The Digital Deluge: Investigating Privacy Through GDPR

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

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 [3], 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 [4]. 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 Apple. For these companies, more specific data analysis will be 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 as 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 [2]. 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 [5]. In [5], 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 Preliminary Progress

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 preliminary quantitative and qualitative analysis is shown in Appendix 2.

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 plan is to have a post-analysis followup to see how opinion has changed after confronting the evidence. The surveys taken also include data from the original responses of the public so the author’s views will be compared against the global opinion distribution, after all the deep dive analysis has been completed.

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

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

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

[5] 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.

5 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		0.00198	1	10.221	No	No	No	No	No	No	50	[redacted]
	compliant	< 1		0.01040	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/Intrusive
Facebook	compliant	< 1		4.42603	13,182	0.344	Yes	Yes	Yes	Yes	Yes	Yes	20	Extensive/Intrusive
Telegram	compliant	< 1	6,051,734,078	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 1: 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 2: Appendix 1b - Dataset overview metadata

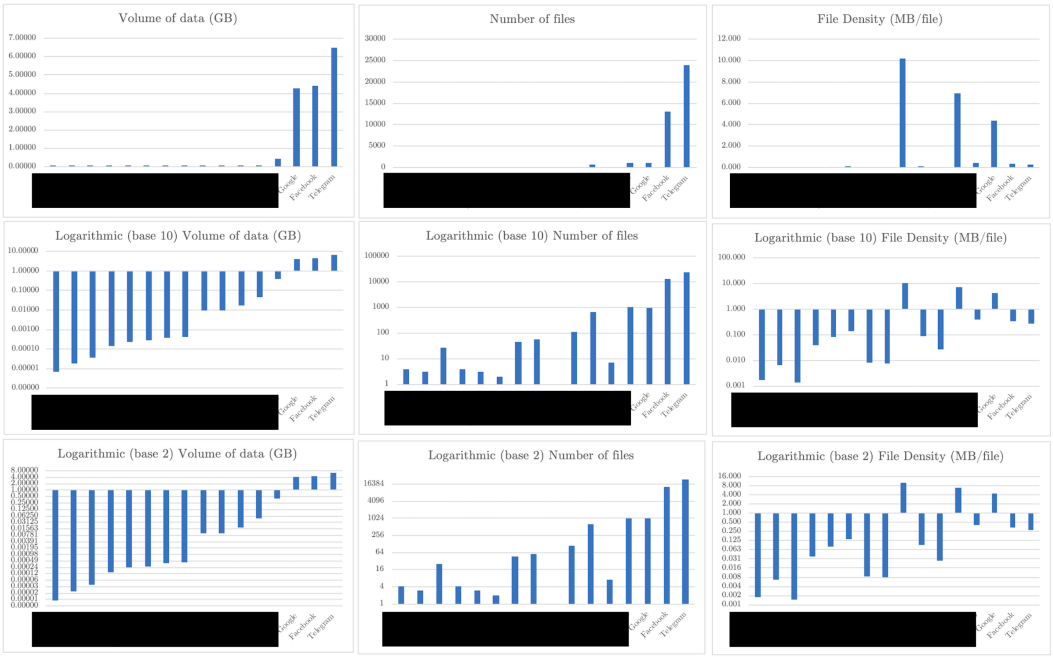

Figure 3: Appendix 2a - Dataset overview quantitative analysis (Privacy-Obfuscated)

Figure 4: 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 5: 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 6: Appendix 3b - 2018 We Fight Fraud Personal Data and Privacy Awareness Survey (Part 2)

