Hestia.ai
sustainable data solutions

#digipower investigation

Technical Reports

AUDITING THE DATA ECONOMY THROUGH PERSONAL DATA ACCESS

A Methodology and Case Studies Report by Hestia.ai, May 2022

Authors: Alex Bowyer, Jessica Pidoux, Jacob Gursky, Paul-Olivier Dehaye

Acknowledgements

The authors want to thank, for their trust, their time and their courage in engaging in something new, the study participants:

- Anders Adlercreutz,
- Leïla Chaibi,
- Filomena Chirico,
- Christian D'Cunha,
- Stephane Duguin,
- Atte Harjanne,
- Jyrki Katainen,
- Miapetra Kumpula-Natri,
- Dan Koivulaakso,
- Markus Lohi,
- Tom Packalén,
- Sirpa Pietkainen,
- Mark Scott,
- Niclas Storås, and
- Sari Tanus.

The authors also want to thank

- for his help, Konrad Kollnig (developer of Tracker Control),
- for their support, Tiina Härkönen, Kristo Lehtonen, Jukka Vahti, Riitta Vänskä and the SITRA staff,
- for their expertise, PersonalData.IO members and volunteers, and in particular Judith Herzog,
- for their dedication, members of data collectives *The Eyeballs*,
- the Migros Pioneer Fund and Robin Born, for their understanding of the relevance of #digipower to the ongoing *HestiaLabs* project, and their willingness to accommodate priority changes accordingly,

and of course

- Hestia.ai staff, who have all contributed differently to the investigation: Emmanuel Eckard, Charles Foucault-Dumas, Hugo Hueber, Andreas Kündig, Valentin Loftsson, François Quellec, Florian Singer, Luã Streit, Marie-Pierre Vidonne, Thomas Wilde

Executive Summary

ABSTRACT

There is an imbalance of control over personal data between service providers and civil society. While service providers acquire knowledge and influence individuals' behaviour through data, individuals do not own their data, and the personal data ecosystem lacks the transparency necessary to be understood. The #digipower investigation explores the practical realities of power and influence in the data economy. This report, one of two, is a detail-oriented report examining the experiences of 15 high-profile participants (among them members of the European and Finnish parliaments, EU and Finnish civil servants, NGO directors, and journalists) in exploring their own personal data ecosystems. Each participant targeted around six companies, selecting common service providers they use in their daily lives, and were coached to use Subject Access Requests, data download portals and technical audits to obtain personal data and data handling information from those targets. Obtained data and responses were discussed with coaches and subsequently analysed by experts at Hestia.ai, who then used the collected data and the participants' experiences to understand and map out patterns of data usage by the service providers. With over 83 individual/service provider relationships explored, we present four overarching case studies, illustrated using participants' experiences and data. These case studies focus respectively on (1) influences by service providers on the physical world, and (2) the online world (2), on influencing by infrastructure providers (3), and on the challenges of obtaining held data (4). Through this investigation we contribute a pedagogical approach for research and examination of data economy practices, and provide a map of some common data flows between companies. Our accompanying narrative report takes a top down approach to map out the uncovered mechanisms of influence and power we reveal in the data economy, and considers their societal implications.

GOAL

The #digipower investigation seeks to understand the distribution of power in the data economy, and in particular how that distribution of power changes as more of the economy becomes digitised, and ever larger amounts of personal data get collected.

METHODOLOGY

We assumed that it would be particularly interesting to consider our research question around test subjects who were already decision makers in a traditional form of power. We settled on fifteen participants that we selected amongst members of the Finnish and European Parliaments, civil servants, directors of NGOs, journalists, etc.

We then built a participatory investigation, which required coaching the participants on how to retrieve their personal data, based on three lenses available to each of us:

- subject access requests, i.e., the right to view one's own data under the General Data Protection Regulation;
- data download portals, i.e., self-checkout transparency portals built by companies on a voluntary basis;
- a technical audit, through an Android app installed on a loan phone.

Each lens offered a different but complementary perspective on data flows surrounding our participants.

INDIVIDUAL FINDINGS

Through that process, the participants collected numerous facts about the data economy surrounding them. For instance, we found evidence of

- online retailer Gigantti sharing individuals' physical purchase information with Meta (Facebook),
- Boeing seeking to influence Finnish MPs to sell fighter planes (including the criteria Boeing used on Twitter),
- the UK Labour Party and data broker Bisnode using Meta (Facebook)'s targeting tools,
- how Google blends data from searches with wifi signals to infer at which exact location one might be,
- systematic non-compliance with General Data Protection Regulation requirements, etc.

CASE STUDIES

Additionally, we weave many of the individual findings into four case studies as when analysed in a transversal way, they contribute together to bigger findings. The case studies are:

- *Who cares about my geolocation and why?*, which simply shows what can be collected and understood about position and movement in the physical space, and the consequences thereof;
- *When you view the web, the web views you*, which contrasts the previous case study with similar digital situations, and highlights the heightened potential for shaping the environment online;
- *Move fast and capture all signals, everywhere*, which is focused on the ability of Facebook to convince others to install their tracking tools, despite going against the long term own interest of those partners ;
- *Participants chasing their personal data*, which is focused on recounting the numerous difficulties that the participants faced.

All of the above is detailed in a methodology and case studies report, entitled *"Auditing the Data Economy through Personal Data Access"*.

BIG PICTURE

We also synthesised all of our heterogeneous, but complementary, findings into one coherent vision of the data economy, as part of a narrative report entitled *"Understanding Influence and Power in the Data Economy"*.

We identified an overarching situation of power (*Infrastructural Power*) that functions through two distinct capabilities working in conjunction with two feedback loops:

- *Technical Capability* opens up the *Accumulating Information and Knowledge to Act* loop, for instance to influence an user of online services to make a purchase through extensive prior profiling.
- *Organisational Capability* enables the *Composing Complex Infrastructures for a Dominating Position* feedback loop, for instance to encourage website owners to transfer data to Facebook.

We also formulated much more precise decompositions of the mechanisms that can be used to acquire positions of power, and provided a taxonomy of the consequences of digital power.

RECOMMENDATIONS

As a guide towards a desirable future for the data economy, we formulate a *#digipower Manifesto*, rooted in values of *transparency through sound engineering design*, *proportionality of data collection*, and *social care*.

In order to reach that future, we make five recommendations:

- *Change the Narrative* of innovation around data, to encourage business practices that are more sustainable and in line with the General Data Protection Regulation;
- *Productivise #digipower*, i.e. replicate this approach to pedagogy situated in the personal data of the study participants, in order to seed multiple communities in line with the manifesto goals;
- *Increase Infrastructural Power of Civil Society*: build technical capability and organisational capability in order to have effective counterpowers in the digital economy;
- *Support Data Collectives* as a vehicle for reaching faster a more fair distribution of value in the data economy;
- *Enforce GDPR Properly* or risk affecting negatively innovative European businesses.

1. Introduction

The #digipower investigation is carried out by Hestia.ai, commissioned by SITRA, on the digital power that controls today's data economy. One of the goals of this investigation is to contrast this digital power to other traditional power forms that high-profile individuals like the #digipower participants would hold, by collecting evidence rooted in those participants' digitised lives.

The results of the investigation are presented in two reports: *(i) a* case studies report entitled *"Auditing the Data Economy through Personal Data Access"* and *(ii)* a narrative report entitled *"Understanding Influence and Power in the Data Economy"*. Both reports are accompanied by a shared *Executive Summary.*

The *Executive Summary* presents the main findings of two perspectives on digital power: the bottom-up results of individual experiences when recovering personal data, the top-down results about the data economy which were produced from a collective analysis of individual experiences – a "big picture". The narrative report gives a theoretical understanding of the big picture – top-down results.

This document, the methodology and case studies report, can be read independently of the narrative report and aims to document our methodology and to present case studies, our bottom-up results. The case studies show illustrations of the data economy involving individual participants. We present those individual illustrations as part of larger coherent wholes (the case studies), each discussing a facet of the data economy. The case studies, which form the heart of this report, are as follows:

1. *Who cares about my geolocation and why?*
2. *When you view the web, the web views you*
3. *Move fast and capture all signals, everywhere*
4. *Participants chasing their data*

First we contrast two situations: when an individual moves in physical space and parts of their life get digitised (Case Study 1), and when an individual moves through a world of digital content (Case Study 2). The contrast reveals that in the digital context there are significant new implications for data usage, which can then be applied back to the physical context. Case Study 3 discusses the introduction by different actors such as Facebook and Google, of data-collecting 'sensors' into service providers' websites and apps. Finally, Case Study 4 reflects on the difficulties encountered by the participants when trying to recover their data from service providers. Data access challenges, SAR blockages, lenses' limitations, and challenges in following up SARs are discussed.

Through these case studies, we aim to present an aggregate and pedagogical view of the data economy. As we will explain in chapter 3, we will use the game FarmVille as a point of reference to judge the various situations we explore through the case studies and to scaffold our pedagogical explanation of case study findings. These case studies can be read alone, but can also provide the reader with an illustrated introduction to the theory presented in the narrative report. In the text we provide further leads into relevant parts of the narrative report.

The methodology and case studies report is structured into four chapters, of which this Introduction is the first. After the present introduction, **Chapter 2** presents the #digipower methodology through eight sections: the study design, study participants, target service providers, and the lenses on the data economy. Additionally, this chapter discusses the resources and foundations of the investigation: data analysis and visualisation tools, data typology, reproducibility, previous work.

Chapter 3 analyses the four case studies as described above, in Sections 3.1, 3.2, 3.3 and 3.4 respectively.

Finally, **Chapter 4** develops the #digipower coaches' reflections about the whole investigation process with participants, describes how we structured this report and approached participant interviews in such a way as to maximise the pedagogical value of the telling of our case studies. It also invites the reader to read the narrative report for a deeper exploration of the big picture of data and power in today's data economy.

2. #digipower Methodology

In this study, we deployed a new methodology that is participatory and relies upon three main principles:

- We helped the participants to develop an understanding of the personal data economy around them, **situated** in their own recovered personal data.
- We aimed to combine multiple participants' experiences in order to provide **mutual learning** between individuals.
- We **modelled** the main facts resulting from the experiences and participant-obtained data to represent the **data economy**, providing a view of the big picture beyond individual observations.

After a couple months of preparation, the heart of this investigation, involving the participants, ran between October 2021 and February 2022.

In this chapter we explain the study methodology and implementation, including study design (Section 2.1) participant details (Section 2.2), target companies (Section 2.3), lenses used (Section 2.4), visualisation tools used (Section 2.5) and our data typology that was used for reference during the study (Section 2.6). In addition we position this investigation relative to prior work (Section 2.7), and explain how it could be replicated (Section 2.8).

2.1. STUDY DESIGN

2.1.1. Justification

The problem space explored by the #digipower investigation is the power imbalance created around data control between civil society and other stakeholders. According to the World Economic Forum in 2014[1], there is "an imbalance in the amount of information about individuals held by, or that is accessible to, industry and governments, and the lack of knowledge and ability of the same individuals to control the use of that information".

Our immediate objectives were to better understand the mechanisms by which data provides power to exert influence over civil society; to explore whether those mechanisms are applied differently towards prominent individuals than laypeople; and to identify specific problematic practices in data processing that might be addressed to redress the power imbalance.

[1] W. Hoffman et al., *Rethinking Personal Data: Trust and Context in User-Centred Data Ecosystems*, May 2014, https://www.weforum.org/reports/rethinking-personal-data

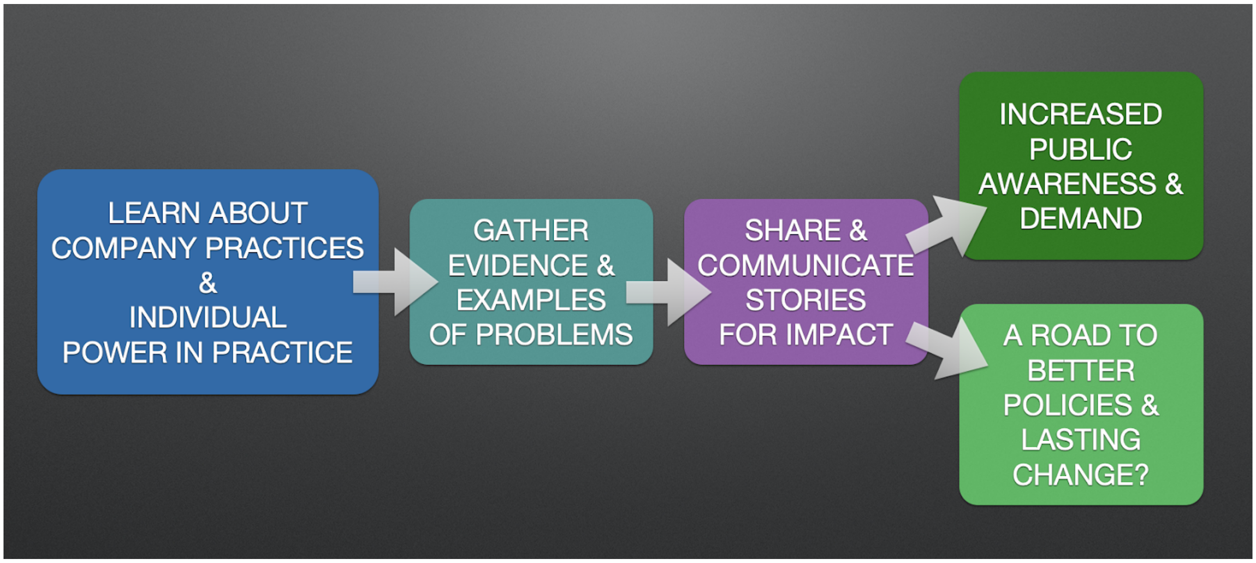

LEARN ABOUT COMPANY PRACTICES & INDIVIDUAL POWER IN PRACTICE
GATHER EVIDENCE & EXAMPLES OF PROBLEMS
SHARE & COMMUNICATE STORIES FOR IMPACT
INCREASED PUBLIC AWARENESS & DEMAND
A ROAD TO BETTER POLICIES & LASTING CHANGE?

Figure 1: Long-term objectives of the #digipower initiative

The long-term objectives of the #digipower initiative are shown in Figure 1: First, public impact - in terms of both raising awareness of data empowerment issues and catalysing demand for a fairer relationship with our service-provider held personal data. Second, to identify a roadmap for how policies, enforcement and practice should change in the future. The process towards these two objectives involves three key steps:

1. For individuals to use their currently available means to access their own held personal data, in order to understand, in a **grounded** way, companies' data practices and understand whether currently available means of data access are effective.
2. To find **evidence**, within individuals' experiences and their retrieved data, of problematic practices that could be addressed, or of previously unseen practices or processes.
3. To enable the sharing of these experiences and evidence as impactful **stories**, that individuals can share or discuss with others, or write about in journalism or on social media.

Our premise is that the strongest understanding about the data economy is gained from one's own experience, the most impactful stories are those which are relatable, and that therefore enabling such stories to be uncovered and evidenced ultimately contributes to spreading the word and sharing knowledge. It produces expertise-networks of people sharing their own experiences with others.

2.1.2. Procedure

We designed a procedure of taking participants on a **journey** of gaining access to their own data, then working with us to uncover evidence, carry out investigations and construct stories.

Our goal throughout the participant interaction is to uncover ‘nuggets’ (or ‘leads’ in journalistic terms), as was explained to participants using the diagram shown in Figure 2 below. These nuggets could be further investigated, or shared with others as stories.

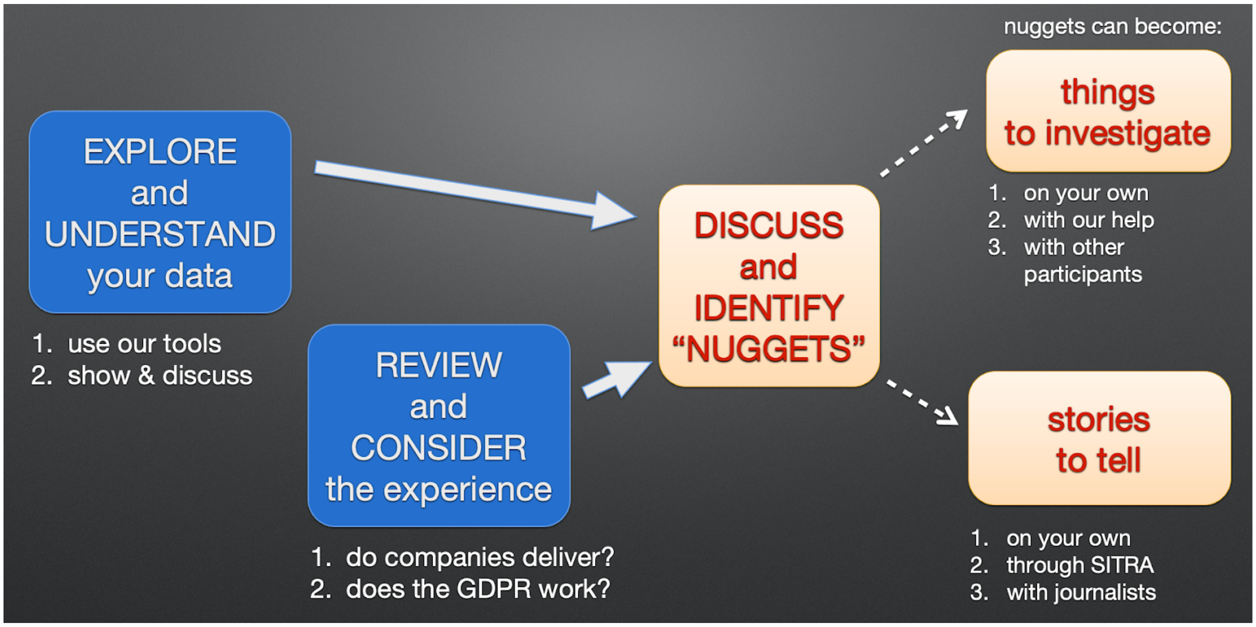


Figure 2: extracting ‘nuggets’ which can become investigations or stories

The individual journey for each participant involved the design of three key phases:

1. **Onboarding**: This phase included recruitment, briefing and introductions, and selection of target companies;
2. **Discovery/Coaching**: This phase included obtaining data from and about target companies, with coaching and support provided; and
3. **Follow up**: This final phase included data analysis and interpretation, interviews, reporting and follow up activities framed around further investigations or storytelling and publicity.

The main components of the three phases are shown in Figure 3, which is a diagram that was used to explain our plans to participants as they embarked on their journey. There were some minor deviations from this approach, which emerge through the explanations below.

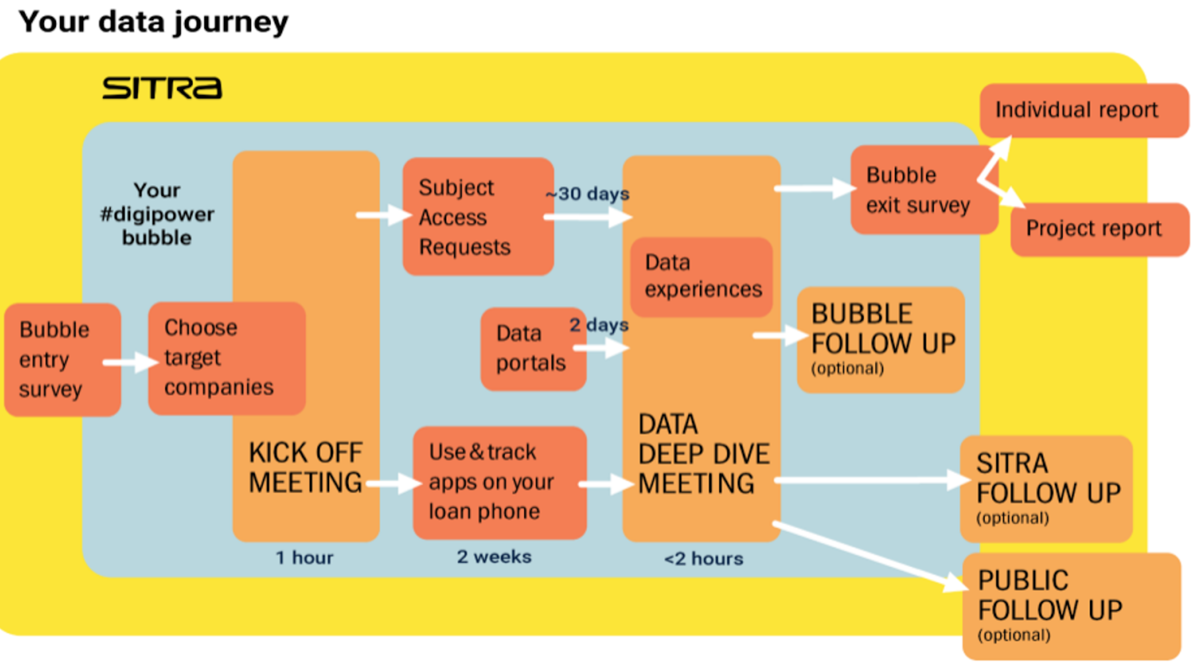


Figure 3: Diagram illustrating the participant's journey through the project, as initially planned

We will now explain the key stages and aspects of participant journeys through the investigation:

Recruitment: Participants were selected and recruited both through Sitra's network and through Paul-Olivier Dehaye's network, selected for their high profile status within or connected to the European data economy and digital policy space. For further details see Section 2.3.

Introduction Video: Recognising that our participants were busy, we prepared an introductory video[2] to explain the project to participants which they could watch in their own time.

Companies Survey: To support the participants in the process of selecting some suitable target companies, we invited participants to complete a survey that helped them to produce a shortlist of candidate companies by suggesting different categories of company (news, entertainment, telco, transport, food delivery etc.). Versions of this survey were created for Finland-, Belgium- and international-based participants.

Kick-Off Meeting: With each participant, the two lead investigators Paul-Olivier Dehaye and Alex Bowyer held a one-hour video call. During that call, following introductions and an explanation of the planned journey, participants outlined their specific profiles, interests and goals, finalised a target set of around 6 companies each, and rated those companies in terms of current trust level and expected transparency.

[2] https://vimeo.com/622388591/1636ecee8b

Discovery/Coaching: During the next 1-2 months (the GDPR response timeframe is 30 days, providing us with a minimum constraint for this phase), participants were guided through the process of using three different lenses to access data from their chosen target companies. The three lenses were as follows:

- Subject Access Requests
- Data Download Portals
- Technical Audit

These lenses are described in Section 2.5. Coaching consisted of regular check-ins with participants, over email, instant message and video calls; fielding questions; advising on how to respond to emails; and providing technical support with data access and with the technical audit (which involved loaning an Android phone to participants with dedicated app-monitoring software installed).

Data Deep Dive Meeting: After the discovery/coaching phase was complete, each participant gathered together their obtained data and either pre-shared it with the investigators, or prepared to describe and show parts of it using screen sharing. A two-hour video call took place between investigators and each participant, during which each of the companies' responses was discussed, including those that did not provide data. A variety of tools and techniques were used to help participants understand the data, including the use of bespoke-developed data visualisation tools (see Section 2.5), or preparing screenshots and extracts in advance. The discussions aimed to identify 'nuggets' as mentioned above and to identify potential future actions. Additionally the prior scores for trust and expected transparency were revisited to see if they had changed.

Followup Activities: The potential for followup activities was left very open-ended, including but not limited to: additional deep dive meetings to look at data in more detail, additional Subject Access Requests to find extra information, preparation of materials for public sharing on social media or reports, interviews with journalists and supporting those journalists, and more. In some cases, followup activities are still ongoing, and we see #digipower and our work with these participants very much as a beginning that could lead on to future investigations or collaborations.

Data Protection Measures: In order to minimise risks with any personal data the participants shared, as well as to protect their privacy, we devised a "bubble" concept. The investigators and the participant would keep any data files and acquired knowledge within a ring-fenced virtual bubble, ensuring that no-one else (including others in Sitra) could access the information without explicit participant consent. At all times, participants were free to share any part of their own data or their own experience outside of this bubble, but Hestia.ai and Sitra were not; we emphasised to participants that "It's your data, and your story to tell".

Consent: By default, all data and information entrusted to the investigators was treated as private. When information or data was shared with the investigators this was taken implicitly consent to view that information, but not to share it. Where we thought there was value in sharing a piece of data more widely, we explicitly sought consent from participants, which was logged, and when consent was given, participants were given choices as to whether information could be specifically attributed to them and whether companies should be named. In general, participants had two opportunities to consent to the publication of facts from their data and experiences: (i) prior to a story, visualisation or report segment being prepared, and (ii) after a visualisation or textual write up had been drafted. Where data was shown, care was taken to hide, blur or remove any elements the participants might or did find sensitive or did not wish to share.

