The Digital Deluge: Investigating Privacy Through GDPR

T. Hollis
Department of Computer Science
University of Toronto
Toronto, ON M5S 3H7
thollis@cs.toronto.edu

Abstract

This paper investigates the use of consumer data in tech companies in 2019, using recently introduced legislation via GDPR data requests. The main findings suggest only approximately half of the audited companies are GDPR-compliant, some companies do not provide all the data they have on users, excessive data is often collected and this data can be used to infer further information about the user without their consent.

1 Introduction

As we approach April 2019, dawn of the first anniversary of the General Data Protection Regulation (GDPR) written into law by the EU in 2018, questions regarding consumer data analytics have reached an all time high. The trust of data collection and privacy in Silicon Valley has plummeted [7], following scandals such as Facebook's Cambridge Analytica debacle. In response to such growing global concern, it seems fitting to exploit GDPR legislation, introduced last year, to uncover the current state of digital data privacy. The main research question that is tackled here is therefore: "*How much can GDPR queries reveal about the use of consumer data in tech companies in 2019?*"

2 Methodology

This paper's methodology is based on the *Armada Strategy* presented in the book Bit By Bit [8]. This investigation is therefore a *digital observational study* broken down into various complementary sub-investigations, preventing decreased *validity* arising from single-point-failure weaknesses.

Firstly, a high level analysis is undertaken to reveal the overarching characteristics of GDPR data, as well as company compliance statistics. This is followed by multiple deep-dive investigations into a handful of the datasets provided by the more notorious data-collecting companies such as Google, Facebook and Microsoft. For these companies, more specific data analysis is later undertaken such as friendship network clustering, sentiment analysis and news feed bias visualisation. This *Armada Strategy* approach is thus highly suited to the GDPR data acquired in this paper since queried companies varied significantly, both in the volume and structure of data that they provided. Indeed, although every single file was analysed, not all companies provided datasets with enough noteworthy information beyond what was already summarised in the initial high-level overview analysis. For example, some companies such as Microsoft returned a negligible amount of data (less than 500KB) by claiming that no other data was collected. While this seems like a dubious claim at best, it is very difficult to prove what tech companies like Microsoft store on their users and this remains a fundamental limitation of the methodology of this study.

The final compiled database used here is a combination of independent datasets provided by each of the 44 companies targeted, under legally controlled GDPR requests. In order to help alleviate ethical

concerns, each company dataset collected was gathered directly by this paper's author through parallel GDPR queries of his own private data. Due to the high sensitivity level of the data, many cautionary approaches were taken throughout the study to obfuscate private information from the presented results. Where privacy obfuscation was required and undertaken, this is consistently labelled in figure captions for maximum transparency. While this approach comes with clear disadvantages, both in terms of secure analysis implementations and time overheads, it seems to be for now the most ethical method of investigation.

Another limitation of this approach is a reduced external validity, due to the inherent bias of looking at only a single user. While this data is not expected to lie within any extremes of population statistics, some particularities in usage were certainly identified. For example, a low volume of Twitter data occurred due to recent registration, as well as particularly large Telegram data due to automated channels. Such external validity issues of bias can only truly be remedied by subsequent studies of other user data, either from observational or experimental supplemental approaches.

The 44 companies targeted correspond to all the companies that have ever stored personally identifiable information (PII) on the aforementioned user, an EU resident entitled to GDPR protections. While it is possible that some companies were overlooked, the probability remains low since the methodology insured many steps were taken to guarantee that the list was as complete as possible. This includes but is not limited to checking through the last two years of browsing history, checking all used passwords for all online accounts and using all manually recorded logs of public profiles.

3 Related Work

Due to the recent introduction of GDPR legislation less than 12 months ago, not all companies have had time to ensure compliance, even if they are legally obliged to have already done so at of the time of writing of this paper. As a consequence, the field remains relatively untouched by academic research. Nonetheless, the financial audit company Deloitte released a survey showing the lack of preparation of companies prior to GDPR enforcement deadlines [4]. In addition, a technical paper proposing a framework for verifying GDPR compliance has been published [1] but no investigation has been made thus far on compliance from a user perspective. There also exist multiple studies that have taken different approaches for investigating consumer data usage in tech corporations, using legislation other than GDPR. The most notable of such studies is perhaps the 2012 work undertaken by Smith, Szongott et. al [10]. In [10], the authors undertake an *observational* study of *ready-made* data by using web crawling techniques to reveal the vulnerabilities of posting and uploading public content on social media sites that use geo-tagging.

As this study is to be the first of its kind, its role is not to show conclusive generalisable trends but rather it is more of a unique opportunity to shed some initial light on a drastically under-researched field of computational social science that is destined to be of critical importance in years to come.

