardless of whether the examples were really mistakes or bugs, or well-intentioned, it is important to pay attention to the events. The fact that the mas-

[11] Nevertheless, the fact that Apple is included in this paper indicates that there is no fundamental difference between services for which is paid and services which are available for free and receive most of their revenues from trading based on data.

[12] See: < http://www.zdnet.com/blog/facebook/mark-zuckerberg-facebook-users-eventually-get-over-privacy-anxiety/1534>.

sive data collections could happen gives an indication that the culture within large ICT companies is probably not enough focused on privacy of individual users. The commercial goal is leading, which is logic, but fundamental rights should be respected. The concepts, like PIA and PbD, mentioned above can be helpful, but will not be the thing on their own. Completely independent inspections on software code are necessary to reveal the implementation of illegal data collection mechanisms. Ultimately, a cultural change might be needed in business cultures which affect an enormous amount of people all over the world. Lack of internal mechanisms to control technical processes facilitates inadvertent data collection. If the processes are not monitored properly, the risks of malicious use of the data or leakage to third parties may be serious as well. Thus, complexity of systems brings greater responsibilities for those who implement them. In order to have companies take these responsibilities, stronger enforcement by data protection authorities and international governmental bodies might be necessary.

References

1. Sayer, P.: Google's Street View Wi-Fi data included passwords, email. InfoWorld (2010)
2. Stroz Friedberg: Source Code Analysis of gstumbler. New York (2010)
3. College Bescherming Persoonsgegevens: Definitieve Bevindingen: Onderzoek CBP naar de verzameling van Wifi-gegevens met Street View auto's door Google. CBP, Den Haag (2010)
4. Masnick, M.: Why Google's Street View WiFi Data Collection Was Almost Certainly An Accident. TechDirt (2010)
5. Federal Trade Commission: Google Letter October 27 2010. Washington D.C. (2010)
6. Roosendaal, A.: Facebook Tracks and Traces Everyone: Like This! Tilburg Law School Research Paper (2010) http://ssrn.com/abstract=1717563
7. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD): Datenschutzrechtliche Bewertung der Reichweitenanalyse durch Facebook. Kiel (2011)
8. Cubrilovic, N.: Facebook Re-Enables Controversial Tracking Cookie. New Web Order (2011)
9. Protalinski, E.: Facebook Denies Cookie Tracking Allegations. ZDNet (2011)
10. Healey, J.: Internet data collection: the privacy line. Los Angeles Times. LA Times, Los Angeles (2011)
11. Efrati, A.: 'Like' Button Follows Web Users. Wall Street Journal. WSJ, New York (2011)
12. Hamburgische Beauftragte für Datenschutz und Informationsfreiheit: Prüfung der nach Abmeldung eines Facebook-Nutzers verbleibenden Cookies. Hamburgische Beauftragte für Datenschutz und Informationsfreiheit, Hamburg (2011)
13. Linden Consulting: Privacy Impact Assessments: International Study of their Application and Effects. In: ICO (ed.). ICO, Bristol (2007)
14. Cavoukian, A.: Privacy by Design. IPC, Ontario (2009)
15. Koops, B.J., R.E.Leenes: 'Code' and the Slow Erosion of Privacy. Michigan Telecommunications and Technology Law Review **12** (2005) 115-188