2.2. STUDY PARTICIPANTS

The #digipower investigation had 15 participants from across Europe, including parliamentary representatives from the Finnish and European Parliaments, national and European civil service, journalism and a non-governmental organisation. In summary, the breakdown of participants was as follows:

- Five Finnish Members of Parliament (MPs),
- Two Finnish Members of the European Parliament (MEPs),
- One French Member of the European Parliament (MEPs),
- One Finnish and two European Commission (EC) civil servants,
- The Directors of SITRA and of an international non-governmental organisation (NGO) in Geneva, and
- Two journalists.

The participants and their backgrounds are detailed in Figure 4.

Participant	Position/Background/Location
Anders Adlercreutz	Finnish MP, The Swedish People's Party of Finland, Uusimaa/Helsinki, Finland.
Leïla Chaibi	French Member of the European Parliament, La France Insoumise, Paris, France and Brussels, Belgium.
Filomena Chirico	Civil Servant, Cabinet of Commissioner Thierry Breton, responsible for digital platform regulation, European Commission, Brussels, Belgium.
Christian D'Cunha	European Civil Servant working on cybersecurity and digital privacy in DG Connect, European Commission, Brussels, Belgium. Former head of private office of the European Data Protection Supervisor.
Stephane Duguin	CEO of non-governmental organisation CyberPeace Institute, Geneva, Switzerland, focusing on protecting vulnerable communities and citizens in the digital space. Background in civil law enforcement.
Atte Harjanne	Finnish MP, The Green League, Helsinki, Finland.
Jyrki Katainen	President of Sitra, Helsinki, Finland. Former European Commission Vice President (2011-2014). Former Prime Minister of Finland (2014-2019).
Miapetra Kumpula-Natri	Finnish Member of the European Parliament, Social Democratic Party, Vaasa, Helsinki, Finland and Brussels, Belgium. Former Finnish MP (2003-2011).
Dan Koivulaakso	State Secretary to the Finnish Minister of Education, Left Alliance, Helsinki, Finland.
Markus Lohi*	Member of Finnish Parliament, Centre Party. Lapland, Finland and Helsinki, Finland. Legal affairs director.
Tom Packalén*	Member of Finnish Parliament, The Finns Party, Helsinki, Finland. Former Chief Inspector of Helsinki Police. Entrepreneur.
Sirpa Pietikainen*	Finnish Member of European Parliament, National Coalition Party, Häme, Finland and Brussels, Belgium. Former Member of Finnish Parliament (1983-2003).
Mark Scott	Chief Technology Journalist at POLITICO, writing about the intersection of technology and politics. London, UK.
Niclas Storås	Journalist at Helsingin Sanomat. Helsinki, Finland.
Sari Tanus	Member of Finnish Parliament, Christian Democrats. Pirkanmaa and Helsinki, Finland.

Figure 4: Participants in the #digipower investigation (= reduced participation)*

These participants were selected as high-profile individuals (considered VIPs in this report) involved with, or closely connected to the space of digital policy in Europe. We started with a national and political perspective in Finland: all political parties were invited to participate in order to represent a broad political spectrum, though not all parties were ultimately able to put forward a representative. The participant pool was then extended beyond Finland in order to include VIPs with significant influence in policy making and thinking in the data economy and European politics.

2.3. TARGET SERVICE PROVIDERS

During the Kick Off meetings, participants discussed their survey answers and finalised a selection of 6 companies they wished to examine and obtain data from. This was largely a free choice by the participants, with some minor influence from the coaches to ensure (a) that our investigation did not unfairly burden any small companies, and (b) to encourage a wide variety of targets across different industries, and especially any unique or unusual targets and (c) to ensure a good balance of large tech companies, Finnish companies, Nordic companies, and smaller companies.

The final selection included the list of 15 companies presented below that were targeted by multiple participants , constituting a total 57 targeting occurrences. The list presents the flag of the specific headquarter country, the company, and the brand or app target in square brackets):

- 🇺🇸 Meta (9 participants) [6 Facebook, 2 WhatsApp, 1 Instagram]
- 🇺🇸 Google (9 participants)
- 🇫🇮 Sanoma Group (news organisation, 6 participants) [4 Helsingin Sanomat, 1 Iltasanomat, 1 Aamulehti]
- 🇺🇸 Apple (5 participants)
- 🇺🇸 Uber (4 participants)
- 🇺🇸 Twitter (4 participants)
- 🇸🇪 Spotify (3 participants)
- 🇺🇸 Netflix (3 participants)
- 🇺🇸 Microsoft (2 participants) [2 LinkedIn]
- 🇸🇪 Telia (Telecom and streaming media provider, 2 participants) [1 Telia, 1 🇫🇮 MTV3]
- 🇫🇮 HSL (Helsinki public transport, 2 participants)
- 🇫🇮 S Group / S-Ryhmä (supermarket chain, 2 participants)
- 🇫🇮 Finnair (2 participants)
- 🇳🇴 🇬🇧 Elkjøp/Currys plc [🇫🇮 Gigantti] (e-commerce/physical retailer, 2 participants)
- 🇺🇸 Signal (messaging app, 2 participants)

The following list below presents 25 service providers that were targeted by 1 participant only:

- 🇬🇧 BBC,
- 🇺🇸 Politico (Brussels-based newspaper),
- 🇺🇸 The New York Times,
- 🇺🇸 The Washington Post,
- 🇫🇷 Le Monde,
- 🇫🇮 YLE (public broadcaster),
- 🇫🇮 Kesko Group / K-Ryhmä (supermarket chain and retail group),
- 🇺🇸 Strava (fitness app),
- 🇳🇱 KLM,
- 🇫🇮 Wolt (food delivery app),
- 🇺🇸 FullContact (identity profiling company),
- 🇸🇪 Voi (e-scooter company),
- 🇧🇪 Mobile Vikings (phone service provider),
- 🇳🇱 Just Eat Takeaway [Thuisbezorgd] (food delivery company),
- 🇫🇷 SNCF (French railway operator),
- 🇫🇮 Stockmann (department store),
- 🇨🇭 Swiss Federal Railways,
- 🇩🇪 Zalando (online retailer),
- 🇬🇧 Deliveroo (food delivery app),
- 🇭🇰 Hutchison 3G [🇬🇧 Three] (mobile phone service provider),
- 🇧🇪 Colruyt (supermarket chain),
- 🇸🇪 Bookbeat (streaming service),
- 🇫🇮 Alma Media [Iltalehti] (news media company),
- 🇺🇸 🇳🇱 Booking (online travel agent),
- 🇸🇪 Bisnode [🇺🇸 Dun & Bradstreet] (data broker).

In total, 40 distinct service providers originating from 10 different countries were targeted, across a total of 83 distinct individual to service provider relationships.

2.4. LENSES ON THE DATA ECONOMY

As described in Section 2.2 above, the basic methodological premise of the investigation was to coach participants through the process of obtaining and examining their own data from each of their selected targets. The rationale was that the data economy should be more understandable when made relatable through one's own data. Furthermore, it is easier to share stories and learnings with others when they draw from your own experience. To achieve this, we relied on the legal rights available to any individual to access the personal data held by service providers via Subject Access Requests (SARs), primarily the EU & UK General Data Protection Regulations, which place a legal obligation on data holders to provide copies of personal data as well as explanations around data processing[3].

We wanted to maximise the potential insights for each participant on their data, and, recognising that not all data access methods work, we wanted to build redundancy and access data in multiple ways. For this reason we elected to equip our participants with three separate parallel “lenses" to view and understand their own provider-held data and the data flows of the organisations holding their data: (i) Subject Access Requests, (ii) Data Download Portals and (iii) Technical Audit with Tracker Control.

The differences between the three lenses are illustrated in Figure 5. Participants used all available lenses (typically all 3 but always Subject Access Request at a minimum, as this is a legal obligation for all companies to provide) according to what was available with each target. While some service providers oblige users to access their data via automated data download portals, others offer a contact email to a data protection officer. Each option provides different transparency and amounts of accessible data.

[3] Your data matters (overview of individual rights), Information Commissioner’s Office, UK, https://ico.org.uk/your-data-matters/

Aspect	**Subject Access Requests (SARs)**	**Data Download Portals**	**Tracker Control (technical audit)**
Nature of access	Zip files received in response to email or filled form.	Zip files received in response to button click on a self-service website, usually with a delay.	Information about the observed activity of apps on an Android phone, exported from the phone as a CSV.
Scope of data	Broad or deep data, as requested by user	Limited but easily accessible data, as chosen by company	Free, niche app, limited by operating system constraints and Google Play Store Policies
Response Timeframe	Up to 30 days, or 3 months if declared a complex request	A few hours or days.	Monitoring period is up to the user. (We recommended a minimum of 2 weeks).
Suitable Targets that can be Observed	Any organisation or service provider	Only those large tech companies that offer a portal	Any app installed on the Android operating system of the device where TrackerControl is installed
Type of Data that should be Available	Per GDPR, should include all data that could identify an individual, including historical data, as well as explanations and contextual information.	Companies present the idea that it is “all your data”, including historical data.	Only information about the data sharing from each app to third parties, from the phone, during the usage period. Companies contacted are listed, but information about what data is sent and received is not available.

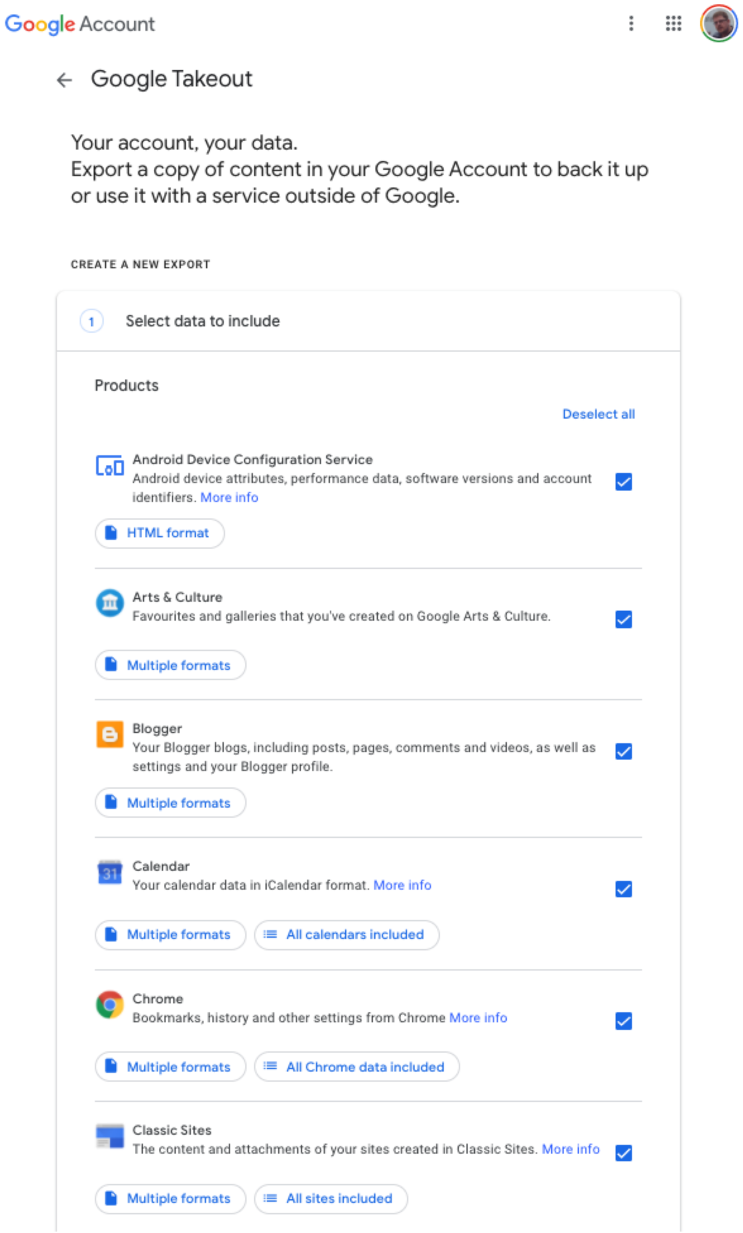
Figure 5: Summary of differences between the three lenses on the data economy

A key part of the investigation approach was that the participants would be *coached* through the use of these three approaches, which involved giving instructions, answering questions, giving advice in light of company actions, responses, or non-responses, following up on participants to check progress and prompt next actions, and, in the Data Deep Dive Meeting (i.e., where each participant pre-shared data with the investigators, or prepared to describe and show parts of it using screen sharing), understanding and assessing responses and any data and/or information returned.

The three lenses are described in the following three subsections. Section 2.4.4 comments on our approach for dealing with differing data formats, and 2.4.5 then explains the value drawn from combining these different lenses.

2.4.1. Data Download Portals

As a response to a growing number of data access requests, many companies have sought to circumvent the otherwise time-consuming and labour-intensive process of satisfying those requests by providing self-service tools for users to access their own data without the need for staff intervention. While the percentage of users actually making actionable GDPR requests is understood to be tiny, this percentage can still represent an unmanageable number of requests for companies with millions of users and small data protection teams. This is why companies such as Google and Facebook have implemented online portals: websites where users can log in and download zip files of their own data, which are typically available to all users regardless of the legal jurisdiction in which they reside. In most cases a choice of particular products and of data types is offered. A choice between machine-readable files in digital formats such as JSON, CSV or XML (useful for data portability or script-based analysis) versus human-readable HTML pages (more suitable for browsing and understanding) is usually offered. The two most well-known download portals are Google Takeout (Figure 6) and Facebook's "Download Your Information" tool (Figure 7).



Google Account

Google Takeout

Your account, your data.
Export a copy of content in your Google Account to back it up or use it with a service outside of Google.

CREATE A NEW EXPORT

1 Select data to include

Products

Deselect all

Android Device Configuration Service
Android device attributes, performance data, software versions and account identifiers. More info ☑
HTML format

Arts & Culture
Favourites and galleries that you've created on Google Arts & Culture. ☑
Multiple formats

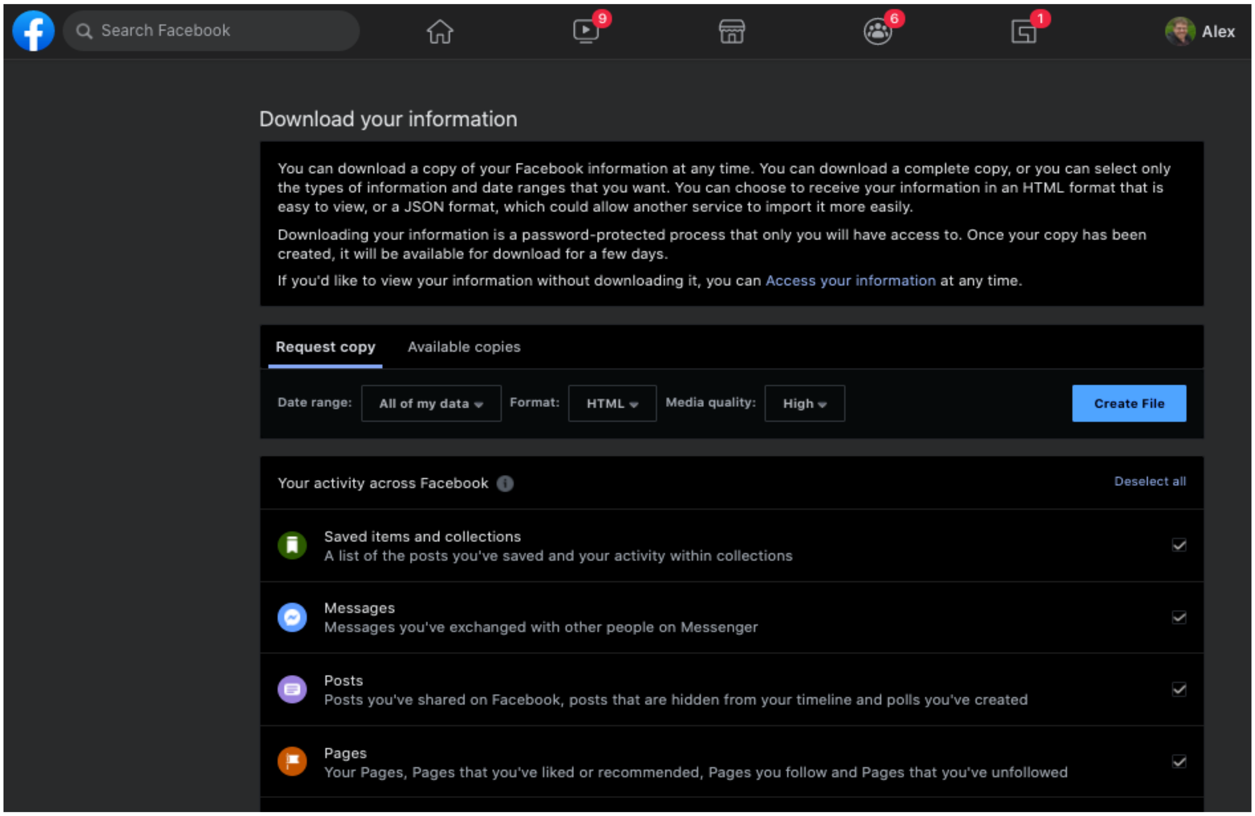
Blogger
Your Blogger blogs, including posts, pages, comments and videos, as well as settings and your Blogger profile. ☑
Multiple formats

Calendar
Your calendar data in iCalendar format. More info ☑
Multiple formats | All calendars included

Chrome
Bookmarks, history and other settings from Chrome More info ☑
Multiple formats | All Chrome data included

Classic Sites
The content and attachments of your sites created in Classic Sites. More info ☑
Multiple formats | All sites included

Figure 6. Google Takeout Download Portal.



Download your information

You can download a copy of your Facebook information at any time. You can download a complete copy, or you can select only the types of information and date ranges that you want. You can choose to receive your information in an HTML format that is easy to view, or a JSON format, which could allow another service to import it more easily.

Downloading your information is a password-protected process that only you will have access to. Once your copy has been created, it will be available for download for a few days.

If you'd like to view your information without downloading it, you can Access your information at any time.

Date range: All of my data Format: HTML Media quality: High

Your activity across Facebook

Saved items and collections
A list of the posts you've saved and your activity within collections ☑

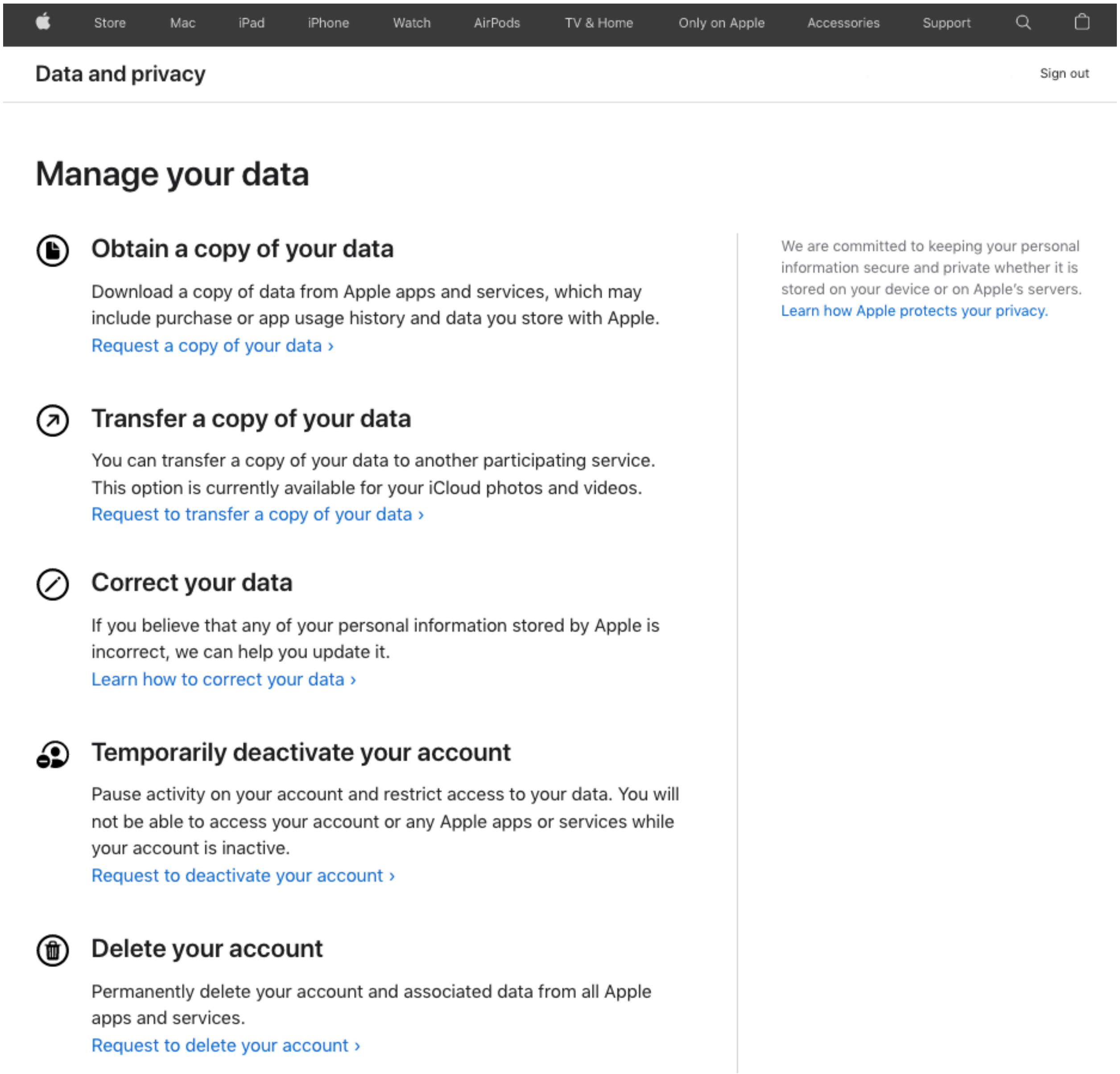
Messages
Messages you've exchanged with other people on Messenger ☑

Posts
Posts you've shared on Facebook, posts that are hidden from your timeline and polls you've created ☑

Pages
Your Pages, Pages that you've liked or recommended, Pages you follow and Pages that you've unfollowed ☑

Figure 7. Facebook's "Download Your Information" portal.

The exact parameters of what constitutes a download portal is blurry. Some companies, such as Netflix, Apple and Spotify, have a semi-automatic process for requesting data via a self-service button click, but this triggers a GDPR request in the background that probably still involves some human processing. Most data download portals focus on the right of access, but some, such as Apple's, allow the user to access some of their other GDPR rights as well, as shown in Figure 8:



Manage your data

Obtain a copy of your data

Download a copy of data from Apple apps and services, which may include purchase or app usage history and data you store with Apple.
Request a copy of your data ›

Transfer a copy of your data

You can transfer a copy of your data to another participating service. This option is currently available for your iCloud photos and videos.
Request to transfer a copy of your data ›

Correct your data

If you believe that any of your personal information stored by Apple is incorrect, we can help you update it.
Learn how to correct your data ›

Temporarily deactivate your account

Pause activity on your account and restrict access to your data. You will not be able to access your account or any Apple apps or services while your account is inactive.
Request to deactivate your account ›

Delete your account

Permanently delete your account and associated data from all Apple apps and services.
Request to delete your account ›

We are committed to keeping your personal information secure and private whether it is stored on your device or on Apple's servers.
Learn how Apple protects your privacy.

Figure 8. Apple's Data and Privacy portal, which offers more than just access.

Download portals are typically only available for major Internet-based companies, and typically most small or local/regional companies or companies focused more on bricks-and-mortar operations tend not to have the resources to offer download portals. For the purposes of this investigation, we considered a download portal to be any self-service offering that would allow users to ultimately receive a copy of their own data, regardless of the time period. In practice, some of these may actually have been GDPR Subject Access Requests behind the scenes, but standard GDPR requests were still conducted as well (see Section 2.4.2).

From the selected 40 target organisations, only the following 9 organisations had download portals available:

- Meta [for Facebook, Instagram and WhatsApp],
- Apple,
- Twitter,
- Spotify,
- Netflix,
- Bookbeat,
- Twitter,
- Microsoft [for LinkedIn], and
- Telia [for MTV3].

FullContact does have a data viewing portal, but we excluded this one as being useful as a lens for two reasons: firstly, this does not allow data download, only viewing, and secondly, testing of the portal revealed that usage of this portal could inadvertently provide additional data to the company, allowing them to link up different email addresses and phone numbers as corresponding to the same person, which would not be desirable. To avoid this risk, this portal was not used for the study as a download portal (though it was used to obtain information to help construct the Subject Access Request).