4 General Results

Over the course of six hours, all 44 companies suspected to hold PII were queried for GDPR data. Of those 44, four companies decided to exceed GDPR targets by simply refusing to store any data of any kind on their users. This was usually done by encrypting all data end-to-end and destroying all previously kept logs. Such companies were marked as having "excellent" GDPR compliance and were omitted from further investigation hereafter as they offer no data to analyse. Of the remaining 40 companies, only 16 allowed for the direct download of data in some form. The others showed no visible GDPR compliance thus required direct contact in the form of a legally-binding email making a formal GDPR query for data. All of the 28 companies that did not offer direct downloads are legally obliged to provide said data upon receiving this email request. However, these companies can take excessively long to respond and make the process unnecessarily complicated to deter requests. For example, Amazon ask for passports to be posted within a strict time-frame and Paypal refuse to cooperate. As such, these 28 companies were labelled as 'poorly compliant'.

The resulting overview of the final database is shown in Appendix 1. This *custom-made* database shows how much data each company gathered, how many files were present and which types of user data was logged. More specifically, the datasets provided by each company were checked for IP logging, location logging, sentiment tracking, device fingerprinting and search logging. Some general

quantitative and qualitative analysis is shown in Appendix 2. This analysis reveals IP/location history, search history, sentiment tracking and device fingerprinting was logged in over 40% of companies audited. In addition, before downloading the GDPR data and undertaking the deep-dive analysis, two surveys were taken by the author to serve as a benchmark for a before-and-after comparison. These surveys are presented in Appendix 3 and Appendix 4. The changes in survey responses reveal that the author's concern increased subsequently to undertaking this study. In addition, the author overestimated both the volume of data and number of companies storing PII data. Initial estimates of GDPR compliance rates were quite accurate as around 50% of companies audited were compliant.

5 Google Results

As shown in Figure 1, Google's data collection is mostly from breadcrumb data. While Mail and Google Drive make the largest cut (by volume), this is because they store user files. Most of the data (by number of files) actually resides in the 31 other Google services, many of which are services that are active unbeknownst to the user.

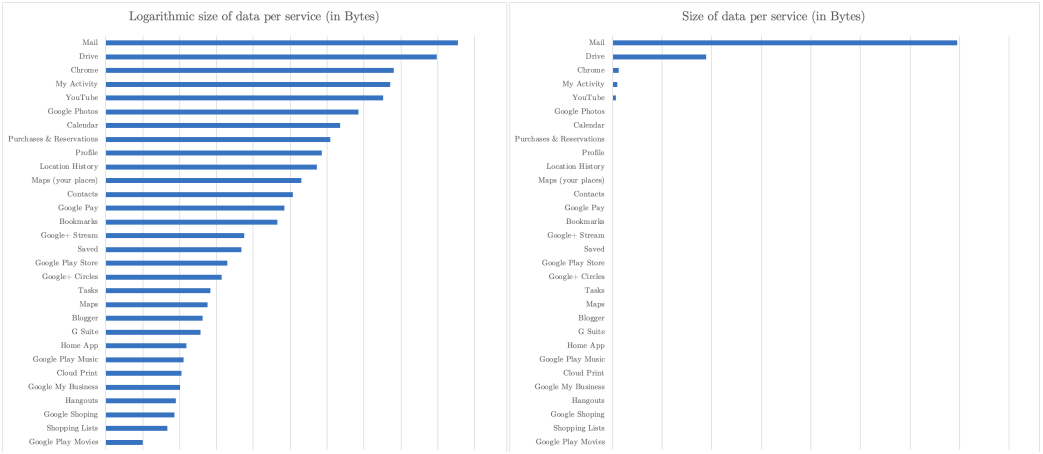

Figure 1: Google GDPR data overview (exact numbers obfuscated

While many services collected excessive and unnecessary data (such as every single YouTube and Google search since opening the account), the service most interesting to investigate here is Google Maps. Parsing the GDPR data through an R script with API calls to the Google Maps service helps shed light on the volume and nature of the location data collected by Google.

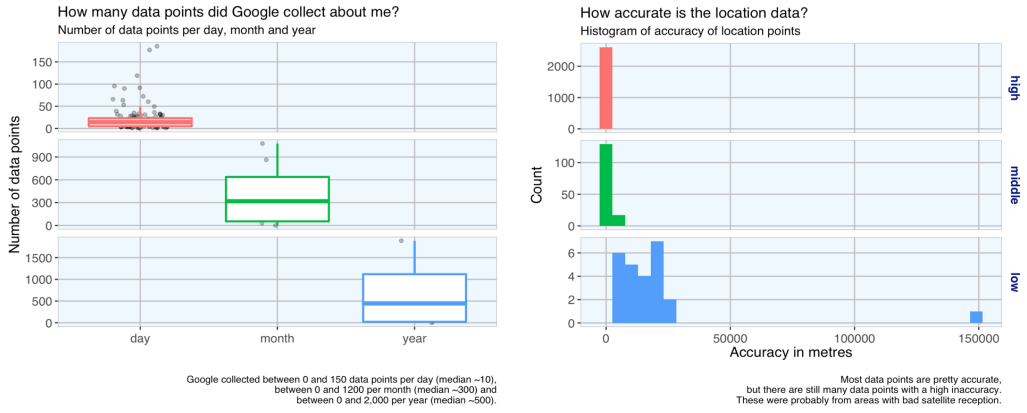

Figure 2: Google GDPR location data volume

As shown in Figure 2, location tracking logs occur at a rate of 300 per month or around 10 times per day. This data is only collected when Google Maps Timeline is activated (a setting that was on by