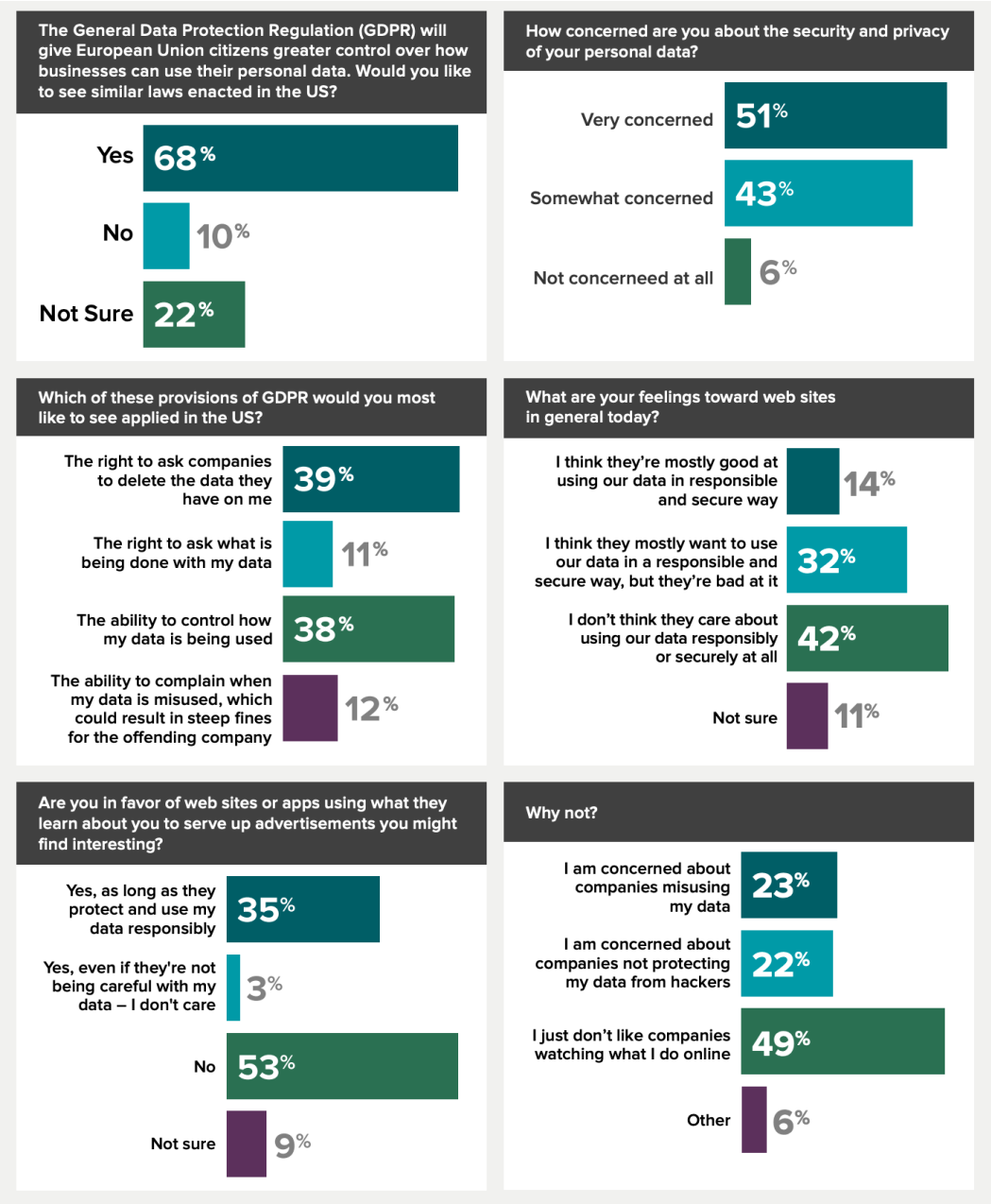

Figure 7: Appendix 3c - Janrain Research Consumer Attitudes Toward Data Privacy Survey (Part 1)