2.4.2. Subject Access Requests (SARs)

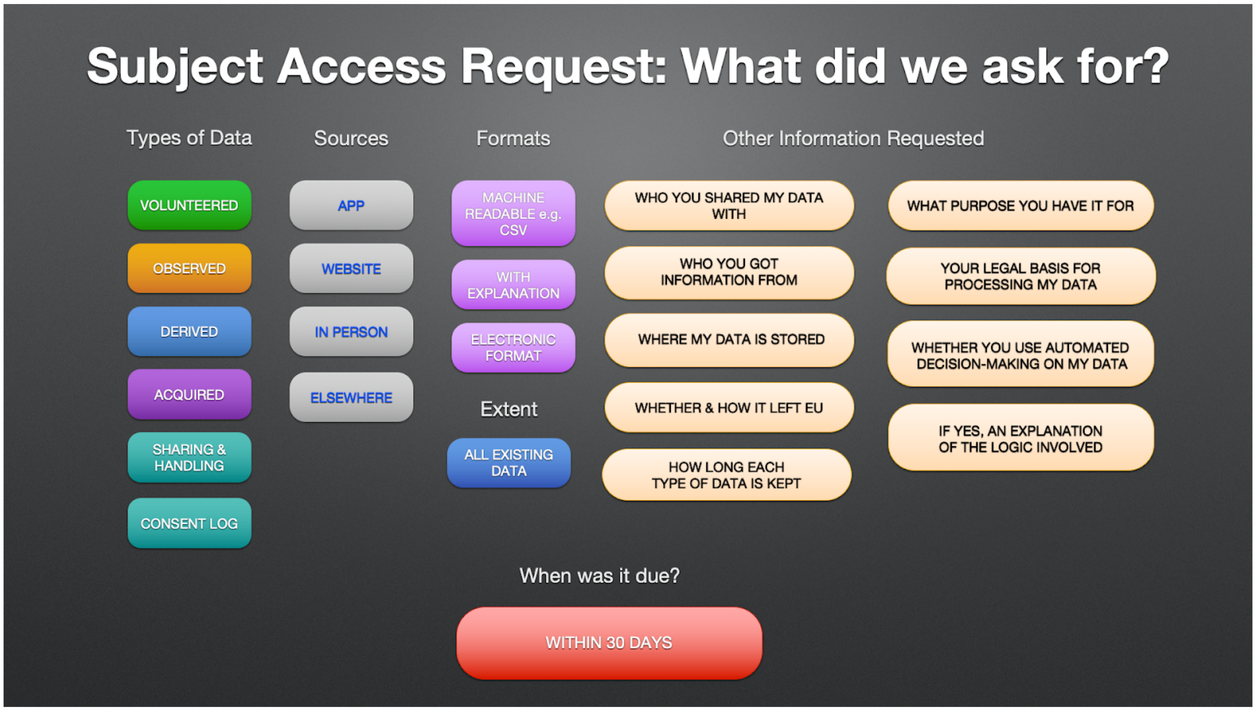
Subject Access Requests are a mechanism established in EU (and subsequently UK) law by the GDPR[4], which allow users to request a copy of their data. This is typically done by email, though in some cases, companies request or require the filling in of a paper or electronic form to initiate the process. After a request has been made, and the service provider has successfully identified the user's account and/or verified their identity, they then have 30 working days to respond with files and/or information per the user's request. A common issue encountered is to receive the answer "We do not have any personal data about you" after a SAR, which is a valid and legal response. However, based on our experience, this response can sometimes be sent in error by the service provider if the user does not make the request from the correct email address used when creating an account.

In order to produce consistent and effective SARs, we used a standard template email, developed by PersonalData.IO[5], that we extended and personalised according to the investigation and our knowledge of current data structures used and data types collected by service providers in parallel work (see Section 2.8). The SAR template mail was designed to

[4] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 27 April 2016
https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 EUR-LEX. Accessed March 2022.

[5] GDPR email template, personaldata.io, https://wiki.personaldata.io/wiki/Template:Access

thoroughly execute all of the GDPR's provisions relating to data access, information provision and explanation, as summarised in Figure 9 below:



Subject Access Request: What did we ask for?

Types of Data

- VOLUNTEERED
- OBSERVED
- DERIVED
- ACQUIRED
- SHARING & HANDLING
- CONSENT LOG

Sources

- APP
- WEBSITE
- IN PERSON
- ELSEWHERE

Formats

- MACHINE READABLE e.g. CSV
- WITH EXPLANATION
- ELECTRONIC FORMAT

Extent

- ALL EXISTING DATA

Other Information Requested

- WHO YOU SHARED MY DATA WITH
- WHO YOU GOT INFORMATION FROM
- WHERE MY DATA IS STORED
- WHETHER & HOW IT LEFT EU
- HOW LONG EACH TYPE OF DATA IS KEPT
- WHAT PURPOSE YOU HAVE IT FOR
- YOUR LEGAL BASIS FOR PROCESSING MY DATA
- WHETHER YOU USE AUTOMATED DECISION-MAKING ON MY DATA
- IF YES, AN EXPLANATION OF THE LOGIC INVOLVED

When was it due?

- WITHIN 30 DAYS

Figure 9: A summary of the different types, sources and formats of data our GDPR requests asked for, along with the other information requested

We identified specific categories of data that should be returned (see Section 2.6). We were explicit that all data should be returned, from all sources - apps, websites, devices, in-person visits and external sources, in technical machine readable formats with explanations. We also were explicit to make the request both a Subject Access Request and a Data Portability Request, which further increases the scope of what should be returned.

This set of requests combines the legal obligations from a number of different Articles of the GDPR. The GDPR currently applies to any customer who resides in the European Union, and any customer of a business that operates to any degree in the European Union.

The United Kingdom also maintains a copy of European Union GDPR law so the regulations apply identically there. In the case of a participant living in the UK we adapted the wording of our template accordingly.

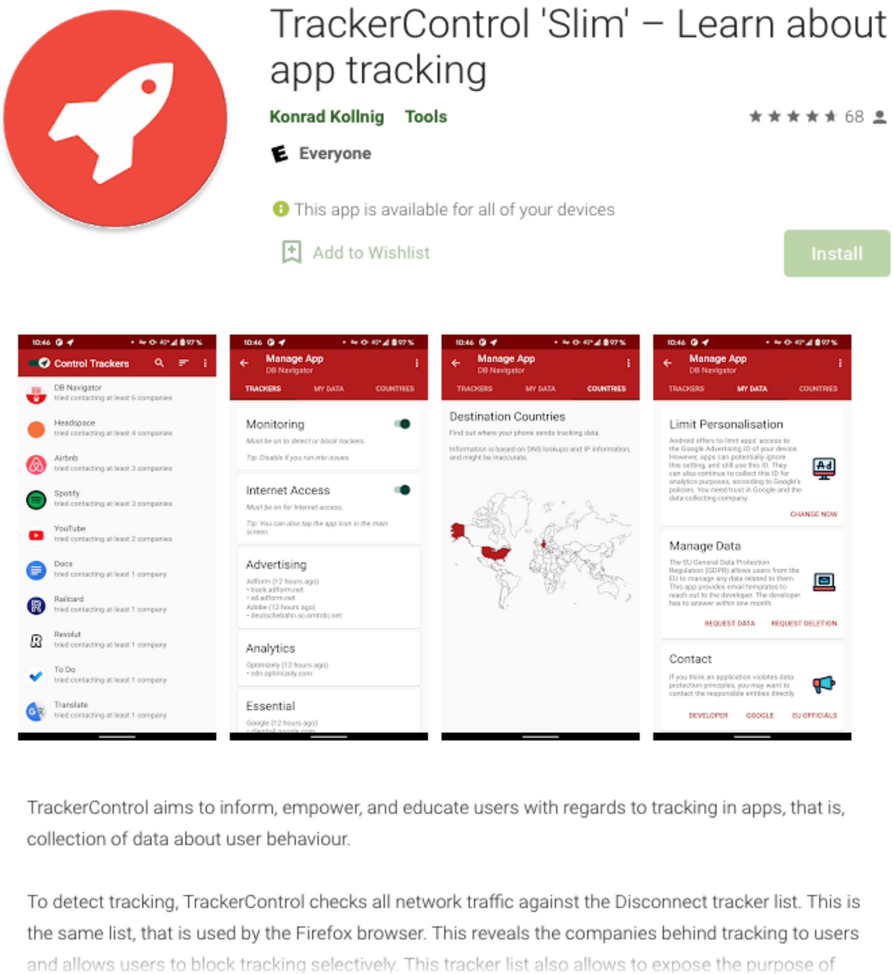
In the case of a participant living in Switzerland (outside the EU) we adapted our template to make reference to the LPD (Ordnance/Loi Fédérale sur la protection des Données[6]) which is the nearest equivalent law. We note that typically many companies will respond to GDPR requests

[6] https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en

even from non-EU, non-UK residents, so in practice the ability to make Subject Access Requests is usually available to the global population, or at least the population of the European continent.

2.4.3. Technical Audit: Tracker Control

Our third lens is a technical audit based on the use of the app TrackerControl installed on Android phones (Figure 10). Here we took a different approach: Rather than ask a target company to produce data or see what data they willingly share, we would explore how much we could learn by observing the behaviour of each target company's mobile app. Each participant was loaned an Android phone with this software pre-installed, and invited to use it for some common tasks for a minimum of 2 weeks alongside their regular phone.



TrackerControl 'Slim' – Learn about app tracking

Konrad Kollnig   Tools

68

Everyone

This app is available for all of your devices

Control Trackers

DB Navigator
tried contacting at least 6 companies

Headspace
tried contacting at least 4 companies

Airbnb
tried contacting at least 3 companies

Spotify
tried contacting at least 3 companies

YouTube
tried contacting at least 2 companies

Docs
tried contacting at least 1 company

Railcard
tried contacting at least 1 company

Revolut
tried contacting at least 1 company

To Do
tried contacting at least 1 company

Translate
tried contacting at least 1 company

Manage App
DB Navigator

Monitoring
Must be on to detect or block trackers.
Tip: Disable if you run into issues.

Internet Access
Must be on for Internet access.
Tip: You can also tap the app icon in the main screen.

Advertising
Adform (12 hours ago)
• track.adform.net
• ad.adform.net
Adobe (12 hours ago)
• deutschebahn.sc.omtrdc.net

Analytics
Optimizely (12 hours ago)
• cdn.optimizely.com

Essential
Google (12 hours ago)

Destination Countries
Find out where your phone sends tracking data.
Information is based on DNS lookups and IP information, and might be inaccurate.

Limit Personalisation
Android offers to limit apps' access to the Google Advertising ID of your device. However, apps can potentially ignore this setting, and still use this ID. They can also continue to collect this ID for analytics purposes, according to Google's policies. You need trust in Google and the data collecting company.

Manage Data
The EU General Data Protection Regulation (GDPR) allows users from the EU to manage any data related to them. This app provides email templates to reach out to the developer. The developer has to answer within one month.

Contact
If you think an application violates data protection principles, you may want to contact the responsible entities directly.

TrackerControl aims to inform, empower, and educate users with regards to tracking in apps, that is, collection of data about user behaviour.

To detect tracking, TrackerControl checks all network traffic against the Disconnect tracker list. This is the same list, that is used by the Firefox browser. This reveals the companies behind tracking to users and allows users to block tracking selectively. This tracker list also allows to expose the purpose of

Figure 10. TrackerControl App Installation.

This approach was enabled by an Android app called Tracker Control, developed by Oxford University researcher Konrad Kollnig[7]. What this app does is simple: it watches the outgoing connections made by all apps on the phone, and sees what domains those apps are connecting to. Seeing those connections and domains is done by acting as a local Virtual Private Network (VPN), inserting itself into the address resolution loop at an operating system level.

Subsequently similar functionality has been integrated into Tracker Control iOS 14.5[8], and another similar app, AppChk[9] is now available on iOS, but these emerged too late to be used by this study, therefore our focus was on tracking app behaviours on Android phones using TrackerControl.

TrackerControl's output lists the times of network activity, and the domains contacted, and which app made contact. Based on commonly-established knowledge and databases such as Exodus Privacy[10], the domains can be identified as to their likely purpose: Content, Fingerprinting, Advertising, Analytics, Social etc. It is important to note that TrackerControl cannot see the content of communications or what data is transferred in which direction, so the presence of a link between an app and a particular domain is only conclusive as to the purpose of the communication where the domain is absolutely unambiguous.

We used data exploration interfaces developed by Hestia.ai[11] (see Section 2.5) to review and explore the TrackerControl data for our participants.

2.4.4. Working with Disparate Data Formats

Every service provider provides different data formats to users. Moreover, every data access lens we exploited multiplied the quantity and variety of data formats received. This disparity of data formats required a lot of additional work in order to process the data and analyse it.

The most structured data came from Tracker Control. It produces CSV data, which can easily be exploited. The formats of all the data download portals are the same for everyone at any time, but we encountered the difficulty of getting a first dataset for producing the first visualisation, for example in the case of Bookbeat where a premium subscription was required to access download capabilities, or Telia's MTV3, whose download portal is optimised for Finns in Finland. Additionally, the formats from an individual provider were observed to change from month to month in several cases. The Subject Access Requests provided data in an ever more creative range of data formats. One of our tools (see Section 2.5) was built to address the "generic" case of visualising previously unseen data.

[7] TrackerControl, https://trackercontrol.org/
[8] Alex Hern, Apple iOS 14.5 updated includes tracking features, The Guardian, April 2022, https://www.theguardian.com/technology/2021/apr/27/apple-ios-145-update-includes-app-tracking-transparency-feature
[9] AppChk, https://appchk.de/
[10] Exodus Privacy, https://exodus-privacy.eu.org/en/
[11] HestiaLabs Experiences, https://digipower.hestialabs.org/ , accessed March 2022.

2.4.5. Combining the Lenses

The three lenses on the data economy, previously described, were complementary to each other in enabling this investigation. Indeed, separately, each lens only provides a partial view on each individual's personal data held by a service provider. There are three observations drawn from the combination of lenses:

- Data download portals only exist for the largest service providers, but Subject Access Requests would enable us to access data for a broad range of companies (in theory, because not all service providers respond in the time period stipulated by law)
- For both data download portals and Subject Access Requests, the response might be incomplete (or even non-existent), and the TrackerControl lens provides cross-checking redundancy with responses obtained from SARs, for instance.
- TrackerControl provides a cross-cutting view of all apps in a standardised form, enabling direct comparisons between different apps and between different participants even where they have targeted different companies.

However, the lenses' complementarity does not fully solve a methodological challenge: in general the personal data ecosystem has many intermediaries that intervene between the raw data collection (outbound) and the eventual usage of some derived data to take decisions back at an individual level (inbound). Generally speaking, between the outbound and inbound there is a profiling step, and data might transit through intermediaries. It might be difficult to surface this information through Subject Access Requests, Data Download Portals and TrackerControl since those lenses are fundamentally centred on the individual. This individual-centric approach has two drawbacks:

- Data holding services are mostly silent about what happens in the intermediary ecosystem, and
- These services are systematically biassed towards providing more information about the outbound part than the inbound part because their expertise, often protected by commercial/industrial trade secrets or intellectual property, and the value exploited from data, relies on the inbound part.

The resulting personal data visibility gap occurs because of a preference of data controllers to be more transparent about what the individual has provided directly than whatever has been derived subsequently to that initial data collection, despite similar legal obligation covering both types. So in theory and practice, better enforcement of the GDPR should ensure a broader scope of transparency, a concept as we develop further in Sections 3.3 (*Participants chasing their data*) and 4 (*Coaches' Reflections*).

Similarly, the first drawback could be mitigated through better enforcement of the GDPR: in theory during transfers of data (in the situation of two organisations acting as joint controllers, for instance) the GDPR is supposed to guarantee access to the information needed to be able to trace the chain of intermediaries.

2.5. DATA ANALYSIS AND VISUALISATION TOOLS FOR DATA UNDERSTANDING

Hestia.ai developed a range of web-based tools in order to provide data analysis and viewing experiences. They were used within #digipower Data Deep Dive meetings (see Section 2.1.2) to provide participants with visualisations for better understanding their data. These experiences are available online to the public on an 'as-is' basis[12]. In Figure 11, an overview of some of the available experiences are shown. Each experience allows the user to load in their local computer their data files (typically those returned by GDPR Subject Access Requests or Data Download Portals) and explore the information within without sharing the data with a third-party, not even Hestia.ai.

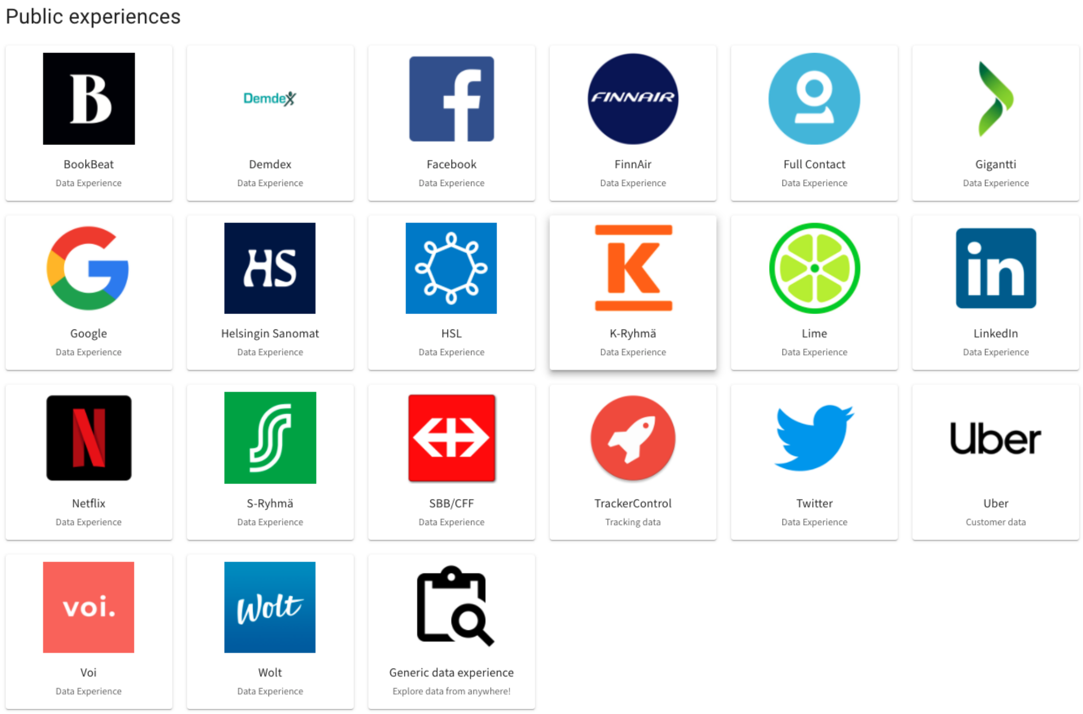


Figure 11. A dashboard showing some of the data-viewing experiences developed by HestiaLabs, which were used in the Deep Dive to help participants understand their data. Other data-viewing experiences not shown here include Apple, Spotify, Signal and WhatsApp.

It is important to note that these tools do not send any data to any server on the Internet. All data remains local within the browser, in order to maximise privacy. At their most basic level, these tools provide capabilities to view CSVs as tables and JSON files as collapsible tree

[12] https://digipower.hestialabs.org/

structures, as well as providing some searching and filtering capabilities. Given that with GDPR requests people will often be viewing data in formats that have never been seen before by the coaches or by Hestia.ai's developers, the viewing experiences were built to fall back to this generic 'file viewing' capability as a minimum, to ensure we could support all companies' returned data.

As part of the #digipower investigation, we improved on those visualisations as new requirements emerged for analysing the participants' data. This is an additional contribution of the investigation: providing new tool features so civil society can also analyse the data obtained from data access requests. The additional features available to the public are:

i) to look for events on data files according to date;

ii) to provide a time-based data overview in the form of a timeline and time-series-graph;

iii) to look for geolocated data;

iv) provide a geographical data overview on a map;

v) in the case of certain companies, namely Google, Facebook, Twitter, Netflix, LinkedIn, and Uber, to offer additional specific functionalities giving views and insights from those specific companies' returned data formats;

vi) an interface for viewing TrackerControl data as shown in Figure 12.

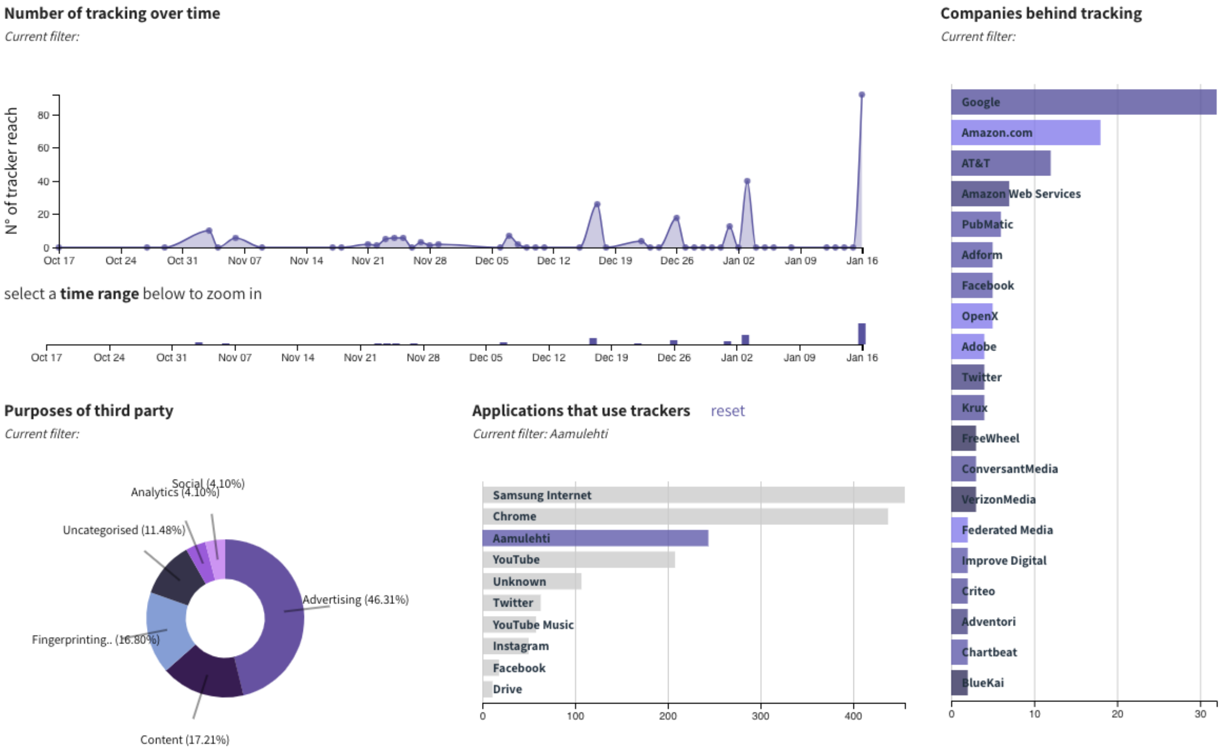


Figure 12: HestiaLabs' viewing experience of Tracker Control showing data collected from Sari Tanus' #digipower loan phone.

In Figure 12 above one can see the highlighted bar in the centre middle showing that we are looking at tracked communication events for Sanoma Group's Aamulehti news app. On the right, the domains contacted are grouped by company (and by type in the bottom left pie); this allows detailed technical audit to be carried out, including validation of Subject Access Request claims and cross comparison between participants.

These experiences, with their respective tool features, have an important pedagogical function, in providing a means for analysing data collection patterns so they are understood by an individual and explained to others, which are ultimately critical to legibility[13].