default but which was only active for one period of a few months in this dataset). Around 90% of the location data collected by Google was done using the highest available accuracy, such that most coordinates are accurate to a few tens of meters.

Figure 3: Google GDPR location data accuracy (privacy obfuscated)

As shown in Figure 3, this data can be used to infer speed (thus method of transport) and can be used to track down precise coordinates for homes and holiday accommodations. It is worth noting the accuracy of the data was found to be mostly higher near these locations.

Figures 1-3, as well as other unreported qualitative analyses of Google data, seem to suggest that data is collected with the goal of gathering as much of it as possible even if certain logs are of limited value to the user and of questionable value to the company or company clients such as advertisers.

6 Facebook Results

Facebook GDPR data was also distributed in a similar "breadcrumb-like" fashion as Google GDPR data. However, Facebook data concerns arose mostly with respect to alternative potential uses rather than their sheer number, which exposes users to having information inferred about themselves or their friends without their consent.

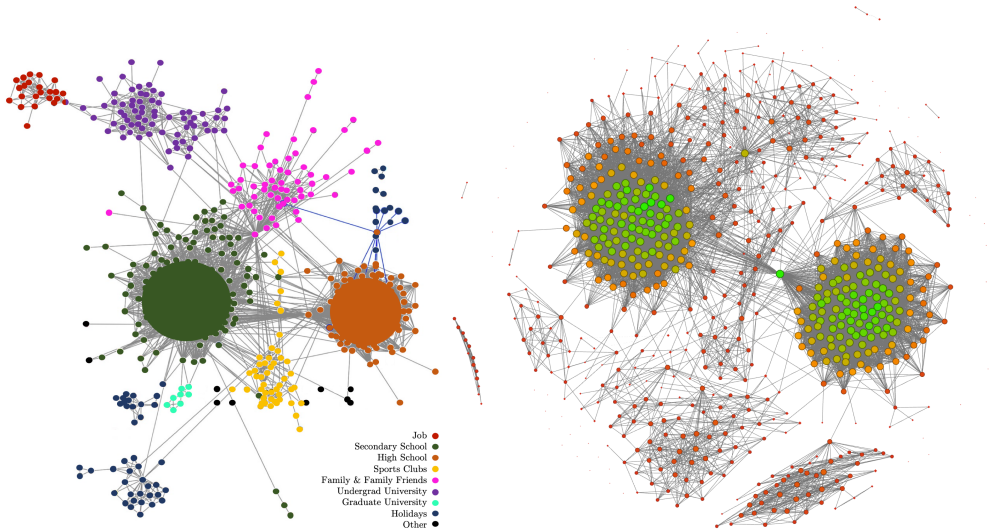

Figure 4: Facebook GDPR data (friends - privacy obfuscated)

Figure 4 reveals how the clustering of friends and mutual friends extracted from the GDPR dataset can be used to infer critical social characteristics. Using an R script and the open source Gephi application [2], the social network was easily clustered to reveal friendship origin (secondary school, high school, university, family, holiday, sports club...). A single Facebook friend within these clusters

that list their education or hometown information is enough to map such information to all other users within that mutual friend cluster. Thus, Facebook algorithms can implicitly learn features, such as where you went to school or holiday and which sports you undertake, only from a single friend listing their information.