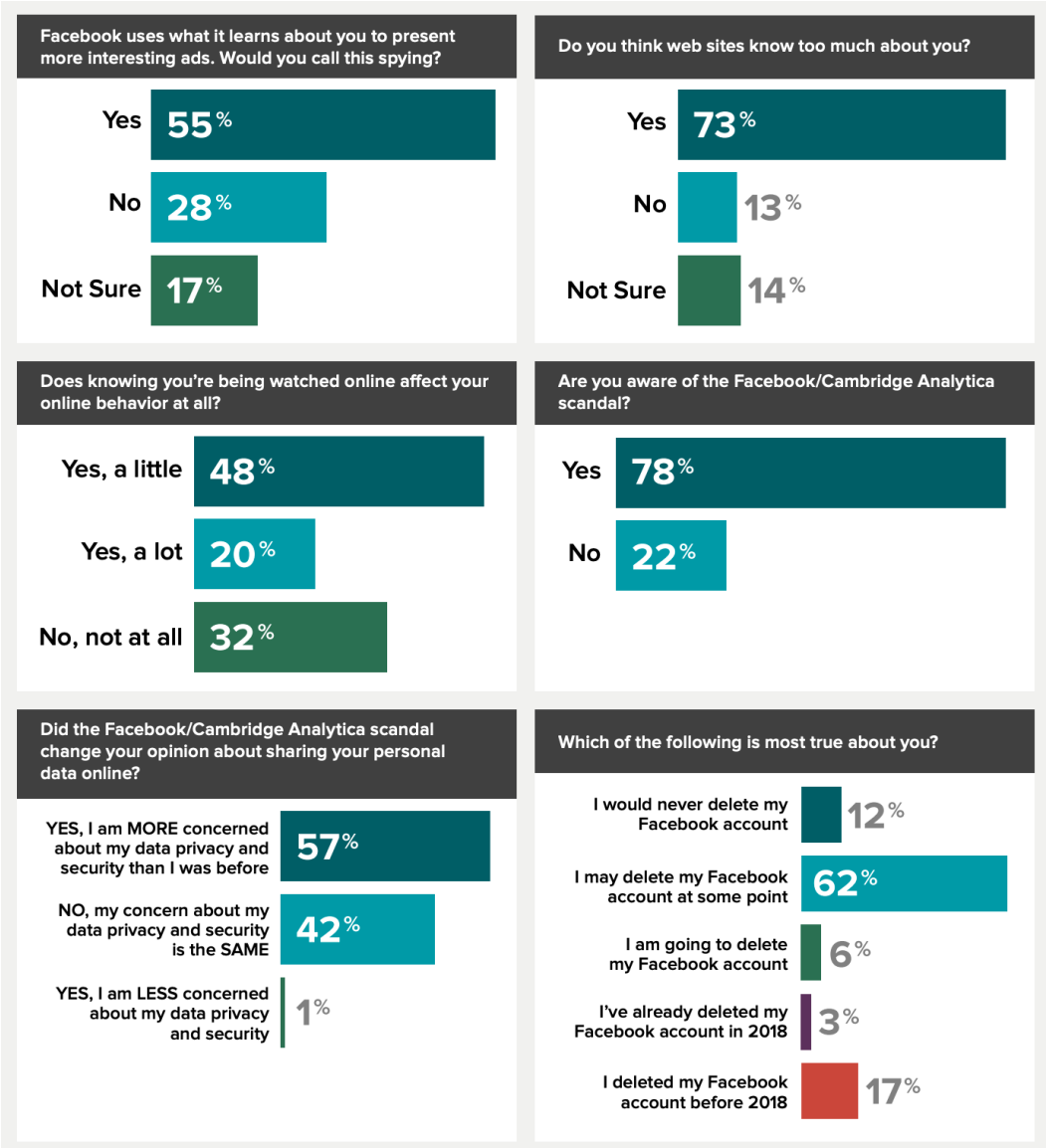

Figure 8: Appendix 3d - Janrain Research Consumer Attitudes Toward Data Privacy Survey (Part 2)