2.6. DATA TYPOLOGY

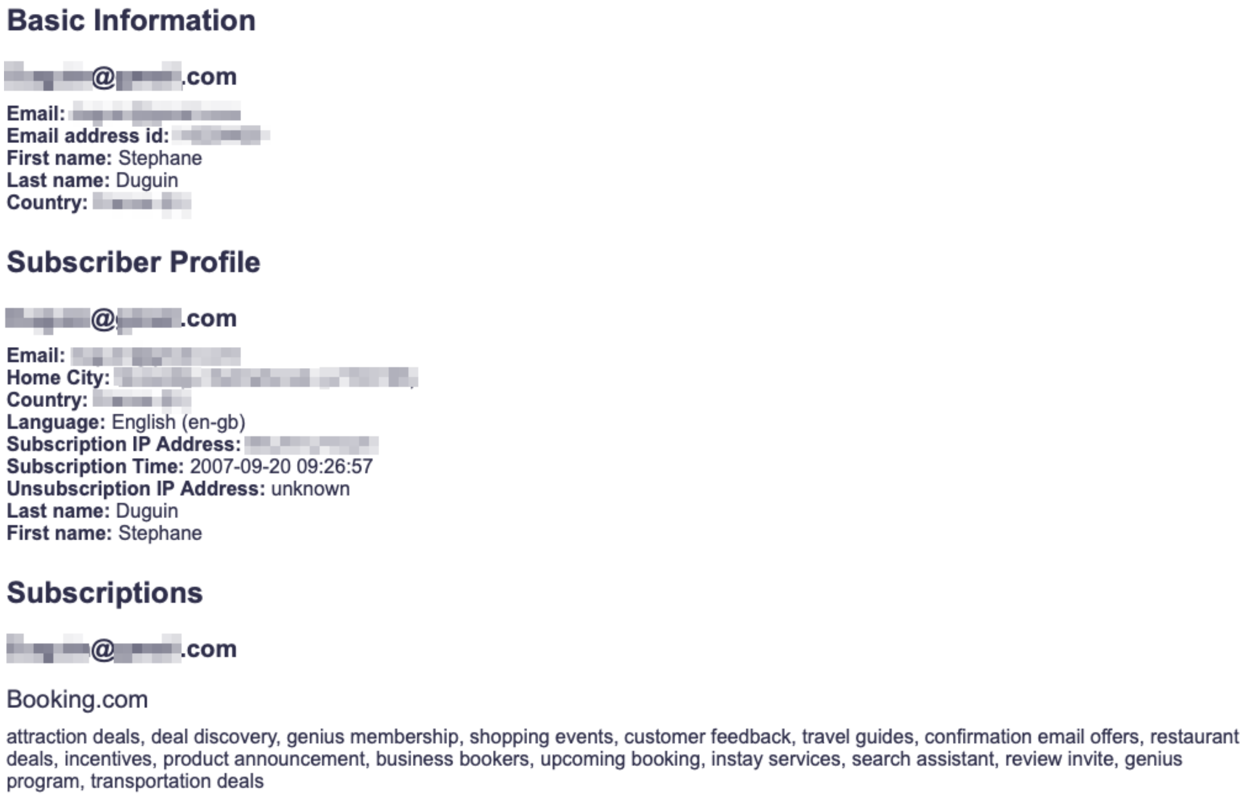
To understand and differentiate the various types of data that companies hold, and to communicate to participants the data they received, we made use of the data typology offered

[13] Mortier, Richard, et al. "Human-data interaction: The human face of the data-driven society." Available at SSRN 2508051 (2014). https://arxiv.org/pdf/1412.6159.pdf

in Bowyer *et al.* 2022[14] which identifies five distinct data types: volunteered data, observed data, derived data, acquired data and metadata. We found examples of all five types amongst participants' returned data, which we will now use to illustrate these terms:

Volunteered data refers to information knowingly provided by the user. This could include registration information as well as user profiles they have filled in. It also includes user-generated content, such as social media posts, shared photographs or cloud file backups, and a user's stored preferences.

For example, in Figure 13 one can see volunteered data held by travel firm Booking about Stephane Duguin, which he was able to retrieve via SAR. His volunteered data includes basic contact details (e.g., first and last name, home city), as well as his email subscription preferences.



Basic Information

[redacted]@[redacted].com

Email: [redacted]
Email address id: [redacted]
First name: Stephane
Last name: Duguin
Country: [redacted]

Subscriber Profile

[redacted]@[redacted].com

Email: [redacted]
Home City: [redacted]
Country: [redacted]
Language: English (en-gb)
Subscription IP Address: [redacted]
Subscription Time: 2007-09-20 09:26:57
Unsubscription IP Address: unknown
Last name: Duguin
First name: Stephane

Subscriptions

[redacted]@[redacted].com

Booking.com

attraction deals, deal discovery, genius membership, shopping events, customer feedback, travel guides, confirmation email offers, restaurant deals, incentives, product announcement, business bookers, upcoming booking, instay services, search assistant, review invite, genius program, transportation deals

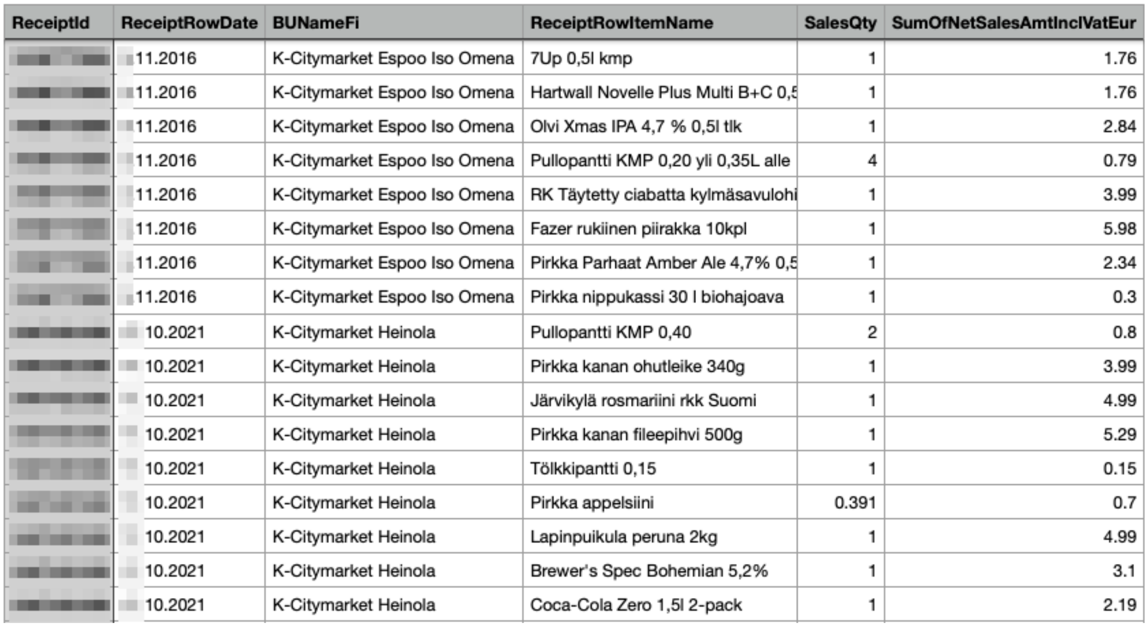
Figure 13: Volunteered data that Booking hold about Stephane Duguin, according to SAR return.

Observed data is data that is directly observed about the users or their devices, typically automatically or as an indirect result of a user's action. Such data collection can be explicit and consented, but sometimes is collected in the background or is unknowingly provided. Examples would be capturing a user's location when they perform certain actions, or as they move about

[14] Alex Bowyer *et al*. 2022. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In CHI Conference on Human Factors in Computing Systems (CHI '22), April 29–May 05, 2022, New Orleans, LA, USA. https://doi.org/10.1145/3491102.3501947

their day; or keeping a record of a user's purchases, physical store or website visits or support interactions. It would also include staff notes and observations from physical visits or calls.

For example, Jyrki Katainen's SAR return from Kesko Group shows that across all Kesko outlets, detailed records are kept of the purchases he and his household make, both at an itemised level (see Figure 14) as well as grouped by product type and by Kesko sub-brand.

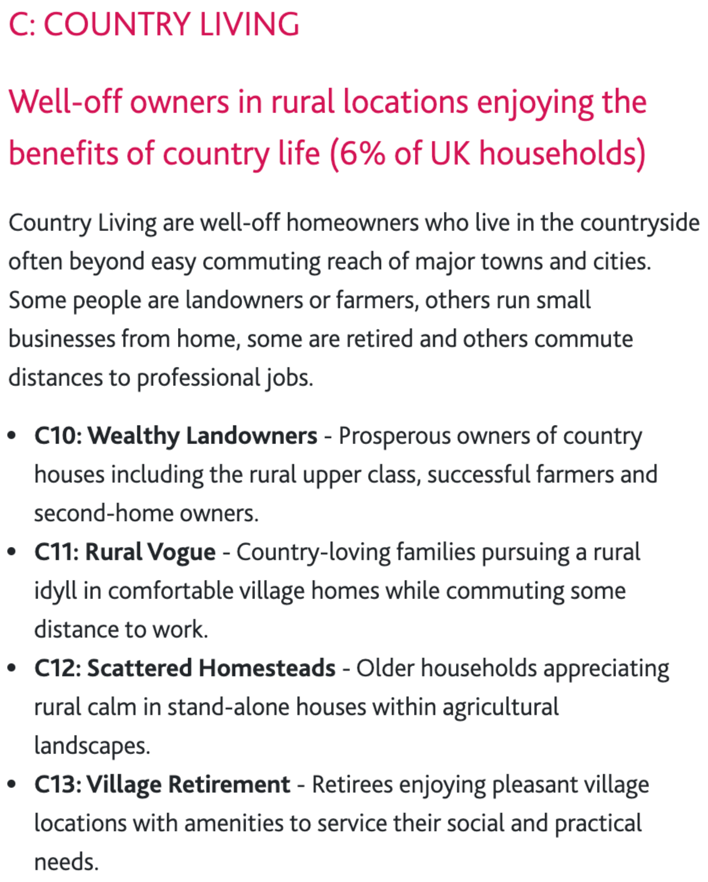


ReceiptId	ReceiptRowDate	BUNameFi	ReceiptRowItemName	SalesQty	SumOfNetSalesAmtInclVatEur
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	7Up 0,5l kmp	1	1.76
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	Hartwall Novelle Plus Multi B+C 0,5	1	1.76
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	Olvi Xmas IPA 4,7 % 0,5l tlk	1	2.84
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	Pullopantti KMP 0,20 yli 0,35L alle	4	0.79
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	RK Täytetty ciabatta kylmäsavulohi	1	3.99
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	Fazer rukiinen piirakka 10kpl	1	5.98
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	Pirkka Parhaat Amber Ale 4,7% 0,5	1	2.34
[redacted]	[redacted]11.2016	K-Citymarket Espoo Iso Omena	Pirkka nippukassi 30 l biohajoava	1	0.3
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Pullopantti KMP 0,40	2	0.8
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Pirkka kanan ohutleike 340g	1	3.99
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Järvikylä rosmariini rkk Suomi	1	4.99
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Pirkka kanan fileepihvi 500g	1	5.29
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Tölkkipantti 0,15	1	0.15
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Pirkka appelsiini	0.391	0.7
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Lapinpuikula peruna 2kg	1	4.99
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Brewer's Spec Bohemian 5,2%	1	3.1
[redacted]	[redacted]10.2021	K-Citymarket Heinola	Coca-Cola Zero 1,5l 2-pack	1	2.19

Figure 14: As with every interaction they make with Kesko, the details of grocery purchases from Kesko stores made by the members of Jyrki Katainen's household are kept in Kesko's database, as visible in his SAR return.

Acquired data is data that has been obtained by the service provider from a third party, often in exchange for payment or access to the customer's data. This might include civic checks such as electoral records, criminal record checks or child protection checks, as well as credit checks, personal recommendations, or information provided by marketers and advertisers.

For example, in Gigantti's SAR return to Miapetra Kumpula-Natri, they explained that as a loyalty club member, they had acquired third party data to place her in a segment for marketing purposes. They did not explain this further, but from the PNG image they sent we could see that they had contacted credit check firm Experian, which had told them that her postcode had a Mosaic segmentation of **C10.** This is data that Experian makes available to companies wishing to better understand their customers; it is their analysis of the residents of the postal code area identified from the customer's address. As Figure 15 shows, C10 corresponds to "Wealthy Landowners".



C: COUNTRY LIVING

Well-off owners in rural locations enjoying the benefits of country life (6% of UK households)

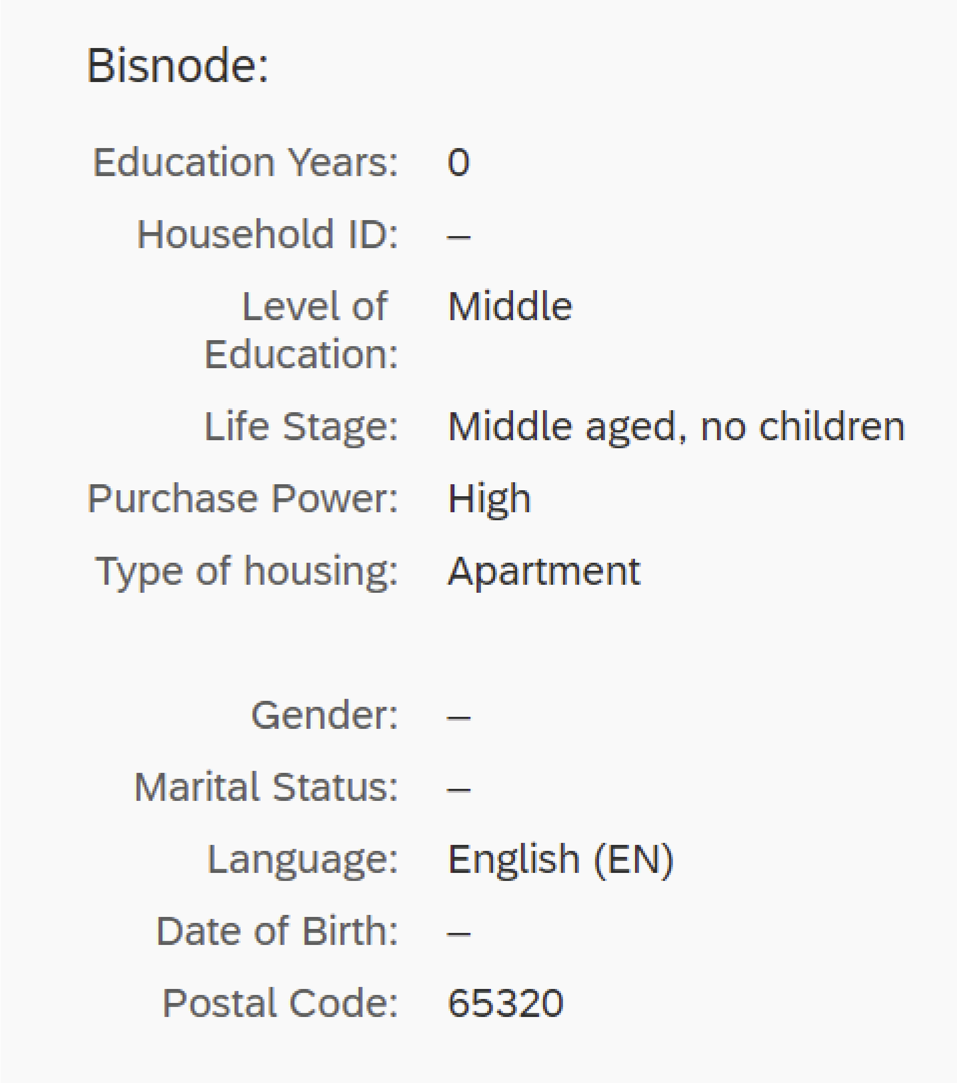
Country Living are well-off homeowners who live in the countryside often beyond easy commuting reach of major towns and cities. Some people are landowners or farmers, others run small businesses from home, some are retired and others commute distances to professional jobs.

- **C10: Wealthy Landowners** - Prosperous owners of country houses including the rural upper class, successful farmers and second-home owners.
- **C11: Rural Vogue** - Country-loving families pursuing a rural idyll in comfortable village homes while commuting some distance to work.
- **C12: Scattered Homesteads** - Older households appreciating rural calm in stand-alone houses within agricultural landscapes.
- **C13: Village Retirement** - Retirees enjoying pleasant village locations with amenities to service their social and practical needs.

Figure 15: Extract of Experian's Mosaic consumer segmentation model[15], which Gigantti used to understand that Miapetra Kumpula-Natri was a second home owner.

Gigantti also acquired data calculated by third party Bisnode about her (as shown in Figure 16).

[15] https://www.theaudienceagency.org/insight/mosaic



Bisnode:

Education Years:	0
Household ID:	–
Level of Education:	Middle
Life Stage:	Middle aged, no children
Purchase Power:	High
Type of housing:	Apartment
Gender:	–
Marital Status:	–
Language:	English (EN)
Date of Birth:	–
Postal Code:	65320

Figure 16: Acquired profiling data reported to Miapetra Kumpula-Natri within her Gigantti SAR return.

Derived data is new data about users, created by the data holder or a third party acting on their behalf, from the analysis of the data they have collected about a customer. This might include inferring a geographical location from an IP address, or trying to analyse the pattern of a customer's spending to determine their likely future spending, or examining the news articles a customer reads in order to determine what topics they are interested in.

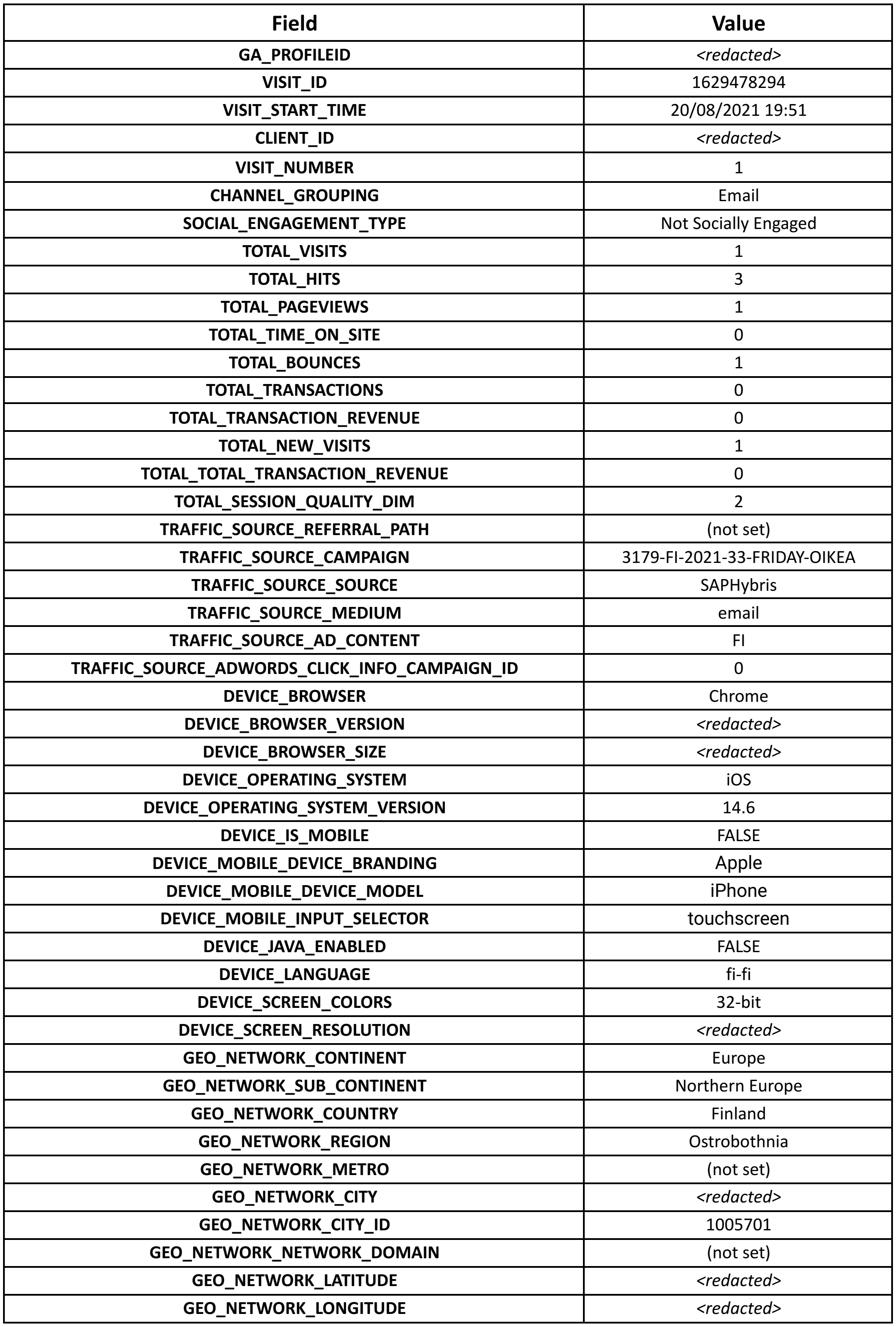
For example, each month, the retail group Kesko calculates, for each Plussa loyalty scheme member, their affinity with five different 'types' of users, based on their spending habits. For example, for December 2019 Kesko assigned Jyrki Katainen (more precisely, to, the aggregate spending of his household) the following customer types with their respective probability percentage:

Customer Type	**Probability of being this type of customer**
Enthusiast	17%
Indulger	12%
Woke	20%
Comfort-seeker	38%
Established/conservative	13%

Figure 17: Kesko's classification of customer categories applied to Jyrki Katainen's household for the month of December 2019, as seen in SAR return data.

Metadata is the final data type covered in the investigation. It encompasses additional data about the data collected from the users and their devices, which would include additional details captured at the time of a user action, such as the details of the device or browser used, the network connection used, the location and precise time at which the action occurred. Metadata provides more information about the context from which data was collected, as well as information about how the stored data is processed.

For example, one of the files returned in Miapetra Kumpula-Natri's Gigantti data shows that Google Analytics SDKs are used to capture a large volume of data about her actions as a customer on the Gigantti website. Figure 18 shows an example of the different metadata captured about just one visit she made to the Gigantti website. These details and more, 74 field values in total, are present for every Gigantti website visit she has made over a 3 year period.

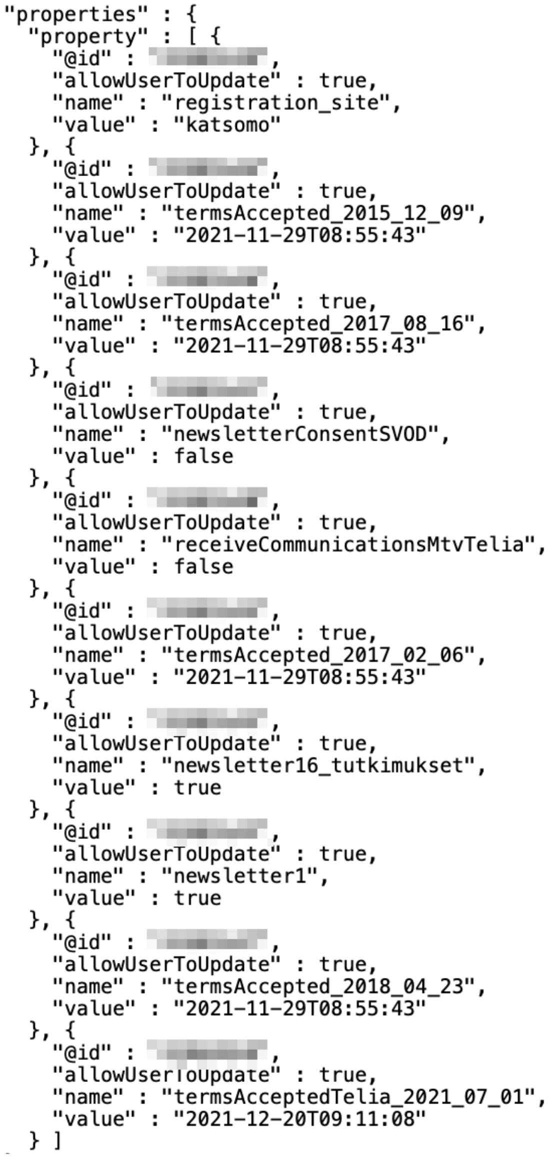


Field	Value
GA_PROFILEID	*<redacted>*
VISIT_ID	1629478294
VISIT_START_TIME	20/08/2021 19:51
CLIENT_ID	*<redacted>*
VISIT_NUMBER	1
CHANNEL_GROUPING	Email
SOCIAL_ENGAGEMENT_TYPE	Not Socially Engaged
TOTAL_VISITS	1
TOTAL_HITS	3
TOTAL_PAGEVIEWS	1
TOTAL_TIME_ON_SITE	0
TOTAL_BOUNCES	1
TOTAL_TRANSACTIONS	0
TOTAL_TRANSACTION_REVENUE	0
TOTAL_NEW_VISITS	1
TOTAL_TOTAL_TRANSACTION_REVENUE	0
TOTAL_SESSION_QUALITY_DIM	2
TRAFFIC_SOURCE_REFERRAL_PATH	(not set)
TRAFFIC_SOURCE_CAMPAIGN	3179-FI-2021-33-FRIDAY-OIKEA
TRAFFIC_SOURCE_SOURCE	SAPHybris
TRAFFIC_SOURCE_MEDIUM	email
TRAFFIC_SOURCE_AD_CONTENT	FI
TRAFFIC_SOURCE_ADWORDS_CLICK_INFO_CAMPAIGN_ID	0
DEVICE_BROWSER	Chrome
DEVICE_BROWSER_VERSION	*<redacted>*
DEVICE_BROWSER_SIZE	*<redacted>*
DEVICE_OPERATING_SYSTEM	iOS
DEVICE_OPERATING_SYSTEM_VERSION	14.6
DEVICE_IS_MOBILE	FALSE
DEVICE_MOBILE_DEVICE_BRANDING	Apple
DEVICE_MOBILE_DEVICE_MODEL	iPhone
DEVICE_MOBILE_INPUT_SELECTOR	touchscreen
DEVICE_JAVA_ENABLED	FALSE
DEVICE_LANGUAGE	fi-fi
DEVICE_SCREEN_COLORS	32-bit
DEVICE_SCREEN_RESOLUTION	*<redacted>*
GEO_NETWORK_CONTINENT	Europe
GEO_NETWORK_SUB_CONTINENT	Northern Europe
GEO_NETWORK_COUNTRY	Finland
GEO_NETWORK_REGION	Ostrobothnia
GEO_NETWORK_METRO	(not set)
GEO_NETWORK_CITY	*<redacted>*
GEO_NETWORK_CITY_ID	1005701
GEO_NETWORK_NETWORK_DOMAIN	(not set)
GEO_NETWORK_LATITUDE	*<redacted>*
GEO_NETWORK_LONGITUDE	*<redacted>*

Figure 18: An extract of the metadata collected about one of Miapetra Kumpula-Natri's visits to the Gigantti website, as shown in one row of a spreadsheet in her SAR return. Only a subset of the 74 fields are shown, and many field values are redacted to protect the participant's privacy.