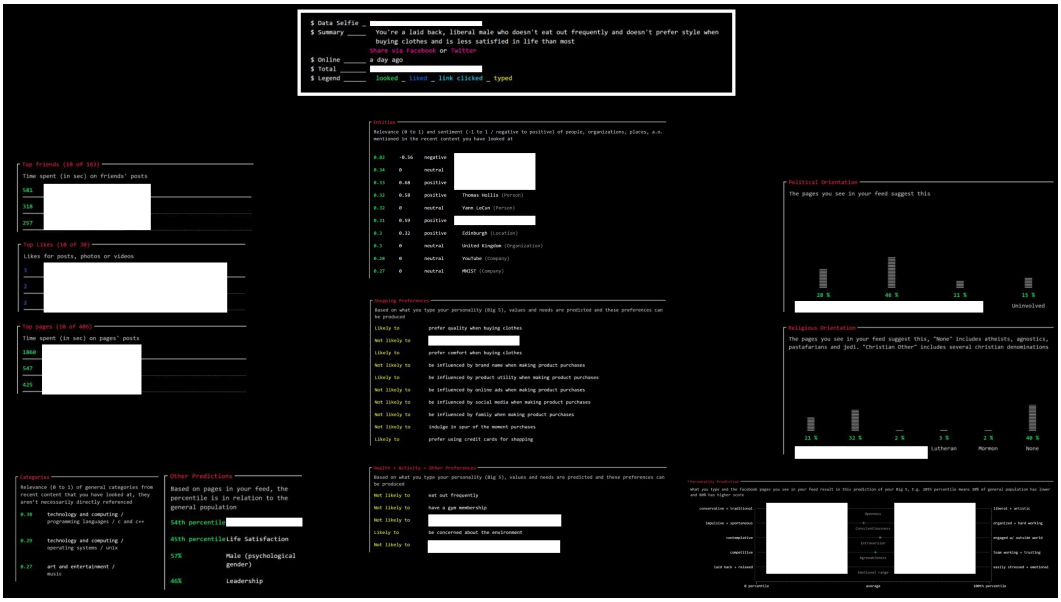

Figure 5: Facebook GDPR data (sentiments - privacy obfuscated)

Figure 5 shows how the timing data of how many seconds are spent looking at which Facebook posts is enough to reconstruct an accurate personality model of the user. The model used to generate figure 5 was trained on only a very small subset of Facebook timing data and relies on API calls to the IBM Watson ML and NLP servers of Data Selfie [9], an open source academic project funded by the New York City Economic Development Corporation. While the data is mostly obfuscated, predictions are troubling in their accuracy.

Appendix 4 shows Facebook Newsfeed and Friends data analysed using the PolitEcho tool [5]. This shows how information from the "Facebook Page likes" of friends can be used to generate a political influence model. This model is able to classify the political stance of friends and thus reveal what type of content is most exposed to the user in their Newsfeed, based on their friends' political influence.

7 Microsoft Results

Microsoft's GDPR data response was amongst the smallest in size of all companies audited. The data was notably minimal and only contained three files with small logs inside them. Indeed, Microsoft list in one of their services (Windows OS) some of the data that they collect on users, which is PII and GDPR-enforceable, but this data was omitted from the GDPR response without justification. Figure 6 does nonetheless show how the logs returned by Microsoft reveal some very interesting findings. The right-most graph shows OneDrive and MSOffice are the Microsoft services that are most useful in generating logs, where each log contains the IP address, location information and device fingerprinting data. While this may be justifiable for security reasons, such regular and granular logs are more of a security risk than protection. Indeed, these logs can be used to accurately model the total time spent on the computer as shown by the left-most graph. This graph shows the correlation between the Microsoft logs and the total time spent on the computer that month, as measured by the RescueTime application [6]. It is worth noting that when moving from Windows to macOS in July 2018, abandoning OneDrive, the logs' correlation dropped slightly.

Another interesting finding arises from training N-gram models using R scripts on Microsoft's LinkedIn message data. This shows that the style of speech is drastically different from that of N-gram models trained on informal communication channels like Telegram. While the outputs are

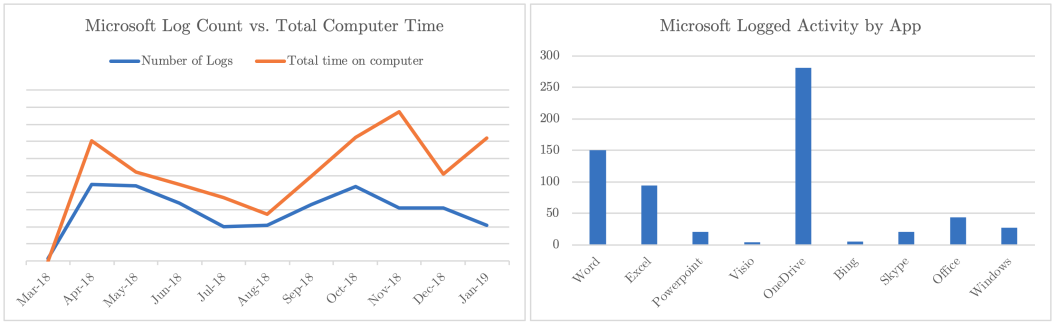

Figure 6: Microsoft GDPR data overview

omitted for privacy reasons, these suggests an encouraging limitation to sentiment tracking: certain websites like LinkedIn only reflect highly unrepresentative interaction data which does not generalise to other websites.

8 Conclusion

The implications of these results to the broader context of consumer data usage in tech companies in 2019 are notable. The general investigation has revealed that almost half of the companies audited track IP/location history, search history, sentiment tracking and device fingerprinting. The survey changes have revealed findings in this investigation are more concerning than reassuring. More specifically, through the three particular case studies of Google, Facebook and Microsoft we can also highlight three key areas of concern regarding the use of consumer data in tech companies in 2019.