The reason why metadata is contextual data is because it encompasses details of what has happened to the data after it was stored or produced: for instance, a history of consents the user has given about how data can be used and shared, a list of edits of that data, or of times the data was used or of times the data was shared with third parties.

Figure 19 presents a technical file in Dan Koivulaakso's SAR return from MTV3 (Telia), which shows a dated list of the times he has accepted certain terms and conditions or expressed preferences.



```
"properties" : {
  "property" : [ {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "registration_site",
    "value" : "katsomo"
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "termsAccepted_2015_12_09",
    "value" : "2021-11-29T08:55:43"
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "termsAccepted_2017_08_16",
    "value" : "2021-11-29T08:55:43"
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "newsletterConsentSVOD",
    "value" : false
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "receiveCommunicationsMtvTelia",
    "value" : false
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "termsAccepted_2017_02_06",
    "value" : "2021-11-29T08:55:43"
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "newsletter16_tutkimukset",
    "value" : true
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "newsletter1",
    "value" : true
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "termsAccepted_2018_04_23",
    "value" : "2021-11-29T08:55:43"
  }, {
    "@id" : [redacted],
    "allowUserToUpdate" : true,
    "name" : "termsAcceptedTelia_2021_07_01",
    "value" : "2021-12-20T09:11:08"
  } ]
```

Figure 19: Extract from a technical file returned in Dan Koivulaakso's SAR return from MTV3/Telia, showing a history of the consents he has given.

2.7. REPRODUCIBILITY

This investigation was conducted in a highly replicable way, from two perspectives.

From an **engineering perspective**, the software used and improved to analyse and visualise data obtained through download portals or subject access requests is available at https://github.com/hestiaai under open licences.

From a **pedagogical perspective**, by situating the experiences in individuals' personal data, we have focused on making it possible for any participant in the *#digipower* methodology to first acquire and then spread knowledge and skills about the data economy. Coupled with the previous software-focused point, this should drastically lower the costs of spreading such active knowledge in the long term, for instance by deploying this methodology further through "train-the-trainers" programs, i.e., a person is trained to be the trainer of somebody else for building communities of experts in civil society.

This program dynamic has proven to be easily adopted with a wide spreading effect, for instance journalist and participant Mark Scott used it his *Digital Bridge* newsletter[16], as well as in the *Tracking Free Ads Coalition*[17], both encouraging European civil servants and elected officials to investigate how they are being targeted on Twitter.

2.8. PREVIOUS WORK

In this section, we outline the various earlier and related works from both academic and industry research that have informed the #digipower investigation.

2.8.1. Sitra's Fair Data Economy Initiative and Digitrail Project

The #digipower investigation continues the work of Sitra's Fair Data Economy initiative[18], building on the previous "Digitrail" project[19] which surveyed companies and monitored six Finnish participants on a journey of exploring their personal data flows while using websites or apps. The GDPR was found to be inadequate to protect the rights of individuals; there is a lack of transparency, insufficient data protection regulation, and individuals cannot see how their data is

[16] https://www.politico.eu/newsletter/digital-bridge/russian-disinformation-me-myself-and-my-data-digital-tax-difficulties/

[17] A coalition of MEPs focused on bringing tracking ads to an end https://trackingfreeads.eu/

[18] https://www.sitra.fi/en/themes/fair-data-economy/

[19] https://www.sitra.fi/en/projects/digitrail/, https://www.sitra.fi/en/publications/on-the-trail-of-personal-data/

circulated. The Digitrail project uncovered a network of companies sharing data for advertising, which has a significant social impact (and which we explore further in 3.3).

#digipower takes some of the concepts of the digitrail project, to apply them on high-profile individuals, to see if these individuals might have greater success than lay people in seeing and understanding the use of their personal data, as well as to uncover the mechanisms of power at play in the data economy. While Digitrail was focused on advertising technology, the scope of #digipower is much bigger: we cover a variety of service providers like physical stores and mobile applications that are used in everyday life both within certain regions or which are accessible worldwide.

2.8.2. Multiplying Autonomous Efforts around Personal Data Rights

Compared to digitrail, the participants were much more encouraged to drive their own investigation. In this shift, the investigation follows a clear continuation of the work started around 2015 by Paul-Olivier Dehaye in developing his data advocacy through documentation of his own data trails[20]. Using his own data rights as they applied to his own data, Dehaye tried not only to understand where his own data was, but also to understand systematically how data rights could be used **as an advocacy method**[21].

In doing so Dehaye was attempting to expand upon the work of Max Schrems. Dehaye materialises more potential from data rights than just judicial action[22]. But while Dehaye's work has had direct input into the regulatory process[23] and led to some press reports over his own data[24], it suffers from two critical problems: it is fundamentally unsustainable (exhausting and unfundable), and fundamentally limited since it brings just one perspective on the data economy because it is an individualised approach.

Therefore, Dehaye has started focusing on multiplying[25] and comparing approaches so together multiple individuals start accumulating a set of proofs coming from their data about the data economy. Dehaye is now dedicated to assisting other actors of civil society to apply this methodology at a large scale, in different contexts where personal data is collected.

[20] Marres and Stark, *Put to the test: For a new sociology of testing*, The British Journal of Sociology, June 2020, https://doi.org/10.7916/d8-kkcr-7s54
[21] See for instance https://www.adexchanger.com/data-driven-thinking/personal-data-equal-law/
[22] *The EU guarantees its citizens' data rights, in theory*, The Economist, https://www.economist.com/europe/2018/04/05/the-eu-guarantees-its-citizens-data-rights-in-theory
[23] For instance, it led to a testimony in the UK Parliament that touched on the completeness of the *Download Your Information* tool. This testimony was directly referred to by Senator Blumenthal in highlighting to Mark Zuckerberg a contradiction in his US Congress testimony. See *Follow up questions to Mark Zuckerberg from the U.S. Senate Committee on Commerce, Science and Transportation*, p. 119/229.
[24] Aliya Ram and Madhumita Murgia, *Data brokers: regulators tackle the "privacy death stars"*, Financial Times https://www.ft.com/content/f1590694-fe68-11e8-aebf-99e208d3e521 .
[25] John Benedicto Krejsler, *Multitude, weaponize ye theories of globalization! Deleuzian strategies to affirm diversity vs predatory capitalism and nationalisms*, Discourse: Studies in the Cultural Politics of Education, Volume 42, 2021 - Issue 5, https://doi.org/10.1080/01596306.2020.1843117

Selected examples are:

- academic David Carroll requested his data from Cambridge Analytica[26], which led to some of the storylines for Netflix documentary *The Great Hack*.
- journalist Judith Duportail asked for her own Tinder data, which led to an extremely popular article in *The Guardian*[27], later expanded into the book *L'Amour sous Algorithme* and a documentary.
- journalist Carole Cadwalladr asked for her car insurance data from *Eldon Insurance*, eventually uncovering flows of data between the car insurer and the Leave.eu campaign[28].

Meanwhile, some methodological insights trickled in from academia, most notably via Jef Ausloos and collaborators[29,30,31]. Ausloos theorised access rights as a **research method**[32] (beyond the clear method of doing Subject Access Requests for the purpose of assessing compliance with them).

Another important influence to the design of this study was #digipower co-lead Alex Bowyer's 2020-2021 study into the human experience of GDPR[33], which took 10 individuals in the UK through a process of targeting 4-5 companies each with GDPR data access requests, and scrutinising privacy policies and data returns. This work assessed the effectiveness of the GDPR as perceived by individuals and sought to understand the experience of using Subject Access Requests from an individual's perspective. Bowyer's study had a particular focus on human-centric thinking, drawing on the Human-Data Interaction[34] and MyData[35] ideologies,

[26] Jackie Flynn Mogensen, *A Groundbreaking Case May Force Controversial Data Firm Cambridge Analytica to Reveal Trump Secrets*, Mother Jones, December 2017, https://www.motherjones.com/politics/2017/12/a-groundbreaking-case-may-force-controversial-data-firm-cambridge-analytica-to-reveal-trump-secrets/

[27] Judith Duportail, *I asked Tinder for my data. It sent me 800 pages of my deepest, darkest secrets,* The Guardian September 2017 https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold

[28] Carole Cadwalladr, *Arron Banks, the insurers and my strange data trail*, The Guardian, April 2018, https://www.theguardian.com/technology/2018/apr/21/arron-banks-insurance-personal-data-leave-eu

[29] Jef Ausloos & Pierre Dewitte. Shattering One-Way Mirrors. Data Subject Access Rights in Practice, International Data Privacy Law, 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3106632

[30] Veale, Binns & Ausloos, When Data Protection by Design and Data Subject Rights Clash, 2017, https://doi.org/10.2139/ssrn.3081069

[31] Jef Ausloos, *Paul-Olivier Dehaye and the Raiders of the Lost Data*, April 2018, https://www.law.kuleuven.be/citip/blog/paul-olivier-dehaye-and-the-raiders-of-the-lost-data/

[32] Jeff Ausloos, *GDPR Transparency as a Research Method*, 2019, https://doi.org/10.2139/ssrn.3465680

[33] Alex Bowyer, Jack Holt, *et al.*, *Human-GDPR Interaction: Practical Experiences of Accessing Personal Data*, 2022, https://doi.org/10.1145/3491102.3501947

[34] Mortier, Richard, et al. "Human-data interaction: The human face of the data-driven society." Available at SSRN 2508051 (2014). https://arxiv.org/pdf/1412.6159.pdf

[35] The MyData declaration, https://mydata.org/declaration/

which allowed it to go beyond the process and compliance aspects of SARs and draw conclusions around the *effects* of data-holders' power and the impacts of deficiencies of GDPR upon individuals. Understanding the areas identified by Bowyer where participants needed most support was very helpful in structuring our coaching and study design, so that we could go deeper than his study and focus on deconstructing the *causes and mechanisms* of data-holders' power from an individual perspective. This was highly complementary to the group investigative approach deployed in this investigation.

2.8.3. Collective Dimension and Exhaustive Qualitative Insights

A further dimension of the #digipower investigation is its social and collective dimension. This dimension has been particularly present in the work of Hadi Asghari, René Mahieu and collaborators[36,37,38]. Some of their practical and theoretical work informed this study, in particular in designing the right balance between a systematic research study with some standardisation (e.g, in the study design) across participants and a more open approach (e.g., personalised coaching with the participants in order to select *their* targets) that might lead to less quantified but more qualitative insights. These qualitative insights are powerful because they tackle and inform in-depth the concerns of individuals and their groups of belonging in the data economy.

We also build upon Jessica Pidoux's investigation[39] on the development and usage of dating apps and matching algorithms. We first adopted her analytical approach on computing systems from a sociotechnical perspective, which complements engineering and legal approaches described above and can be applied to the social and commercial practices of platforms with matching and recommendation systems (such as Uber, Airbnb, and Amazon). Secondly, we drew on her insights into the sociopolitical issues affecting individual autonomy, social cohesion and personal data sovereignty, as well as the idea that a community of practice (users of apps) can be understood through their own interests.

For background on the ecosystem of digital tracking and profiling by service providers, this investigation draws on the exhaustive work of Cracked Labs and Wolfie Christl[40], who

[36] Hadi Asghari, Thomas van Biemen and Martijn Warnier*, Amplifying Privacy: Scaling Up Transparency Research Through Delegated Access Requests* https://doi.org/10.48550/arXiv.2106.06844

[37] René L. P. Mahieu, Hadi Asghari and Michel van Eeten, *Collectively exercising the right of access: individual effort, societal effect*, Delft University of Technology, 13 July 2018, https://policyreview.info/articles/analysis/collectively-exercising-right-access-individual-effort-societal-effect

[38] René Mahieu and Jef Ausloos, *Harnessing the collective potential of GDPR access rights: towards an ecology of transparency*, https://dare.uva.nl/search?identifier=db9c6685-f3bb-4a3b-ae5d-4bb96501f79c

[39] Pidoux, Jessica. "Online Dating Quantification Practices: A Human-Machine Learning Process". EPFL, 2021. http://infoscience.epfl.ch/record/288400 .
Pidoux, Jessica. « Toi et moi, une distance calculée. Les pratiques de quantification algorithmiques sur Tinder ». In Carte d'identités. L'espace au singulier, édité par Yann Calbérac, Olivier Lazzarotti, Jacques Lévy, et Michel Lussault, Hermann., 249-67. Paris, 2019. https://infoscience.epfl.ch/record/283981?ln=fr

[40] Wolfie Christl, Digital Profiling in the Online Gambling Industry, January 2022, https://crackedlabs.org/en/gambling-data

investigated data flows in the online gambling industry and found evidence of highly sensitive personal data being exchanged between companies, with a lack of transparency and risk of vulnerable individuals being exploited through behavioural profiling. In particular, the exhaustivity of the coverage for that investigation is particularly masterful, even if centred on just one individual and one theme.

3. Case Studies

We now illustrate our findings through four case studies written from the combination of facts drawn from the individual experiences of the #digipower participants. The first three case studies are divided into two parts. They present a variety of contexts (more physical or more digital) where service providers such as Uber, Google, Meta (Facebook), Sanoma and Gigantti collect data about individuals. The fourth case study contains six parts and details the multiple efforts made by participants to exert their personal data rights with their chosen target service providers.

Together the case studies illustrate, in a pedagogical way, the processes within which personal data moves around in different contexts in the same manner as objects, humans and ideas forming daily life interactions move.

The case studies are:

1. *Who cares about my geolocation and why?*
2. *When you view the web, the web views you*
3. *Move fast and capture all signals, everywhere*
4. *Participants chasing their data*

Case study 1 (Section 3.1) explains the data that we (users) generate and which is collected about us in the physical context as we move through: a city, a road, a village, a building, a workspace. It shows how service providers attempt to understand and influence us to act differently by responding to our perceived intents (i.e., how users' interests are interpreted statistically). This influence is done through targeted offerings, and highlighting information about particular products, services and venues in order to persuade us to frequent specific stores or buy specific items.

Case study 2 (Section 3.2) examines the digital contexts we visit - websites, apps, or any other online space that, in comparison to a physical context service providers, can in this digital context observe to a greater degree, our actions, understand our interests, and attempt to influence us even more to e.g., click on specific content, download certain apps, spend time consuming content in a particular site or app, or subscribe.

Case study 3 (Section 3.3) looks in greater detail at *who* is collecting and influencing the user, and discusses Facebook and Google use of data-collecting 'sensors' into service providers' websites and apps, allowing them to monitor user's behaviour beyond the context where data is collected. We go into detail in particular on Facebook Custom Audiences in order to better illustrate the stakeholders involved in personal data flows and how those actors understand us and try to influence us.

Case study 4 (Section 3.4) differs from the other three case studies, as it does not focus on the data flows controlled by service providers. Instead, it focuses on how to gain transparency over those data flows. The case study concerns the gargantuan efforts made by the participants of the study to retrieve their data, the effectiveness of the GDPR and other lenses available. It provides further evidence, in comparison to previous work, of established issues that exist with data access and exerting data rights.

Our goal in presenting the first three case studies is not only to provide a series of examples. The case studies should, more importantly, enable the reader to reason about those examples, see the relationships between them within a case study, make associative leaps *between* case studies, and identify where those parallels break down. This is hard, as the examples will be pulled from many different situations in the data economy. For instance, how can we compare a Strava runner and the monetization of their data, with a Twitter user and the influence that results from their use of the platform?

Although the task is hard, only when an individual understands how their data is shared and used in ways beyond those originally consented, and how powerful data becomes when it is aggregated with other users' data, can they clearly see how the data economy affects one's everyday life. To aid the reader in being able to do this, we will first introduce a simpler situation, which we will use for reference throughout: the mobile game FarmVille. While at first sight this might seem irreverent to a serious investigation of data economy issues, the reader will see as our explanation progresses, how FarmVille can serve as a useful analogy for all the ways in which we act and are influenced in both physical and digital realms.

FARMVILLE: A TOTALITARIAN DATA-DRIVEN WORLD

Introduction to FarmVille

Let us introduce a context that only exists through data. **Context** is the surroundings of a user: e.g., a shop, a neighbourhood, a group of friends, an article or a social network app[41]. The context we want to present is FarmVille, a mobile game developed by the service provider Zynga. It is an entirely constructed game "world" where the player can expand their virtual farmstead, grow and nurture digital crops, trade produce and use tools, and generally spend time working on improving their digital farmstead. It is in many ways a simple and self-contained world, where only a finite number of activities are possible, and only a finite number of 'places' and objects exist. The FarmVille game and its spinoffs clearly provide a rewarding and fun experience to many people; at its peak in 2011, the game had an active player population of over

[41] For more details on this concept, see the narrative report (not required).

83 million[42]. While the game has declined from that level of popularity, Zynga continue to make revenue in the hundreds of millions from their game catalogue[43].

But there is more to this happy game than meets the eye. While players are free to go there any time, the world of FarmVille does not belong to the players; players are not free and it is not a democracy. It is a context created and designed by Zynga for the purpose of making money. The more time the player spends in FarmVille, the more emotionally invested they will become, and the more likely they will be to watch more ads that generate revenue for Zynga, and the more likely they will be to spend money on virtual upgrades to their farm. This is how the Free to Play (Freemium) business model[44] works. While the game is downloaded and installed for free, **everyone pays with attention or money**, either by spending time watching adverts (which is why certain in-game items have to be "earned" by logging time in game or waiting for real world time to elapse and visiting multiple times) or by paying in real money to subscribe and remove ads, or to buy in-game virtual upgrades and items. Whether it's money from advertisers or money from players, this is what Zynga cares about above all else.

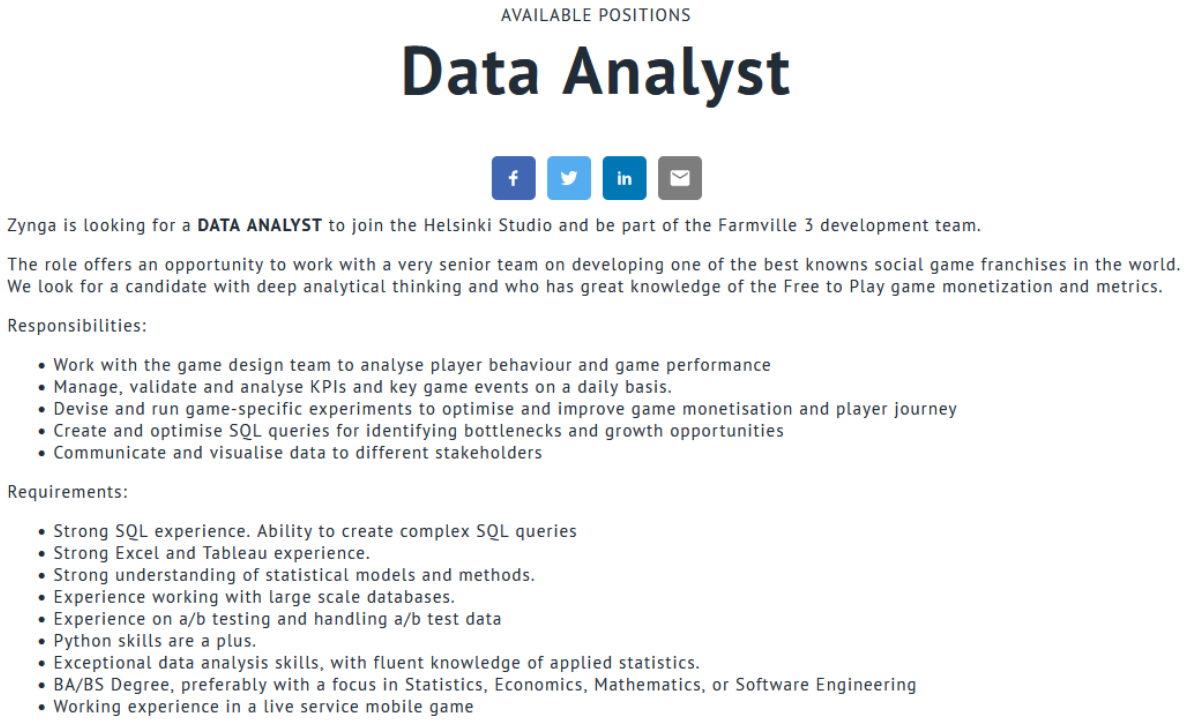
Constructing and Shaping FarmVille

Now let us take a moment to consider who constructs and shapes FarmVille. Figure 20 shows a recent job vacancy at Zynga's Helsinki office.

[42] https://venturebeat.com/2011/01/03/zyngas-cityville-becomes-the-biggest-ever-app-on-facebook/

[43] https://www.thedrum.com/news/2020/05/20/mobile-games-publisher-zynga-hopes-farmville-3-will-return-record-harvest

[44] Daniel Nations. What Are Freemium Games? Dec 2, 2020, https://www.lifewire.com/what-is-freemium-1994347



AVAILABLE POSITIONS

Data Analyst

Zynga is looking for a **DATA ANALYST** to join the Helsinki Studio and be part of the Farmville 3 development team.

The role offers an opportunity to work with a very senior team on developing one of the best knowns social game franchises in the world. We look for a candidate with deep analytical thinking and who has great knowledge of the Free to Play game monetization and metrics.

Responsibilities:

- Work with the game design team to analyse player behaviour and game performance
- Manage, validate and analyse KPIs and key game events on a daily basis.
- Devise and run game-specific experiments to optimise and improve game monetisation and player journey
- Create and optimise SQL queries for identifying bottlenecks and growth opportunities
- Communicate and visualise data to different stakeholders

Requirements:

- Strong SQL experience. Ability to create complex SQL queries
- Strong Excel and Tableau experience.
- Strong understanding of statistical models and methods.
- Experience working with large scale databases.
- Experience on a/b testing and handling a/b test data
- Python skills are a plus.
- Exceptional data analysis skills, with fluent knowledge of applied statistics.
- BA/BS Degree, preferably with a focus in Statistics, Economics, Mathematics, or Software Engineering
- Working experience in a live service mobile game

Figure 20: A job opening at the Helsinki offices of Zynga, the game company behind FarmVille.

Presenting this job vacancy is important because it shows that **data** is what drives FarmVille. Zynga needs to hold a staff of skilled data analysts and programmers so that they can watch the behaviour of players in the FarmVille game, identify changes they can make to increase revenue from those players, then make those changes and monitor their impact. In essence, Zynga runs the FarmVille game world as a totalitarian, Big Brother[45]-style state.

By watching everything, making changes to the world that players see and imposing rules that determine what players can and cannot do, Zynga uses their power to shape the context of players to function exactly as Zynga wants to: as a profit-generating machine. And whoever is hired to this vacancy, will become part of that totalitarian regime. That person will gain powers to use data through which to watch and understand players, and along with Zynga's developers - to use code to change the landscape (i.e., the user's viewpoint on the context) and rules of FarmVille society for influencing players to act differently within the FarmVille context. Collectively, Zynga's team acts (with regard to FarmVille-as-a-society) like a dictator in a totalitarian state. We use highly evocative terms like 'dictator' and totalitarian deliberately, not to cast judgement upon Zynga, but to emphasise the obvious differences between the functioning of the FarmVille society and that of the real world we inhabit.

[45] Novel: George Orwell, Nineteen Eighty-Four (1949)

Explaining the Mechanisms by which FarmVille Operates

Let us now explore the mechanisms by which FarmVille's pseudo-dictatorship operates, and how Zynga sees the game context. As we do so, we will introduce some of the terms used by the engineers who conceive those systems, and by the business executives who trade around such data, so that we can empower the reader and reclaim the jargon, turning it into something meaningful.

Every day, Zynga collects huge volumes of new data about its players. Its players' movements around the FarmVille context are watched, recorded and analysed - which seeds have been planted, which farm buildings are most popular, which virtual clothes are worn, which decorations have been placed. Players' engagements with quests will be carefully logged for later analysis.

All of these data points are known as **signals.** These signals, recorded in data, can be analysed to gain knowledge of a player's motivations, interests, and preferred choices. Statistics about their behaviour in the game context are generated from the billions of signals collected and used to inform Zynga's decision-making: Which quests did players quickly complete? Which ones did they find unrewarding and leave unfinished? Which ones motivated players to buy new upgrades to get them over the finish line?