Firstly, the Google investigation revealed that data is collected in a greedy "collect first, worry later" approach which is directly in contradiction with Salganik's ethical view of "minimising harm". Secondly, the Facebook investigation showed that the potential of this data is usually far more worrisome than initially anticipated. This can result in an unconsented inference of sensitive features possible from information added by friends rather than by the user themselves. Thirdly, as shown in the Microsoft investigation, we cannot trust that the data returned to us by GDPR requests is a complete sample of the PII stored about us by tech companies.

Having examined the data of all 44 audited companies, it seems that the healthiest approach consumers should take would be to assume that, when logged into a service, that company will log every single action undertaken on said service. This data can subsequently be used for inference so we should be mindful as to which services we use, favouring those that do not collect any data (such as Signal) and remaining logged into services only when required.

A limitation of this paper lies within companies that purchase pseudonymous data and try to combine it themselves. Such companies are unfortunately not bound by GDPR in the same way as those companies investigated in this paper. A notable example of this is Palantir which have repeatedly claimed they have not processed any data regarding users that have queried them by email. For this reason, these companies are particularly difficult to audit but this could be the subject of further research in this field. Another limitation of this paper lies within the methodology chosen, which is a trade-off of statistical significance for a more cautious ethical approach. While this means some results may be a statistical artifact, this does not affect the main findings of this exploratory analysis.

Many findings from this GDPR dataset were sensitive and could not be discussed or even mentioned in this paper, thus we encourage others to undertake their own investigations to help grow future work in this field. Indeed, all information that is not sensitive can be open-sourced and a meta analysis of these independent reviews could also constitute future work. To best facilitate this, all tools developed or used during this study have been open-sourced and uploaded to this paper's public GitHub repository [3]. Additional tools such as "Noiszy" and "Random User Agent" were also included which help to fight against excessive data collection by inserting random noise in searches and fingerprinting data.

References

[1] David Basin, Søren Debois, and Thomas Hildebrandt. On purpose and by necessity: compliance under the gdpr. *Proceedings of Financial Cryptography and Data Security*, 18, 2018.

[2] Gephi. The open graph viz platform. `https://gephi.org/`, 2019.

[3] GitHub. The digital deluge - a project exploiting gdpr legislation to uncover the state of consumer data privacy in 2019. `https://github.com/PsiPhiTheta/The-Digital-Deluge`, 2019.

[4] Erik Luysterborg. Deloitte general data protection regulation benchmarking survey. `https://www2.deloitte.com/be/en/pages/risk/articles/gdpr-readiness.html`, 2018.

[5] PolitEcho. Politecho - is your news feed a bubble? `https://politecho.org/`, 2019.

[6] RescueTime. Rescuetime - time management software. `https://www.rescuetime.com/`, 2019.

[7] Thomson Reuters. Who trusts facebook? `http://fingfx.thomsonreuters.com/gfx/rngs/USA-FACEBOOK-POLL/0100619Q2Q6/index.html`, 2018.

[8] Matthew J Salganik. *Bit by bit: social research in the digital age*. Princeton University Press, 2017.

[9] Data Selfie. Data selfie - get back your facebook data. `https://dataselfie.it/#/`, 2019.

[10] Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele Von Voigt. Big data privacy issues in public social media. In *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*, pages 1–6. IEEE, 2012.

9 Appendices

Company	GDPR compliance	Response speed (in hrs)	Volume of data (bytes)	Volume of data (GB)	Number of files	File Density (MB/file)	IP logging	Location logging	Sentiment Ad Sensity	Fingerprinting (device info)	Search history logging	Contact info	Total Score	Comments
[redacted]	compliant	< 1	[redacted]	0.00001	4	0.002	No	No	No	No	No	No	50	Minimal, great
	compliant	< 1		0.00002	3	0.006	Yes	Yes	Yes	Yes	No	No	30	Excessive for usage
	compliant	< 1		0.00004	26	0.001	No	No	Yes	Yes	No	No	40	Minimal, great
Microsoft	compliant	< 1		0.00015	4	0.038	No	Yes	No	No	Yes	No	40	Bad, hiding info
[redacted]	compliant	66		0.00024	3	0.081	Yes	No	No	Yes	No	Yes	35	Minimal, great
	compliant	< 1		0.00028	2	0.142	No	No	No	No	No	No	50	Logged all code
	compliant	< 1		0.00038	45	0.009	Yes	Yes	Yes	Yes	Yes	Yes	20	No chat history?
	compliant	14		0.00043	56	0.008	Yes	No	Yes	Yes	No	Yes	30	Mostly profile info
	compliant	< 1		0.00998	1	10.221	No	No	No	No	No	No	50	[redacted]
	compliant	< 1		0.01060	113	0.091	No	Yes	No	No	No	Yes	40	
	compliant	121		0.01751	643	0.028	Yes	Yes	Yes	Yes	Yes	Yes	20	
	compliant	< 1		0.04763	7	6.967	Yes	Yes	No	Yes	No	No	35	
GitHub	compliant	< 1		0.40376	1046	0.397	No	No	No	No	No	No	50	
Google	compliant	< 1		4.29681	989	4.418	Yes	Yes	Yes	Yes	Yes	Yes	20	Extensive/Invasive
Facebook	compliant	< 1		4.42603	13,182	0.344	Yes	Yes	Yes	Yes	Yes	Yes	20	Extensive/Invasive
Telegram	compliant	< 1	6,951,734,078	6.47431	24,004	0.270	Yes	Yes	No	Yes	No	Yes	30	Only message/security data
[redacted]	compliant	-	-	-	-	-	-	-	-	-	-	-	50	-
	excellent	-	-	-	-	-	-	-	-	-	-	-	100	-
	excellent	-	-	-	-	-	-	-	-	-	-	-	100	-
	excellent	-	-	-	-	-	-	-	-	-	-	-	100	-
	excellent	-	-	-	-	-	-	-	-	-	-	-	100	-
Amazon	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
[redacted]	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
	poor	-	-	-	-	-	-	-	-	-	-	-	0	-
Slack	poor	-	-	-	-	-	-	-	-	-	-	-	0	-

Figure 7: Appendix 1a - Dataset overview of 44 target companies (Privacy-Obfuscated)

Metadata

Compliance	Count
compliant	17
excellent	4
poor	23

Response speed	Count
< 1 hr	14
14 hrs	1
66 hrs	1
121 hrs	1

IP logging	Count
No	7
Yes	9

Location logging	Count
No	7
Yes	9

Sentiment Ad Sensin	Count
No	9
Yes	7

Fingerprinting	Count
No	6
Yes	10

Search history loggin	Count
No	11
Yes	5

Figure 8: Appendix 1b - Dataset overview metadata

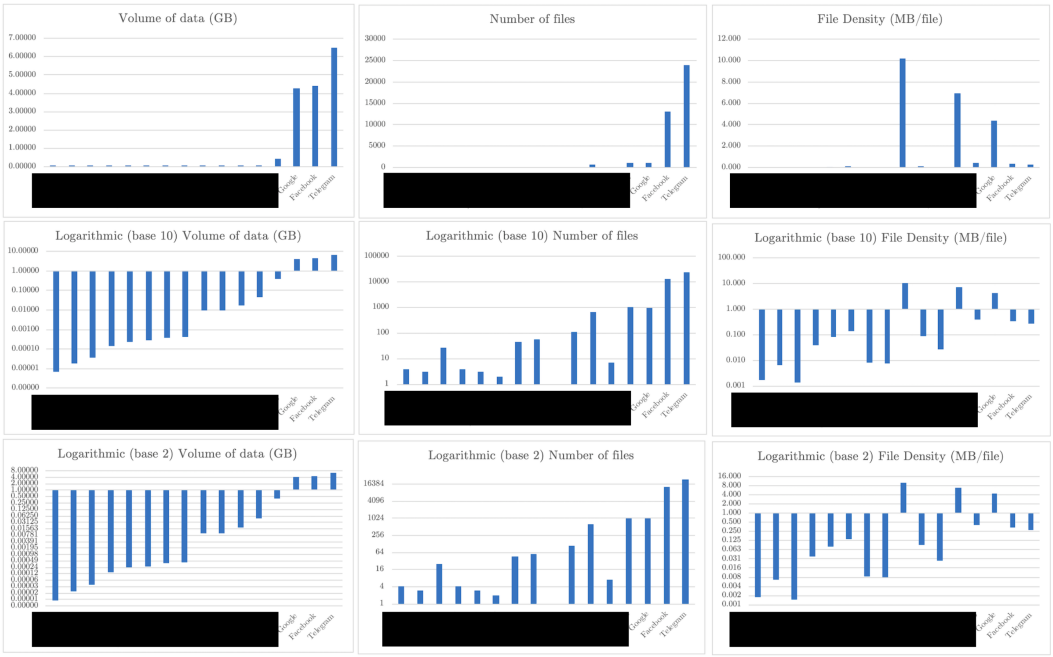

Figure 9: Appendix 2a - Dataset overview quantitative analysis (Privacy-Obfuscated)

Figure 10: Appendix 2b - Dataset overview qualitative analysis

1 Please rank the privacy importance on the following 1 - 12 types of personal data, using a scale of 1 - 5, 1 represents being not important at all in terms of privacy, 5 represents being very important in terms of privacy.