This last point related to buying game upgrades is a key metric (or KPI, key performance indicator, as shown in the job ad) that is used to inform decisions. In sales terms, it is called **conversion**. A conversion is the successful transformation of an offer to a player (e.g. buy this pack of seeds or this upgrade to your farmhouse) into an actual monetary purchase. Every successful conversion is further analysed, using a technique called **attribution**, which allows mapping back from players' conversions to see exactly what offer or information presentation influenced their final decision to buy.

Using statistical analysis, Zynga staff work hard to **optimise** the game experience to maximise conversions. By analysing the data of players' game activity, they can figure out how to do this. For example, if play statistics identify that certain in-game features are leading to users spending less time on site, or buying fewer upgrades, these features can be modified or removed.

Using a well-established technique for researching user interaction features known as **A/B testing**[46] (again, mentioned in the job ad), **experiments** can be run on sections of the player population to determine which features will result in people spending the most time in-game or spending the most money. Both the experimentation and the resulting **data-informed decision making is unseen** by players, they only see the effects of the decisions - the disappearance of an

[46] What Is A/B Testing? A/B Testing and Split Testing Explained, Salesforce, https://www.salesforce.com/blog/what-is-a-b-testing/

interface button they used to use, a new discounted bundle of in-game currency appearing for sale, or the price of an item upgrade changing.

The goal of A/B testing is to optimise a player's journey, influence their choices, seed their desires and **manipulate their experience** of the FarmVille context to better serve Zynga's commercial objectives. In a digital context, it is easy for the context creator, like Zynga, to **change how reality is represented**.

Zynga can even treat specific portions of the player population differently, if Zynga identifies through psychological or **behavioural profiling** that certain groups of players are motivated differently, behave differently and therefore should be targeted in different ways, which means that not everyone experiences the same version of the game. While this treatment leads to personalised features appreciated by players, it can also provide powerful means for individual and massive manipulation.

The attempts to influence players' choices are done through targeted **nudges**[47]. These are appeals to human emotions, presenting what appears to be a fun and rewarding experience. But the reality is that in the context of FarmVille, profit comes first, and human welfare and **agency**[48] comes second.

A whole industry of Free-to-Play game developers have for over a decade been thinking of ways to manipulate human emotions to produce greater profits; threatening to remove a hard-earned reward, for example[49]. This is not new of course, psychology has informed marketing for many years[50]. But while traditional advertisers are limited to broadcast commercials, billboards and magazine ads, those involved in advertising in a digital context can go much further, manipulating aspects of individual experiences, sometimes in highly personalised ways.

While each FarmVille player is building a reality seemingly of their own design, for their own enjoyment, investing huge amounts of their time, mental attention and potentially money to see their virtual farm grow, Zynga (in collaboration with advertisers) is all the while building and growing a context of their design for those same players, configured to hold players' attention and entice them to reveal, through their actions, more about themselves.

[47] Richard H. Thaler & Cass R. Sunstein, Nudge: Improving Decisions about Health, Wealth, and Happiness, 2008

[48] Agency (n): the capacity to act or exert power (Merriam Webster online); Mortier *et al*, Human Data Interaction: The Human Face of the Data-Driven Society (2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2508051

[49] Ramin Shokrizade, The Top F2P Monetization Tricks, Gamedeveloper.com. June 26, 2013 https://www.gamedeveloper.com/business/the-top-f2p-monetization-tricks

[50] Understanding the Psychology of Advertising. Chicago School of Professional Psychology. Nov 3, 2020. https://www.thechicagoschool.edu/insight/psychology/understanding-the-psychology-of-advertising/

The games are structured to motivate players to keep playing, and arguably even to unhealthy levels, sometimes leading to addiction and physical or mental health impacts[51]. Ultimately, players' agency is removed or hindered in order to further the profits of the world's creator.

Considering again the role of the data analysts and engineers, and their ability to shape the entire world through structural changes and the imposition of potent rules, this position of data analyst at Zynga for FarmVille is analogous to being *The Mark Zuckerberg of FarmVille*. That data analyst has at his disposal a huge amount of signals to use as input to understand and watch people's behaviour, just like Big Brother in Orwell's *Nineteen Eighty-Four*, and just like Mark Zuckerberg at Meta (Facebook). That omniscience through data allows the analyst to instruct coders on how to structure players' interactions, just as Mark Zuckerberg does within Facebook. In Section 3.3, we will discuss how Mark Zuckerberg's Facebook is able to convince other actors to collect and share signals about individuals' actions *outside* of Facebook, extending his power beyond the social network platform itself. Mark Zuckerberg can be seen to have an enormous amount of power to shape our world (or at least the digital context within it) in ways we cannot always see. In the following sections, using evidence from our participants, we can now shed a light on some of these powers, not only those of Mark Zuckerberg, but of all the data-holding service providers who exert power over our lives to varying degrees.

[51] Daniel King, Paul Delfabbro, and Mark Griffiths, The Role of Structural Characteristics in Problem Video Game Playing: A Review, Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 2010, https://cyberpsychology.eu/article/view/4229/3272

3.1. WHO CARES ABOUT MY GEOLOCATION, AND WHY?

In the physical world, we take what we see as we go about our lives at face value. A tree is a tree, a spade is a spade. We have full agency to move around, to use our senses to explore the world autonomously. We have freedom of thought to make our decisions. Naively, we generally feel we are not influenced[52].

Yet the reality is that we are subject to influences in the physical world every day, through advertising such as billboards, radio/TV advertisements, mailshots and signs. And now as citizens of the Information Age[53] it has been true for some time that it is no longer possible to operate in the world without engaging with the digital side of that world[54]. Everyday, through our smartphones and computers, we access information about the world: articles, maps, stories - and we cannot always tell which of these were put there with the intent of manipulating our worldview. It is now time to ask, are there some ways in which the physical world is not so different from that of FarmVille?

In this first case study, we will illustrate, with data obtained by our participants through the three lenses (Section 2.4), the extent to which service providers collect data about our physical context, our movements and our intents to act in the physical context: intent to travel to a place or to make a physical purchase. We show, for example, how the capture of physical geolocation data can be statistically analysed by Google to draw inferences about individual activity, and how information from the digital realm (such as from Facebook) can be used to learn more about customers' purchasing behaviour in bricks-and-mortar stores.

First, in Section 3.1.1, we illustrate the ways in which data is captured about our physical context, and how that information is used to infer knowledge about us. We will show that, like in FarmVille, many of our everyday actions are watched, datafied and used to infer our interests and likely future predilections.

Then, in Section 3.1.2, we examine the ways in which those who have gathered data about our physical activity and context can extract highly detailed assertions about our activity, context and likely future behaviours, and subsequently exploit that knowledge to make (limited) changes in the physical world and (more freely) to its digital, informational counterpart. This is done in order to influence our thinking and behaviour, and to put these companies into a more advantageous position with regard to their ability to watch and learn even more from us in future.

[52] Jonah Berger, Invisible Influence: The Hidden Forces that Shape Behavior, 2016
[53] Julian Birkinshaw, Beyond the Information Age, Wired, 2014, https://www.wired.com/insights/2014/06/beyond-information-age/
[54] Adam Wagner, Is Internet access a human right? Guardian, 2012, https://www.theguardian.com/law/2012/jan/11/is-internet-access-a-human-right

Through this outline we will show that powerful data-holding organisations act like the FarmVille 'dictator' and are already influencing our choices over what to buy, where to go, and how to act in the physical world, by manipulating our worldview. Our choices and agency are already limited in ways we might not realise.

3.1.1. Location Data Reveals Interests and Intentions

What many users see as an annoyance (being constantly asked by apps for permission to know your location) in fact reveals that location data collection has become a normal part of business operations for many consumer-facing businesses. Myriad service providers like Deliveroo, Booking, Uber and Strava map user's movements. They collect location data and use it to acquire Big Brother-like knowledge about where their users are (as we will illustrate in the first subsection below). These companies will then use that knowledge to try and influence those users, which can have significant implications when populations' locations are tracked (as we will describe in the second subsection).

The most obvious form of geolocation tracking occurs in plain sight, when an individual consciously and deliberately uses technology to track their movements, using a smartphone, smartwatch or health tracker. For example, Figure 21, a visualisation constructed from data returned by Strava in response to Christian D'Cunha's SAR, shows a run he did in 2021, with red dots indicating the sample points taken by the device after he presses Start, up until the moment he stops logging. These records are kept primarily for the user's reference and are made available in Strava's user interface. But as a side effect, Strava gain the capability to know about their users' physical activities, and have the potential to analyse that data to extract further value from it[55].

[55] This Extracting behaviour is described further in the narrative report, Section 3.6, lever 1.

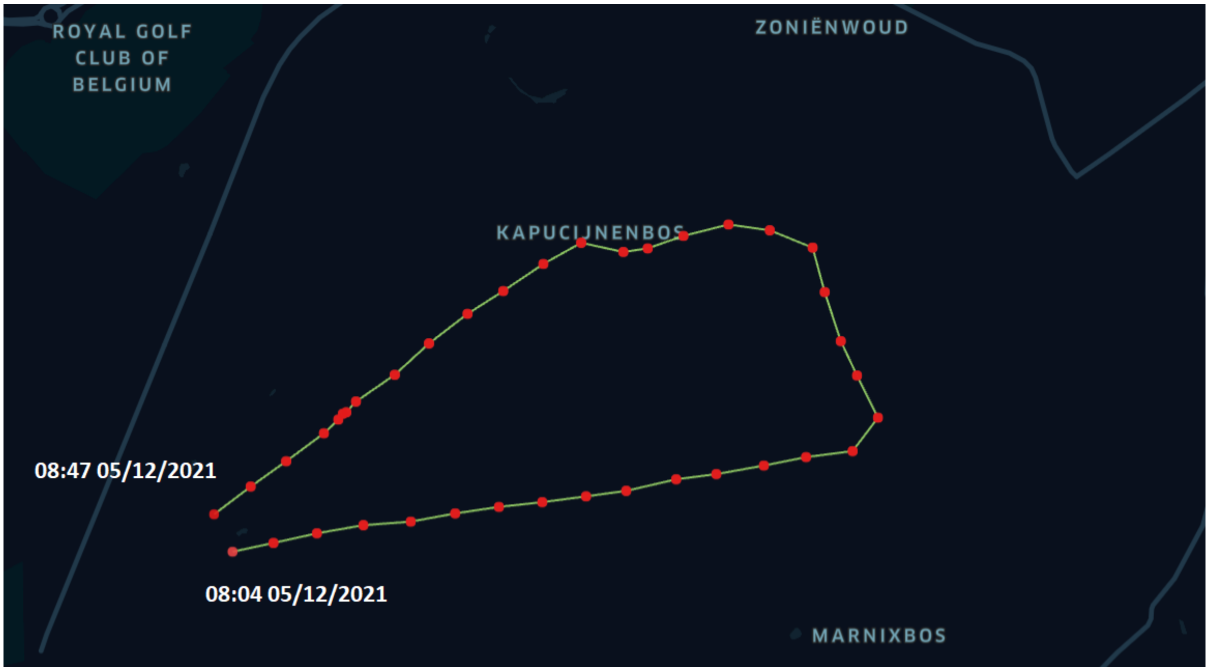
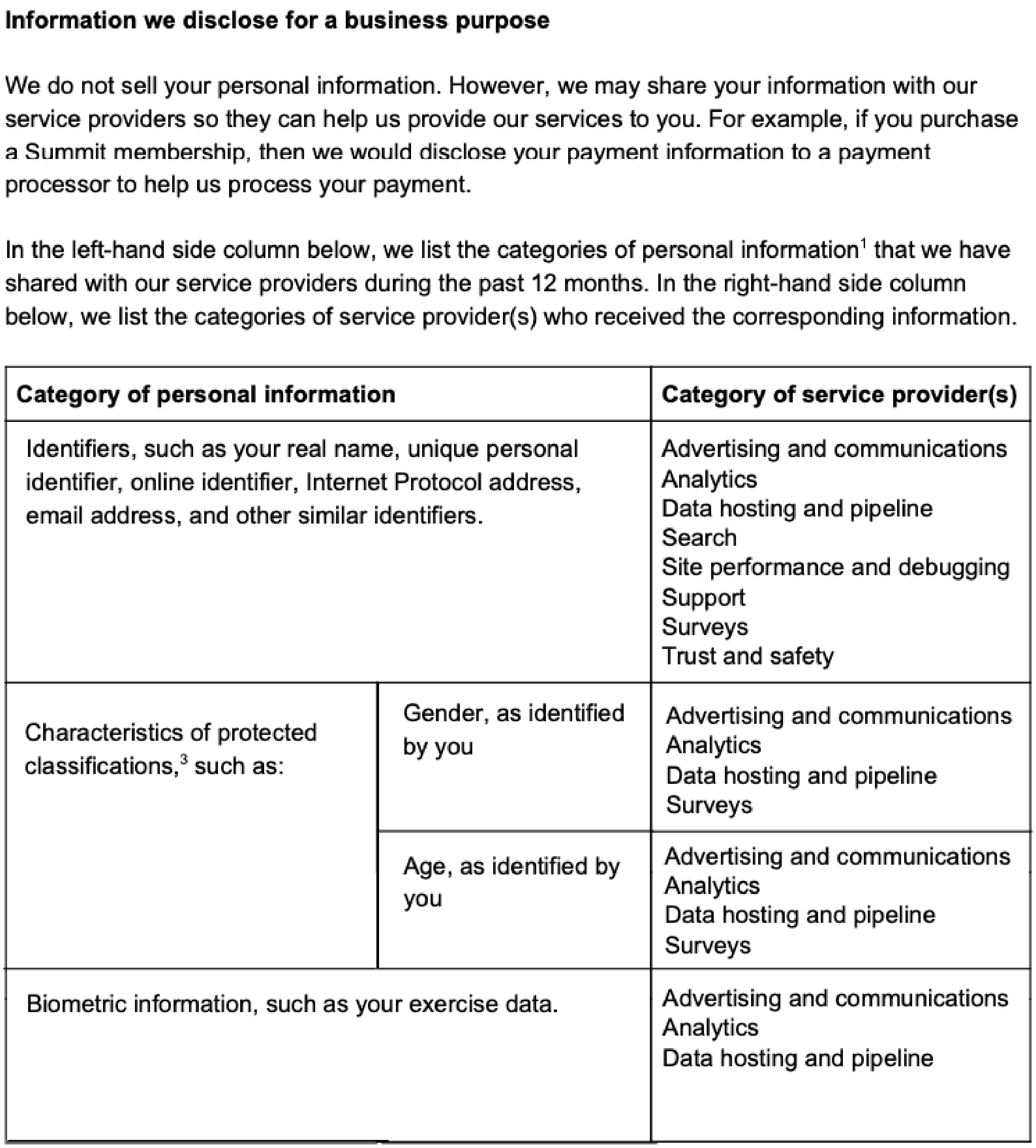


Figure 21: Christian D'Cunha goes for a run in Brussels, viewed through his Strava data.

This volunteered (see Section 2.6) location data can be used internally by the data holder to understand the user's locality and movements better, so that they might target any features or related products accordingly. Biometric data is not merely used to feed back to users information about their exercise performance and health condition, but when looked at over time can reveal much more about a person's habits and surroundings, which can be used for different purposes.

The information Strava provided in Christian D'Cunha's SAR return (see Figure 22 below) shows that Strava have found a new way to exploit this data about individuals for commercial benefit: they routinely share their users' biometric (exercise) data with third parties for advertising and communications purposes. In practice this means that when you log a run with Strava, you indirectly enable other actors to exploit your exercise data for their own ends. One can imagine that some companies, for example those advertising fitness products or health products would find this information very useful for targeting individuals, and insurance providers could potentially use such data to inform risk calculations, and research has already shown that health data can be very valuable in this way[56]. This could result in serious consequences for the individual concerned, from being bombarded with advertisements for fitness, health, or medical products they do not want or need, or facing increased insurance premiums.

[56] Libert, T. (2015). Privacy implications of health information seeking on the web. Communications of the ACM, 58(3), 68-77.



Information we disclose for a business purpose

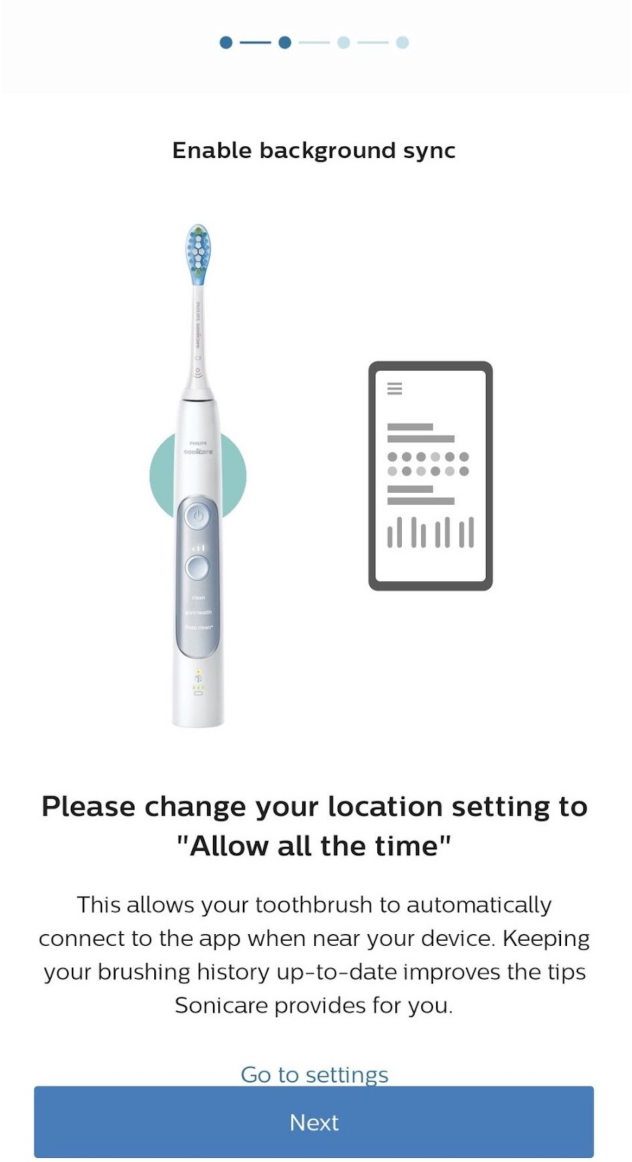
We do not sell your personal information. However, we may share your information with our service providers so they can help us provide our services to you. For example, if you purchase a Summit membership, then we would disclose your payment information to a payment processor to help us process your payment.

In the left-hand side column below, we list the categories of personal information[1] that we have shared with our service providers during the past 12 months. In the right-hand side column below, we list the categories of service provider(s) who received the corresponding information.

Category of personal information		**Category of service provider(s)**
Identifiers, such as your real name, unique personal identifier, online identifier, Internet Protocol address, email address, and other similar identifiers.		Advertising and communications Analytics Data hosting and pipeline Search Site performance and debugging Support Surveys Trust and safety
Characteristics of protected classifications,[3] such as:	Gender, as identified by you	Advertising and communications Analytics Data hosting and pipeline Surveys
	Age, as identified by you	Advertising and communications Analytics Data hosting and pipeline Surveys
Biometric information, such as your exercise data.		Advertising and communications Analytics Data hosting and pipeline

Figure 22: Extract from policy information returned by Strava, showing that they share identifying information, gender, age and biometric information such as exercise data with third parties for various purposes including advertising and analytics.

Sometimes geolocation data is collected by ongoing and continuous observation (see Section 2.6) rather than being reliant upon user action. Through various means, including Android operating systems and the use of Google Maps apps and websites, Google monitors the locations of many of its users 24/7 and stores this data on their servers. This is done by encouraging users to enable **background location tracking**, sometimes presenting a misleading impression that such sacrifice of location data is a routine or required part of normal operation, as in the case of some Philips Sonicare toothbrushes, as shown in Figure 23:



Enable background sync

Please change your location setting to "Allow all the time"

This allows your toothbrush to automatically connect to the app when near your device. Keeping your brushing history up-to-date improves the tips Sonicare provides for you.

Figure 23: A setup screen for a Philips Sonicare toothbrush encourages the user to enable background location tracking[57].

[57] For a discussion with Philips about this: https://twitter.com/PhilipsSonicare/status/1466427422514069515

Once background location data tracking is operational, the user's device sends geolocated coordinates to the provider at hundreds of intervals throughout the day. For example, we saw participant Mark Scott had previously been background location tracked by Google. In Figure 24, we show a reconstruction of a small excerpt of Mark Scott's movements in Berlin, where his Google Takeout-downloaded Location History data shows the precise route he took and time taken while walking between two buildings. In the case of Google, such granular location data is recorded all the time, except where users have explicitly disallowed Location tracking on their device or in Google settings, or when the user's devices has location tracking off by default, as is the case on some non-Google Android phones. In some cases, location tracking by Google was found to take place even after users had opted out[58]. The parallels to Big Brother and the FarmVille 'dictator' are clear - your location is watched, whether you like it or not.

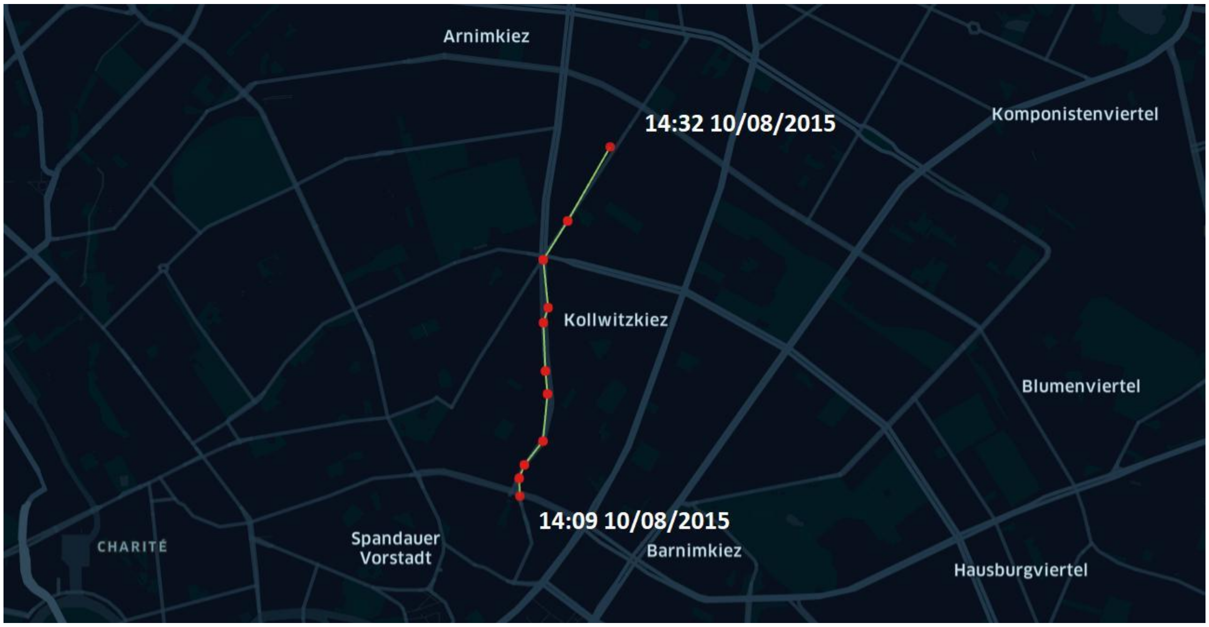


Figure 24: Mark Scott walks through Berlin; his movements and his activity are captured in Google Location History.

This **individual-level surveillance data** represents a rich source of information that Google can use to adapt content and advertisements based on user location. Typically, for example, the search results returned by Google's search engine will be tailored to the city or country where the user's location data shows they have recently been located.

More significantly this capability also gives **population-level surveillance capabilities** to Google, which they are then free to exploit. For example, by in effect turning every smartphone into a **surveillance device** (like *Nineteen Eighty-Four*'s omnipresent and always surveilling telescreens), Google now has the ability to know the movement speed and trajectory of millions of individuals in real time. This allows them to measure (and detect variations in) traffic flow, which they use

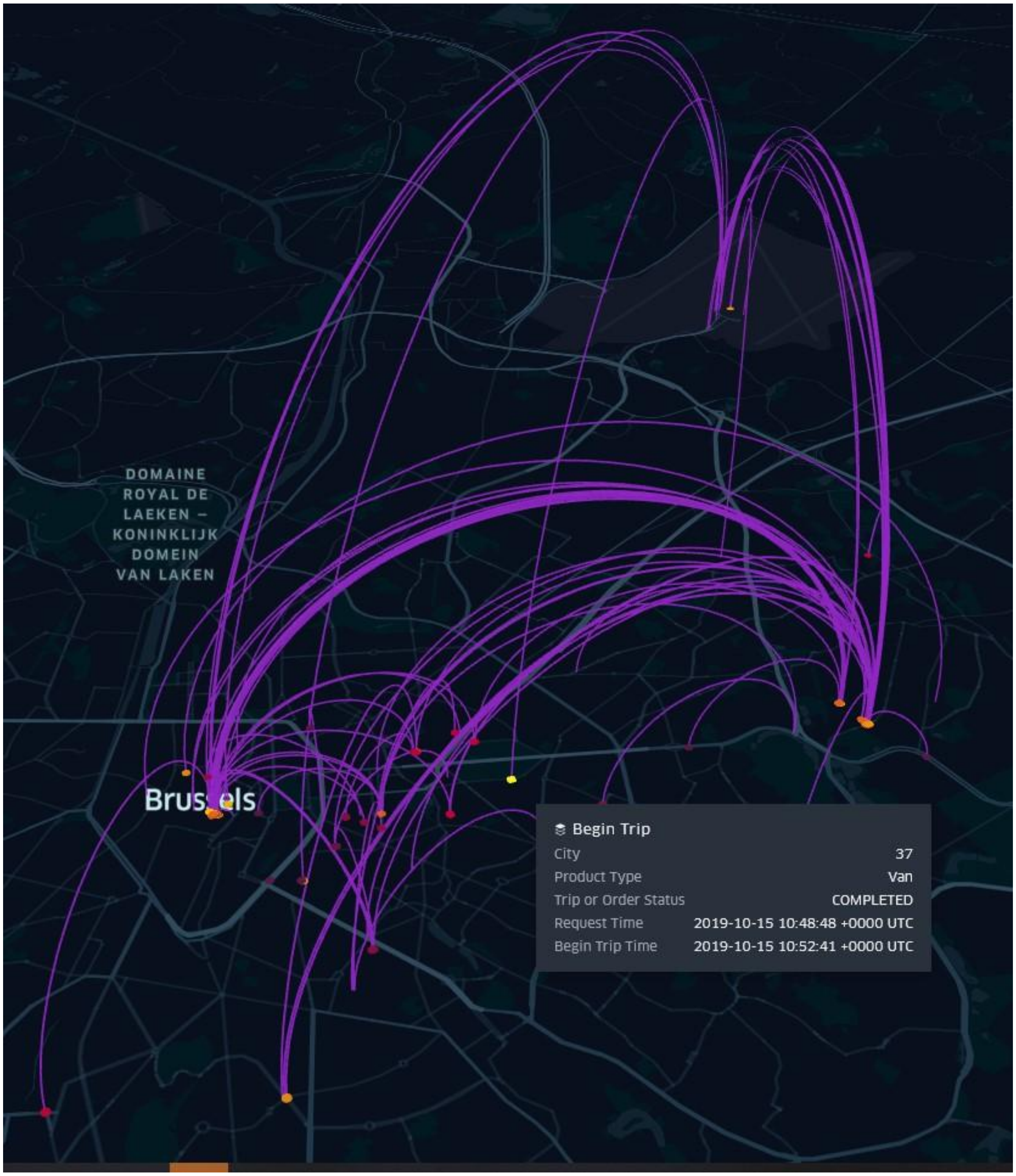
[58] Google records your location even when you tell it not to, The Guardian, 13 Aug 2018, https://www.theguardian.com/technology/2018/aug/13/google-location-tracking-android-iphone-mobile

to indicate traffic on Google Maps with a red, yellow or green line according to traffic density; their navigation features can then re-route users around roadblocks or congestion. This exploitation of the location data provides a benefit to users, but it is not difficult to imagine other ways in which that data could be exploited by Google. There have already been calls for greater regulation of location data[59], given the potential risks it exposes to individual privacy.

Data retrieved by our participants allows us to observe and infer more deeply about what companies do with all this location data, and how they take advantage of it. Indeed, companies do invest significant technical effort in analysing collected geolocation data, in order to gain a better understanding of the user's context.

For example, an individual's movements over time can be even more valuable to companies than just knowing their instantaneous location. Patterns of activity over weeks, months and years reveal the locations with which an individual is most strongly associated; for example their home, workplace or frequently used travel hubs. In the Uber trips data downloaded by Miapetra Kumpula-Natri, it is easy to identify her workplace, her former residence, and her commonly used drop-off/pick-up points at the airport and in the city centre, as our visualisation in Figure 25 below shows. Uber can use this to change the pricing and routing decisions they make for her and customers like her in future.

[59] How to Stop the Abuse of Location Data, The New York Times, 16 Oct 2019. https://www.nytimes.com/2019/10/16/opinion/foursquare-privacy-internet.html



Begin Trip

City	37
Product Type	Van
Trip or Order Status	COMPLETED
Request Time	2019-10-15 10:48:48 +0000 UTC
Begin Trip Time	2019-10-15 10:52:41 +0000 UTC

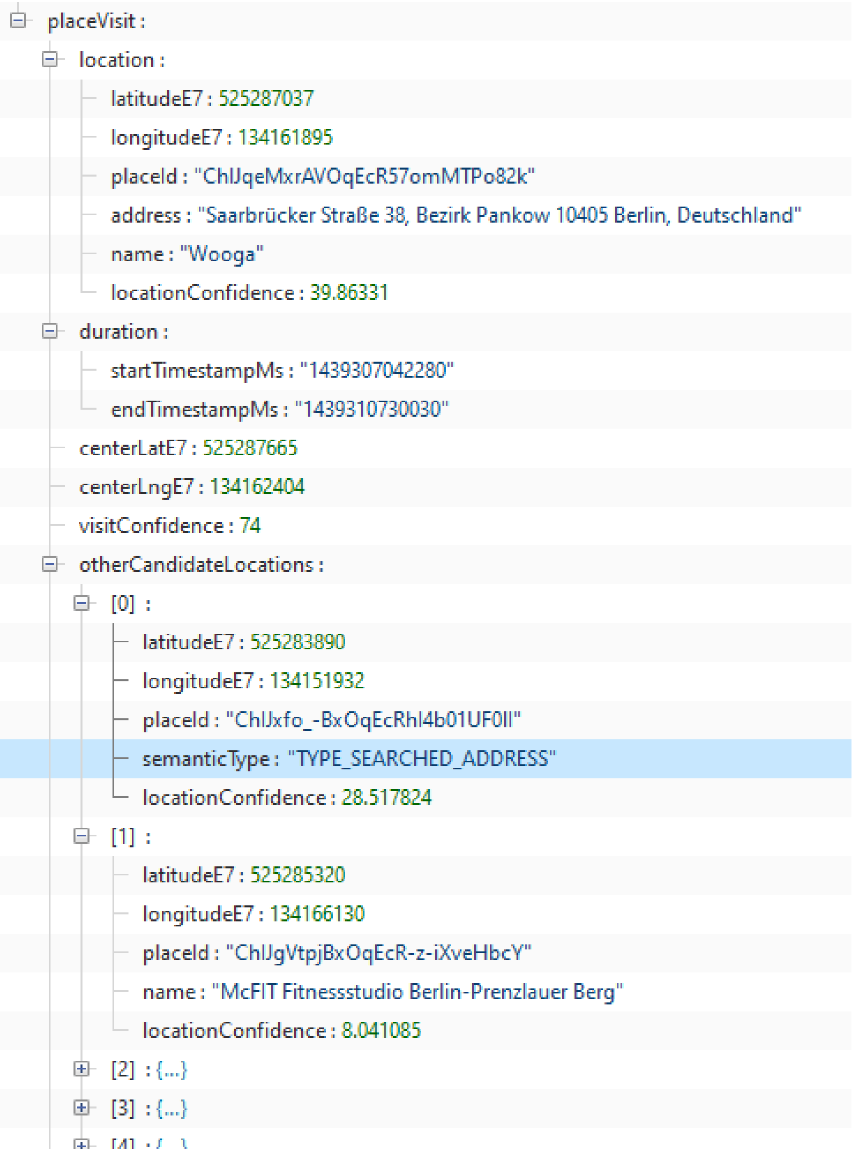
Figure 25: The start and endpoints of Miapetra Kumpula-Natri's Uber journeys in Brussels provide a clear overview of her (taxi-based) travelling habits across multiple visits, between her work area, her old living place, the airport, and the centre of town.

Patterns in location data can be extremely identifying, and make it much easier to make the leap from raw data to knowledge about the individual and how they live their life. Researchers found that just 4 location points were enough to uniquely identify an individual[60]. The existence of multiple location data points about an individual therefore presents further risks, that different actors could combine data points, even when anonymised, and map this data back to identifiable individuals who can then be targeted with advertising. For example, regular visits to a hospital, addiction clinic or psychiatrist might reveal facts about an individual that they would not want their employer or an insurer to know, and even small signals can convey information: for example, a location data point getting stronger for five minutes every two hours might reveal that person is stepping outside to smoke a cigarette. The capability to extract knowledge from location data patterns has significant impacts for privacy, as in the case of the gay Catholic priest who was outed as a result of analysis of geolocation data gathered by dating app Grindr[61].

Google takes the analysis of users' movement data even further. In Takeout-returned data, there is a section called "Semantic Location History", which shows that Google use statistical techniques to infer, for each time the user is stationary, the most likely actual location (venue, building, home, store, restaurant etc.) that a user visited. From a set of known geolocated locations, Google assigns a statistical probability that it was the correct location for this user, based on what they know about that user, and select the most likely to infer that that is where he was. A sample of this data is shown in Figure 26 below. We can see that Google does not only know that Mark Scott was at coordinates 52.5287037 N, 13.4161895 E at a precise point in time, but that there was a 74% chance he visited a particular venue in that area, and a 39.8% chance that this place was the offices of game developer Wooga, at Saarbrücker Straße 38, Berlin. We also see two other candidate locations for that visit: a previously searched address with 28.52% confidence, identified here only by an internal identifier, and a McFIT gym with 8.04%.

[60] Yves-Alexandre de Montjoye *et al.*, Unique in the Crowd: The privacy bounds of human mobility, Nature, 2013, https://www.nature.com/articles/srep01376

[61] Molly Olmstead, A Prominent Priest Was Outed for Using Grindr. Experts Say It's a Warning Sign, Slate, July 2021, https://slate.com/technology/2021/07/catholic-priest-grindr-data-privacy.html



```
placeVisit :
    location :
        latitudeE7 : 525287037
        longitudeE7 : 134161895
        placeId : "ChIJqeMxrAVOqEcR57omMTPo82k"
        address : "Saarbrücker Straße 38, Bezirk Pankow 10405 Berlin, Deutschland"
        name : "Wooga"
        locationConfidence : 39.86331
    duration :
        startTimestampMs : "1439307042280"
        endTimestampMs : "1439310730030"
    centerLatE7 : 525287665
    centerLngE7 : 134162404
    visitConfidence : 74
    otherCandidateLocations :
        [0] :
            latitudeE7 : 525283890
            longitudeE7 : 134151932
            placeId : "ChIJxfo_-BxOqEcRhl4b01UF0lI"
            semanticType : "TYPE_SEARCHED_ADDRESS"
            locationConfidence : 28.517824
        [1] :
            latitudeE7 : 525285320
            longitudeE7 : 134166130
            placeId : "ChIJgVtpjBxOqEcR-z-iXveHbcY"
            name : "McFIT Fitnessstudio Berlin-Prenzlauer Berg"
            locationConfidence : 8.041085
        [2] : {...}
        [3] : {...}
        [4] : {...}
```

Figure 26: Google Takeout-downloaded Semantic Location History data for Mark Scott.

Google can combine signals from its many data sources to improve their statistical modelling: In Figure 27 below, we have plotted the candidate locations for this single geolocation data point onto a map to illustrate this. The central node indicates the most likely location (the offices of game developer Wooga), and the thickness of the lines indicate the likelihood for the secondary guesses. The green circle indicates a previously searched address, which is allocated by Google a greater likelihood of being correct (28.52%) despite its greater distance, presumably precisely because it had been searched by Mark Scott before.

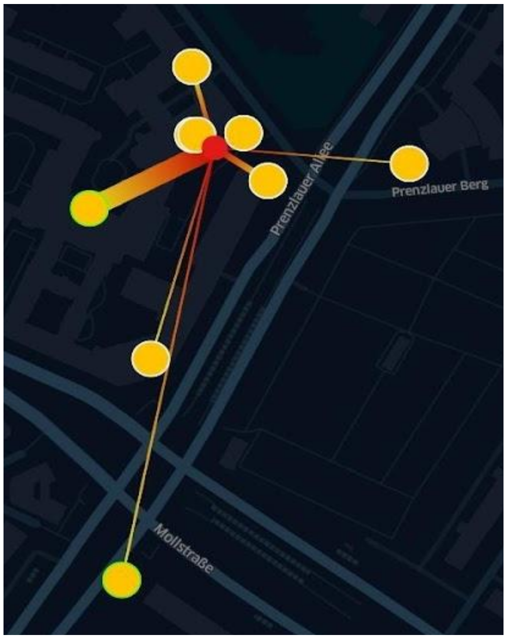


Figure 27: Geolocated illustration of the information from the previous figure (Mark Scott in Berlin),

Google also performs similar analysis while users are moving, to determine their most likely mode of transport, for example walking, driving, taking a bus, train, taxi or tram. Analysis of returned Semantic Location History files reveal that Google is now able to classify, from background location data, the user's most likely movement activity at any given moment, in nearly 40 different ways, as Figure 28 indicates:

BOATING
CATCHING_POKEMON
CYCLING
FLYING
HIKING
HORSEBACK_RIDING
IN_BUS
IN_CABLECAR
IN_FERRY
IN_FUNICULAR
IN_GONDOLA_LIFT
IN_PASSENGER_VEHICLE
IN_SUBWAY
IN_TAXI
IN_TRAIN
IN_TRAM
IN_VEHICLE
IN_WHEELCHAIR
KAYAKING
KITESURFING
MOTORCYCLING
ROWING
RUNNING
SAILING
SKATEBOARDING
SKATING
SKIING
SLEDDING
SNOWBOARDING
SNOWMOBILE
SNOWSHOEING
STILL
SURFING
SWIMMING
WALKING
WALKING_NORDIC

Figure 28: The different modes of transport or movement to which Google is able to assign a statistical significance in derived Semantic Location History data.

Taking these two semantic inference capabilities (the capability to infer **where people spend time** and **how and where they travel**) together, Google can build up an entire picture of how users spend each moment of their daily lives, in the process uncovering some highly marketable facts about those individuals. They infer detailed knowledge about users' behaviour, translating raw location data into exploitable facts that tie the users to specific addresses, businesses, services, leisure pastimes or routes. Google performs this knowledge extraction from location data in real time, which can be verified by visiting https://timeline.google.com. Only the largest tech platforms have the capability to perform such detailed capture and analysis of people's locations. As a consequence, those companies with the greatest data breadth and volume and the most advanced processing infrastructure have the most knowledge of users' context (which can be exploited directly by that company, or sold to advertisers as knowledge about users).

From a service provider or advertiser's perspective, knowing a user's intent as early and as granularly as possible represents tangible commercial opportunity - the opportunity to convert that intent into a purchase. Geographically localised actions are more informative. For example, when Miapetra Kumpula-Natri opened the Uber app when away from her home in Brussels, as shown in Figure 29 below, this produces a signal of intent that provided Uber (and also her handset provider such as Google, Apple or Samsung, as well any other intermediaries able to access that datapoint) with a clear indication that she wanted a taxi, which is commercially exploitable information.

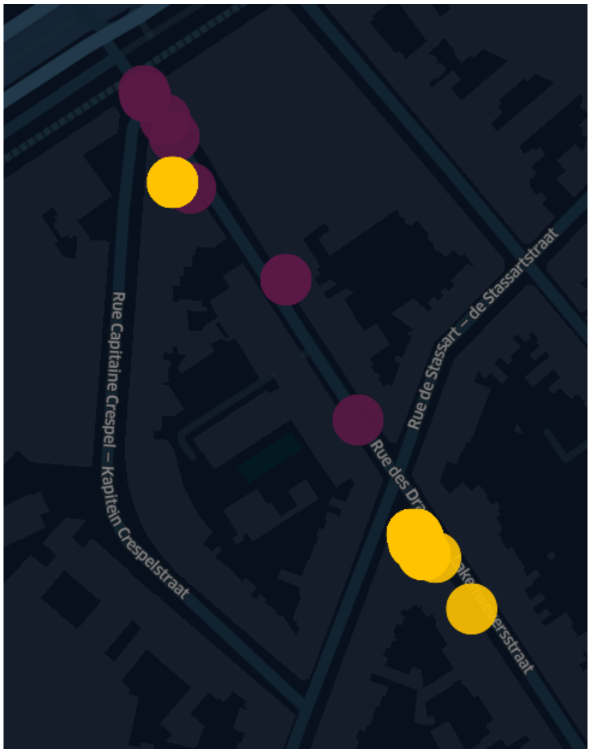

Rue Capitaine Crespel – Kapitein Crespelstraat
Rue de Stassart – de Stassartstraat

Figure 29: Miapetra Kumpula-Natri's app opens on a given street in Brussels. The colour reflects horizontal accuracy, with yellow a precision of ~50m and purple a precision of ~5m.

The dots further South also occurred five minutes earlier than those further North. We deduce that she might have opened the app in the South, quite possibly inside, and walked to the vehicle once it arrived.

There are many ways in which our actions in the physical world can produce data which signals our intentions, even without picking up our phones. For example, each time Jyrki Katainen fills up his car at a Neste K service station, Kesko Group collects data on that purchase, which indicates that he is travelling - the same for S Group when Sari Tanus refuels at an ABC station.

These retail groups can collect a huge number of commercially valuable signals about individuals who shop with them regularly. Purchased food and drinks and household items might reveal intent to host a party or go on a holiday - context that advertisers would like to know. Home improvement purchases such as those we saw in Jyrki Katainen's Kesko data from their K-Rauta stores, could reveal intent to refurbish a bathroom or decorate a living room, which is actionable insight that an advertiser could use.

This **opportunity to gather signals** is of course is the premise behind loyalty schemes such as Kesko's Plussa scheme - valuable data is collected, in exchange for some customer benefits and discounts. The value of such data is well established and data has been used in this way for over a decade; famously US retailer Target deduced a young woman's pregnancy before anyone else in her family knew, purely from the data they had collected about her purchases[62].

There is evidence that providers continue to look for new forms of signals. For example, the format of Google's Semantic Location History (described above) shows fields designed to store details of when a user charges their electric car. These fields do not yet appear to be in use, but suggest that companies wanting to understand their users better are continuing to look for new signals that can tell them something about that person's intentions.

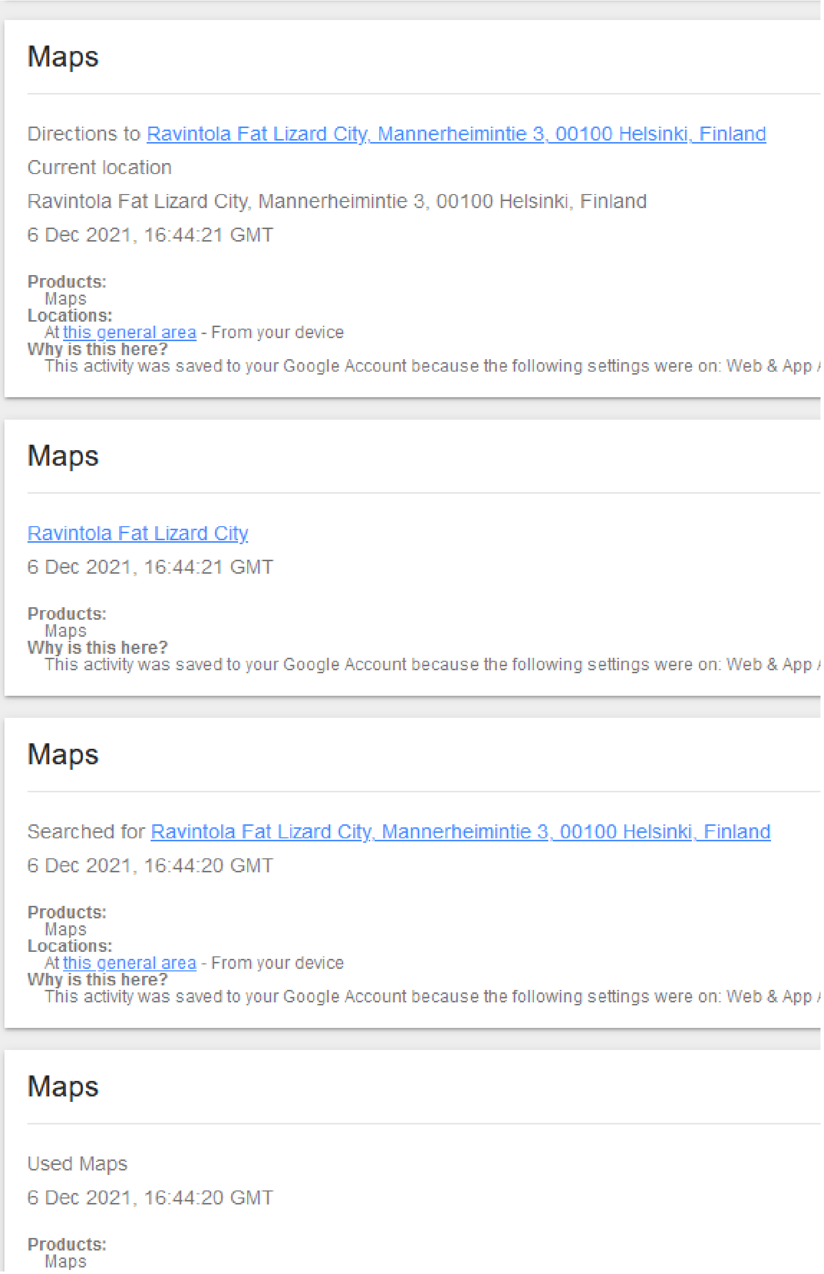
The recent emergence of smart homes and offices, urban sensing projects, and Internet-of-Things devices can create a whole new range of signals from the physical world that could be used by data-collecting companies to determine user intent or context. For example, smart lightbulb and smart TV data could reveal sleep patterns, work/leisure habits, and more, which many organisations would value.

Clearly all commercial actors (not just retailers and service providers, but infrastructure companies and advertisers too) are driven to collect as much data as possible, in order to collect as many signals of intent as possible. An example of this caught public attention this month, when it was uncovered that Google have been using the Dialer and Messages app on many Android phones to collect details about when people are making calls and sending and receiving messages[63].

[62] Kashmir Hill, How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did , Forbes, 2012. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=745beec56668

[63] Thomas Claburn, Android's Messages, Dialer apps quietly sent text, call info to Google, The Register, 21 Mar 2022, https://www.theregister.com/2022/03/21/google_messages_gdpr/

While (thanks to the sensors in our smartphones and fitness trackers) our physical actions in the world can reveal some signals, as soon as we pick up our phones or laptops and look for information about the physical world, we reveal even more. Our searches can produce hundreds of intent signals every day. In the Google Takeout-downloaded My Activity data obtained by Miapetra Kumpula-Natri (see Figure 30 below) we can see the precise times at which she opened the Maps app, searched for a specific restaurant, The Fat Lizard in Helsinki, viewed a map of that restaurant's location, and then searched directions of how to travel there. Each of these steps is a signal that indicates interest or intent… interest in food/restaurants, interest in a particular type of cuisine, intent to eat out, intent to travel to travel to Helsinki's Forum/Mannerheimintie area, intent to travel on foot, and intent to visit a specific restaurant while there.