	Unimportant	Slightly important	Important	Very important	Critical	No opinion
Biometric data (e.g. finger, retinal, facial recognition)	○	○	○	○	●	○
Gender	○	●	○	○	○	○
Religion	○	●	○	○	○	○
Home or mobile phone number	○	○	●	○	○	○
Home address	○	○	○	●	○	○
Personal financial information (bank account; income)	○	○	○	○	●	○
Health conditions (including medical history)	○	○	○	○	●	○
Age	○	●	○	○	○	○
Family financial situation	○	○	○	●	○	○
Identity documents (passport number; drivers license number)	○	○	○	○	●	○
Marital status	○	●	○	○	○	○

Figure 11: Appendix 3a - 2018 We Fight Fraud Personal Data and Privacy Awareness Survey (Part 1)

6 How much trust do you have in the following organisations with regards to how they protect or use your personal data?

	Poor	Fair	Good	Very Good	Excellent	No Opinion
Financial Institutions (Banks; credit)	○	●	○	○	○	○
Insurance Companies (Life; pensions; investments)	○	●	○	○	○	○
General Insurance Firms (Home contents; car; buildings, accident)	○	●	○	○	○	○
Charities	●	○	○	○	○	○
Central Government	●	○	○	○	○	○
Local Government	●	○	○	○	○	○
Public Health Service (NHS)	●	○	○	○	○	○
Private Health Services (Private clinics, dentists, cosmetic/plastics)	○	●	○	○	○	○
Market research organisations	●	○	○	○	○	○
Retailers (High Street)	●	○	○	○	○	○
Online retailers	●	○	○	○	○	○

Figure 12: Appendix 3b - 2018 We Fight Fraud Personal Data and Privacy Awareness Survey (Part 2)

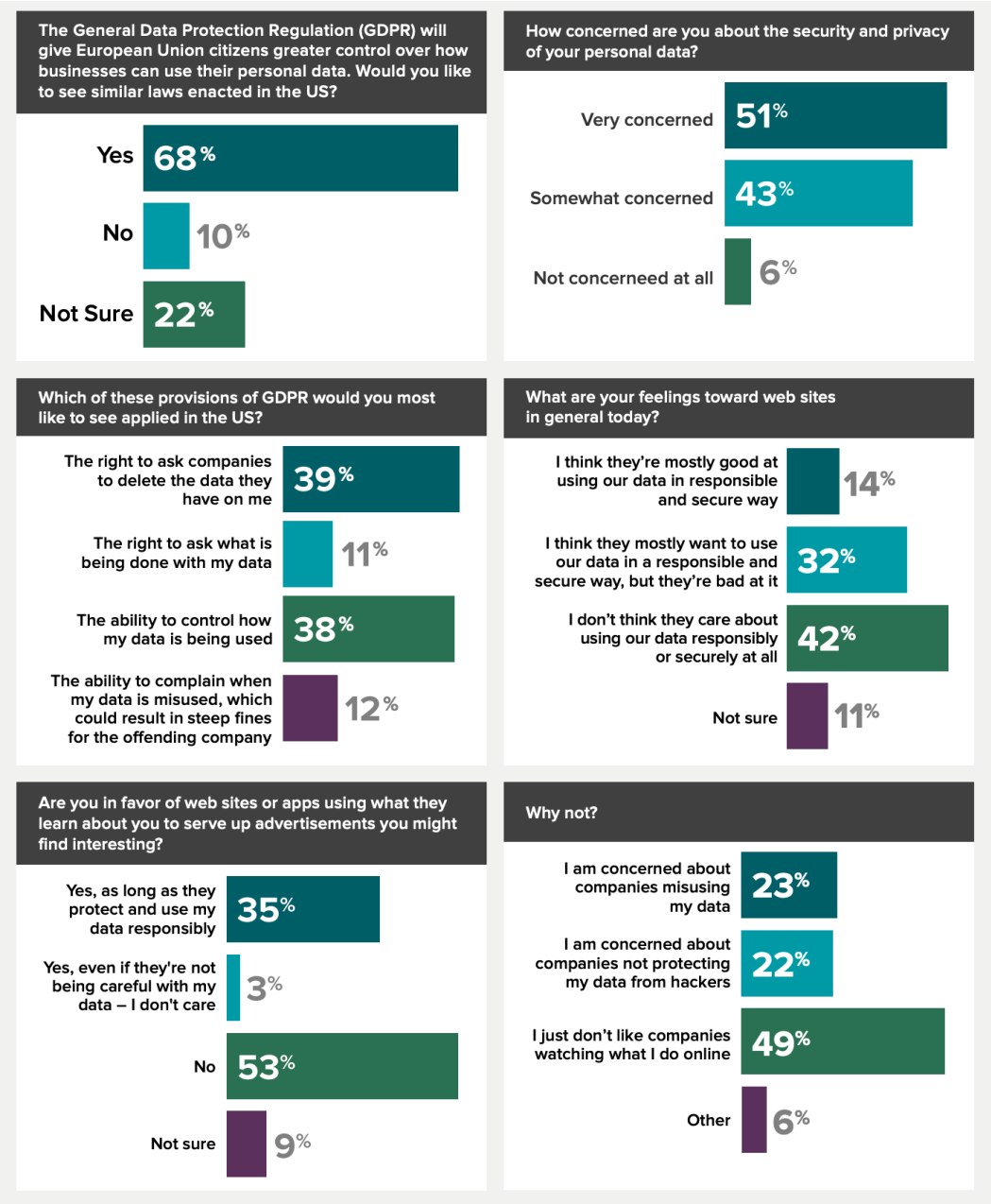

Figure 13: Appendix 3c - Janrain Research Consumer Attitudes Toward Data Privacy Survey (Part 1)

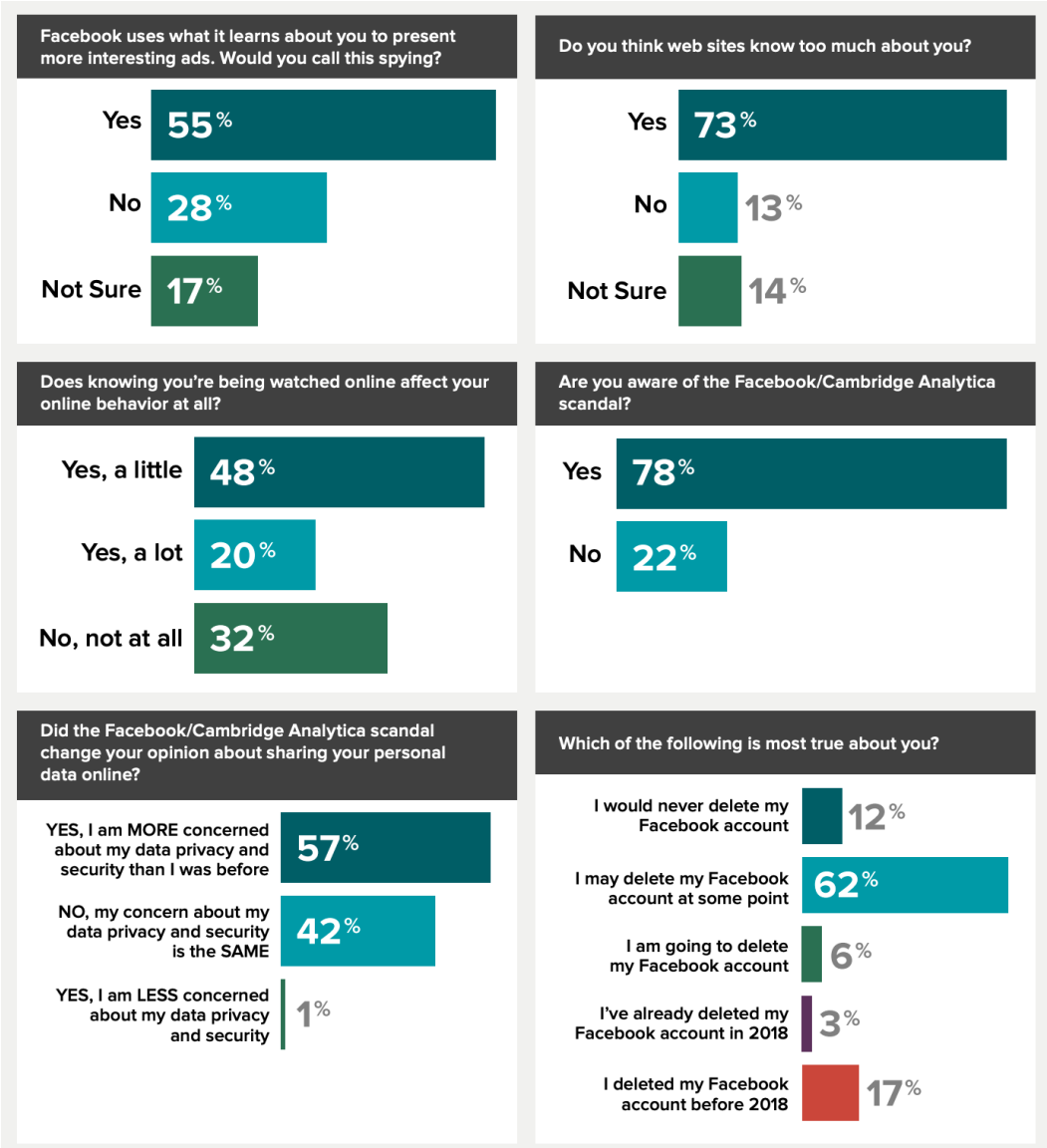

Figure 14: Appendix 3d - Janrain Research Consumer Attitudes Toward Data Privacy Survey (Part 2)

Figure 15: Appendix 4 - Facebook GDPR data (political bias - privacy obfuscated)