Maps

Directions to Ravintola Fat Lizard City, Mannerheimintie 3, 00100 Helsinki, Finland

Current location

Ravintola Fat Lizard City, Mannerheimintie 3, 00100 Helsinki, Finland

6 Dec 2021, 16:44:21 GMT

Products:
Maps
Locations:
At this general area - From your device
Why is this here?
This activity was saved to your Google Account because the following settings were on: Web & App /

Maps

Ravintola Fat Lizard City

6 Dec 2021, 16:44:21 GMT

Products:
Maps
Why is this here?
This activity was saved to your Google Account because the following settings were on: Web & App /

Maps

Searched for Ravintola Fat Lizard City, Mannerheimintie 3, 00100 Helsinki, Finland

6 Dec 2021, 16:44:20 GMT

Products:
Maps
Locations:
At this general area - From your device
Why is this here?
This activity was saved to your Google Account because the following settings were on: Web & App /

Maps

Used Maps

6 Dec 2021, 16:44:20 GMT

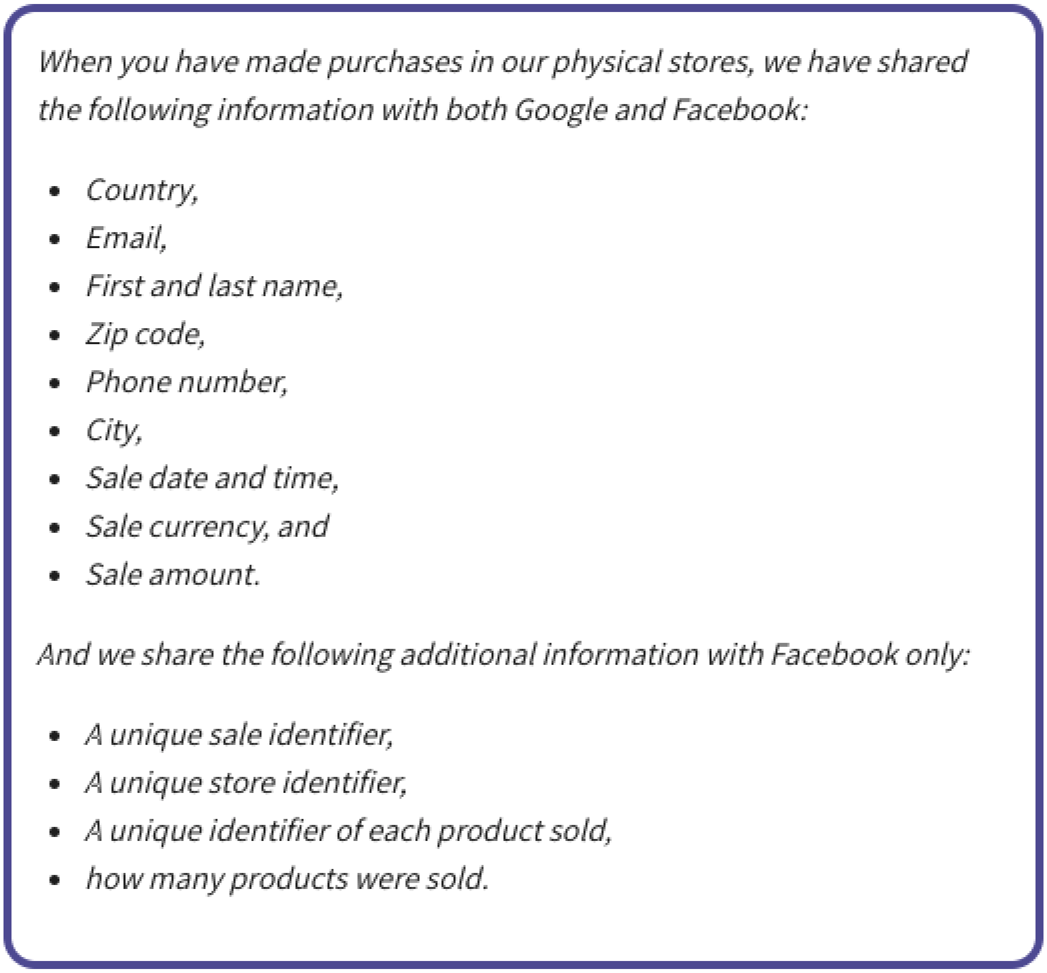
Products:
Maps

Figure 30: Miapetra Kumpula-Natri searches for a restaurant on Google Maps, then gets directions to travel there, as seen in her Google My Activity data. In between, Google confirms she was indeed looking for that restaurant because she (presumably) clicked on its business card within Google Maps.

What we can see from this sort of example is that as we move through the physical world and make use of apps on our phones, our every action reveals an intention; these actions are surveilled and can be used to augment the understanding that commercial organisations have about us. While data sharing and exchange practices were often not transparent in companies'

GDPR responses, we can nonetheless infer that highly granular data about individuals' behaviour is being offered to advertisers as a means to target particular types of individuals more precisely.

More advanced information about individuals' context, can enhance interpretation of the signals of intent. We saw evidence of this in the case of Nordic retail giant Gigantti (owned by Currys plc in the UK) - which has both online and physical bricks-and-mortar stores. The company revealed in their SAR responses to Miapetra Kumpula-Natri and Sari Tanus (see Figure 31 below) that details of Gigantti purchases *in physical stores* are routinely passed to Google and Facebook (additional granular product-level information is passed to Facebook), which was previously unknown and concerning to both participants. **NB.** We do not believe this practice is unique to Gigantti; it would appear to be a systemic practice of retailers, that is visible in Gigantti's case only due to their exceptional candour and transparency in their SAR response (see Section 3.4.4).



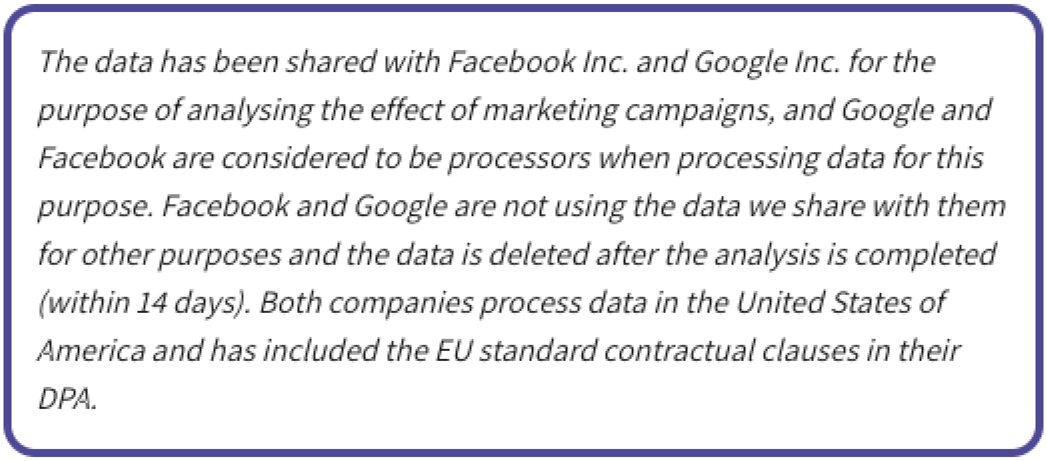
When you have made purchases in our physical stores, we have shared the following information with both Google and Facebook:

- *Country,*
- *Email,*
- *First and last name,*
- *Zip code,*
- *Phone number,*
- *City,*
- *Sale date and time,*
- *Sale currency, and*
- *Sale amount.*

And we share the following additional information with Facebook only:

- *A unique sale identifier,*
- *A unique store identifier,*
- *A unique identifier of each product sold,*
- *how many products were sold.*

Figure 31: An extract from Gigantti's SAR response to Miapetra Kumpula-Natri and Sari Tanus, formatted and reordered for easier readability.



The data has been shared with Facebook Inc. and Google Inc. for the purpose of analysing the effect of marketing campaigns, and Google and Facebook are considered to be processors when processing data for this purpose. Facebook and Google are not using the data we share with them for other purposes and the data is deleted after the analysis is completed (within 14 days). Both companies process data in the United States of America and has included the EU standard contractual clauses in their DPA.

Figure 32: An extract from Gigantti's SAR response to Miapetra Kumpula-Natri and Sari Tanus, where their transfers of data to Facebook and Google are explained.

According to Gigantti (see Figure 32 above), this was done to connect user interactions with online marketing campaigns to judge the effectiveness, and Facebook and Google are not supposed to use the data for other purposes. But the wording does not preclude the possibility that the data is used for other Facebook/Google customers in assessing their marketing campaigns. Furthermore, if Facebook and Google are defined as data processors, this would mean their role here is subservient to the retailer; this is at odds with what is known about Facebook & Google's business models and attempts to amass data from multiple sources about individuals, and thus scepticism and further scrutiny is advised (see Section 3.3 below).

This practice of sharing physical purchases (conversions) into the digital space is significant, and shows the lengths to which businesses are going to build and augment data about individual physical activity in the digiscape. If your routines and habits are known to advertisers, these can be exploited against you. As mentioned at the start of this section, people consider that they are not influenced, but we can see that with such granular inferences such as your hobbies and lifestyle being revealed through your location data, it is only a matter of time until an advertiser succeeds in converting you through a very specifically tailored advertisement that they were only able to offer you because of the location-data-based-inferences that had been made about you. The existence of granular data signals about you carries personal risk too, as now data that previously didn't exist could that could be exploited against you[64]. Providers try to mask these risks that might deter location tracking permission, by providing benefits to individuals, such as Google Timeline, or certain functionality that can only be accessed when location data tracking is enabled. Nonetheless, we can see that in the digitally-connected physical world, your activity can be monitored by those that wish to understand you better. Just like in FarmVille, **everything you do is logged and analysed** so that companies might figure out

[64] Jacob Leon Kröger, Milagros Miceli, and Florian Müller, How Data Can Be Used Against People: A Classification of Personal Data Misuses, December 2021, https://doi.org/10.2139/ssrn.3887097

how to influence you in ways that will generate more profit for them or put them in a stronger position in future.

3.1.2 Knowledge of Population Activity is Exploited to Influence Us

Having established the extent to which companies are gathering signals of intent and information about individual context, we turn our attention in this subsection to a deeper look at how companies can use their acquired knowledge about individuals and groups of individuals to exert influence in the physical world. As we saw above, our usage of digital tools to access information about the physical world can be highly revealing, as in the case of Miapetra Kumpula-Natri searching for directions on Google Maps. Google can then use this acquired knowledge to target advertisements more precisely to those individuals most likely to be interested - and can sell our attention to advertisers for a greater amount when they are able to target adverts more precisely. However such individual-level information is really only the tip of the iceberg, because knowing about the intentions and activity contexts of large groups of people unlocks even greater capabilities. Such acquisition of data and knowledge quickly grants data holders powers over sections of the population, and, as we will show below, the informational landscape available to people in the physical world can be manipulated, using such knowledge, to influence people.

Data analysis can discover new signals of intent about groups as well as individuals. Similar to the pattern described in 3.1.1 above where organisations analysed individuals' data to determine their context, we can expect that data holding organisations will analyse the data they have collected to see if they can make **predictions about groups of people**.

For instance, in 2016, the CEO of Foursquare bragged on his blog[65] of being able to predict the stock market misfortune of Chipotle, a US fast food chain ahead of quarterly reports. This prediction was based on so-called **footfall** (foot traffic into the stores), a proxy for sales. He further described the data collection to be explicit check-ins by users but *also* implicit (i.e. non-voluntary) ones, in a pattern similar to Google's predictions of Mark Scott's activity through background geolocation collection.

Interestingly, the Foursquare CEO was beaten in that game - and by quite a wide margin. Indeed years earlier two fraud analysts at Capital One had bypassed the blog posts stage to directly exploit their privileged access to the master database of Capital One to discover stock market tips. They started to aggregate sales data (Capital One has a sizable portion of all credit card transactions) for multiple listed companies, predicting swings in sales just before the quarterly earnings announcements. For instance in July 2014, they bought 5,500 Chipotle options (coincidentally) for less than $100,000 and made $278,000 in three days. According to the

[65] Matt Turner, *How Foursquare Accurately Predicted That Chipotle's Sales Would Plummet,* Slate https://slate.com/business/2016/04/foursquare-data-accurately-predicts-chipotle-s-sales-dip.html

Securities and Exchange Commission, between 2012 and 2015 they made almost $3 million on a $147,300 initial investment (three year return: 1819%)[66]. Both individuals were fired and faced criminal charges and multi-million-dollar fines, but, illegality aside, this illustrates the knowledge and power available from population-level activity data.

In recognition of the power to draw insights from data about people, some organisations have taken steps to establish business relationships that can **configure data use for mutual advantage**. Realising that more data means more signals, and more signals means more knowledge, and more knowledge means more power to influence, we see an emerging practice known as "**compute together**"[67], where companies, lacking consent to exchange personal data, instead 'compare notes' by performing parallel calculations on the data they do have. The exchange of data between retailers like Gigantti and social networks like Facebook and Google can be seen as a manifestation of this kind of attitude to data. Even where companies do not have consent to share specific data points with each other as in Gigantti/Facebook/Google's case, two companies can both run algorithms on the customer data they do have, match customers, visits and transactions up, and draw new inferences.

A more involved example of 'compute together' was seen in the case of Mastercard and Google, who were revealed to have made a deal to match up credit card purchases to online advertising campaigns without sharing customer data[68], allowing them to attribute particular transactions as successful conversions against a particular advertising campaign. While details of the technique are not revealed, it is likely that customers were matched using an anonymous hash of credit card details and date/amount of purchase, which would be sufficient to enable matching without sharing any personal data (though whether this is safe from an individual privacy perspective is a matter of debate[69]).

What this pattern reveals is that some organisations can have more signals than others, ultimately making them more powerful actors to be able to advertise with precision, or to sell that knowledge to advertisers so that they can. It also shows that working with other parties in "compute together" relationships allows the collection of more signals or the deeper interpretation of signals without even needing to share personal data. This foreshadows the existence of a network of data brokers and adtech companies, which will be explored more in Section 3.3.

[66] Matt Levine, *Capital One Fraud Researchers May Also Have Done Some Fraud*, Bloomberg https://www.bloomberg.com/opinion/articles/2015-01-23/capital-one-fraud-researchers-may-also-have-done-some-fraud

[67] coined by Markus Lohi, one of the #digipower participants

[68] Mark Bergen and Jennifer Surane, Google and Mastercard Cut a Secret Ad Deal to Track Retail Sales, 30 Aug 2018, https://www.bloomberg.com/news/articles/2018-08-30/google-and-mastercard-cut-a-secret-ad-deal-to-track-retail-sales

[69] Guidance for businesses on hashing data, Pinsent Masons, 25 Nov 2019, https://www.pinsentmasons.com/out-law/analysis/guidance-for-businesses-on-hashing-data

Why this matters to society, is that the more users a platform or service has, and the greater their processing/analytical power, the more they can know at individual and population level, which translates into them having huge amounts of power to influence populations (or make money by selling that knowledge, or that ability, to others).

We know that companies advertise to us every day in physical spaces (through billboards, bus ads, and TV and radio ads). Some degree of tailoring to individual demographics can be achieved in postal mailings, where different versions of catalogues can be sent to different customers based on their past purchases. But in the online space, advertisements about physical world products and services can be tailored to specific individuals, not just groups. Advertisers like Facebook, Amazon and Google now offer advertisers very granular ways to target specific types of individuals, where adverts can be served to individuals based on particular demographics, locations, interests and intents. All of this shows an intent for advertisers to achieve successful conversions - i.e. to successfully influence a prospective customer to perform some particular action, such as purchasing a particular product or service.

Returning to our participants' data, we can illustrate this using Miapetra Kumpula-Natri's intended visit to the Fat Lizard restaurant in Helsinki. In Figure 33 below, an extract from Google Maps, we see that the Maps presented to users in Google (and other Maps apps) are not static, objective representations of the world. **The information that is highlighted will be different** depending on individuals' previous signals of intent, and upon the business dealings the map provider (such as Google) has made with different venues. This is evident in this map because alongside the Fat Lizard restaurant (the search target) is the Eat Poke Kaivopiha restaurant, which has a square icon, indicating they are paying for Google Maps ads. This restaurant likely paid for a form of advertising that will show up primarily to users searching for known food types or known city centre restaurants (signals of intent like those her Maps search provided earlier). This paid placement makes their businesses more prominent on the map at different scales.

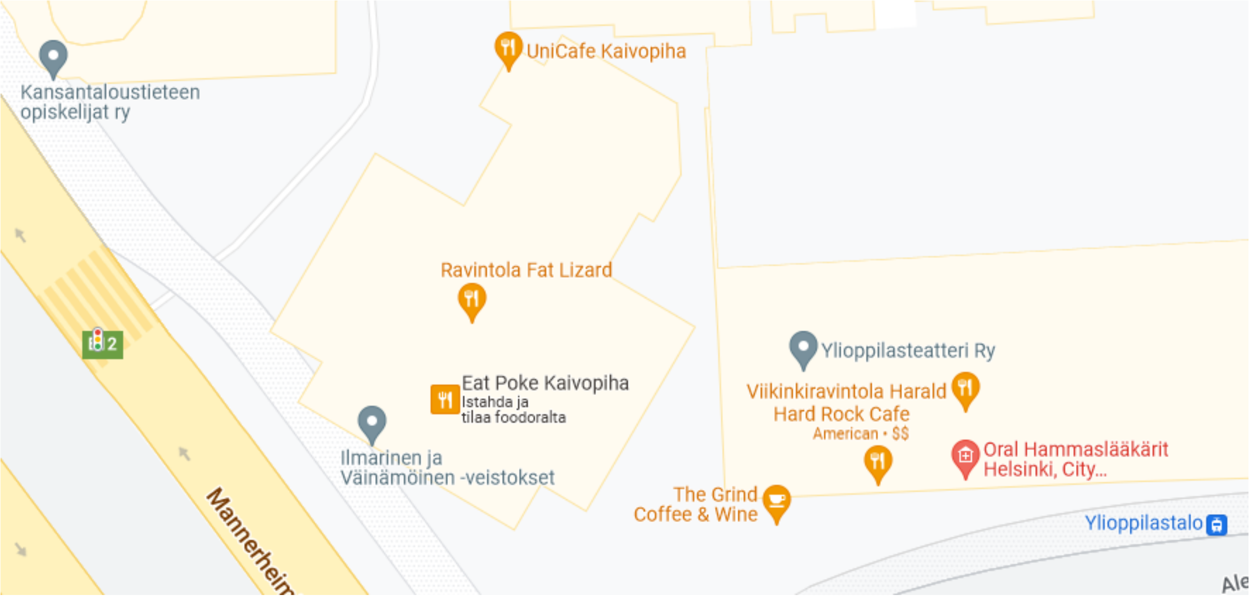


Figure 33: The searched-for Ravintola Fat Lizard *restaurant, sitting alongside the* Eat Poke Kaivopiha *restaurant, which has been indicated to the user as the result of a payment for Google Maps advertising.*

While we all experience the same streets, billboards and adverts, the digital representation of the physical world can be manipulated to influence the choices that individuals make. The most obvious example of this is search engines, where the results are not neutral or objective, but tailored specifically to individuals[70]. This phenomenon is known as the Search Engine Manipulation Effect, and has been shown to have significant impacts on society, not least in its impact upon voting patterns; researchers have argued for greater regulation of search engine results[71], and indeed it makes sense that anything that **alters people's view of reality** should be subject to careful scrutiny. As new technologies such as home hubs and virtual assistants begin to become part of our everyday lives, this question of being able to trust the information you are given to be factual and unbiased is even more critically important. A virtual assistant gives only one answer to a question, and that answer may be susceptible to advertiser influence[72], just like the paid icon on the Google Map.

The practices this section has begun to uncover reveal two important impacts for society: First, the impact on **individual trust.** In the physical world we regularly consult digital information

[70] Eric Goldman, Search engine bias and the demise of search engine utopianism, Yale JL & Tech. 8, 2005, https://digitalcommons.law.scu.edu/facpubs/76 ; James Grimmelmann, The Google Dilemma, New York Law School Law Review, 2009 https://ssrn.com/abstract=1160320
[71] Robert Epstein *et al.*, Suppressing the Search Engine Manipulation Effect (SEME), November 2017, https://doi.org/10.1145/3134677
[72] Valerie K. Jones, Voice-activated change: Marketing in the age of artificial intelligence and virtual assistants, Journal of Brand Strategy, Winter 2018, https://www.ingentaconnect.com/content/hsp/jbs/2018/00000007/00000003/art00005

sources, and must come to terms with the fact that many of them (especially when they are free, and hence advertising-funded) are biassed. We are influenced more than we realise.

The second impact is that **the sacrifice of our data fuels a growing imbalance in power**, the same imbalance of power described by the World Economic Forum that we referenced in 2.1.1. It is no longer a level playing field. Society is reconfigured in favour of advertisers and technology platforms: Those who can afford to advertise get an advantage. And those with the platforms that can use our devices and apps to gather data signals from us get to profit from those advertisers and get to change our world. It is harder for smaller companies to gain influence, which in turn redirects more of our money towards larger players (as seen with Amazon's dominance in book selling and online retail, for example), and it is harder for individuals to operate freely in the face of such dominant forces.

The only saving grace is that the physical world does carry one advantage over the digital though; we can, cost-permitting, travel to a place and **inspect the truth with our own eyes**. As we will see in 3.2, this is not so easy in the digital context.

3.2. WHEN YOU VIEW THE WEB, THE WEB VIEWS YOU

In this section, we draw parallels with the previous section where an individual moving, acting and being tracked in the physical world, which can be related to an individual moving around the web to consume content. At the same time, we will contrast the places where differences occur.

A FarmVille player can also move around. The main difference is, of course, that unlike the physical world, FarmVille is entirely constructed. Zynga can observe all player actions, can seek to influence the player, and of course the available choices are structured in advance by the game designer. In a sense FarmVille has more similarities to an online platform delivering (interactive) content such as news or social media posts than it does to the physical world[73]. This is the tension that we will explore in this section: when we click hyperlinks to move around the web and consume content, when do we have agency that resembles that of the physical world, and when do we have the negligible agency of a FarmVille player?

When our main context is reading or consuming content rather than visiting physical places, we should ask questions such as: which signals are captured? How can these signals be used to acquire knowledge about us? How can that knowledge be used to structure our available action choices within that world?

For instance, the dating app OkCupid runs A/B tests on user's data behind their back, influencing this way what is seen and known by the users about the app and their peers. The app changed the "match rating" (the likely compatibility of two users according to comparison of questions answered in their profiles). "OKCupid actually lied to users about what they were seeing on the

[73] Farmville "Quests" (https://farmville.fandom.com/wiki/Quest) could then be seen as the equivalent of news or social media content that FarmVille player visit, progress through and consume.

website. The company took pairs of users with low match ratings of around 30% (the ideal match is 100%) and told them they were a 90% match. They also did the opposite, giving highly-compatible pairs of users low match ratings”.[74] In this way they were able to engineer new perceptions of their users of each other, and to learn from this. We want to explore similar situations across many other content types.

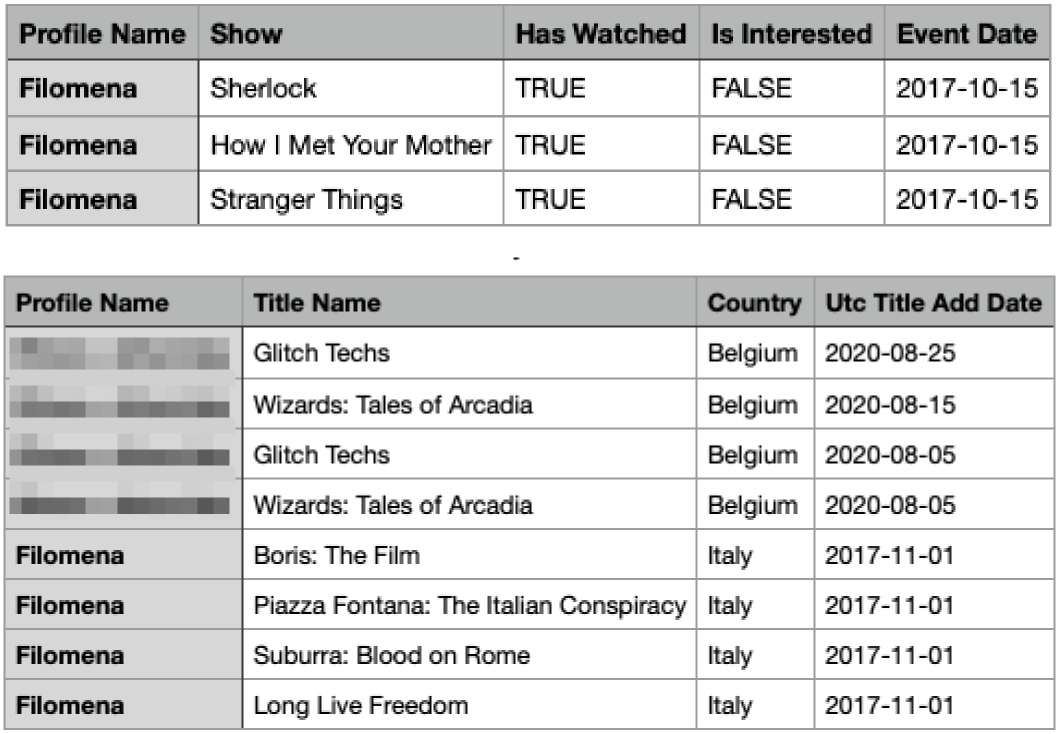
In Section 3.2.1 we examine how, as in Section 3.1.1, companies capture signals of individual activity (in this case, engaging with content rather than visiting physical locations) in order to infer and assert knowledge about their interests and predict future intentions.

Then, in Section 3.2.2, we show how in the digital context, the companies that design the interfaces through which we read and consume information and content have almost unlimited ability to structure and shape the available choices we can make, just like in FarmVille - a capability that is not available in the same way in the physical world described in Section 3.1.

3.2.1. Your Content Choices Define You

In 3.1 we saw multiple examples of users expressing their intent to travel somewhere, more or less explicitly, such as Miapetra Kumpula-Natri moving around Brussels and Helsinki. Some content-based websites and apps similarly enable users to curate their own channels, categories, bookmarks or wishlists, or infer interests themselves, acting as a direct indication of content preference. The examples are varied: for instance Jyrki Katainen’s K-Ryhmä app kept a list of his preferred recipes (a form of content, after all). Another example was the interests data that Netflix reports they hold about the members of Filomena Chirico’s household - including shows that have been added to “My List”, as well as other preferences that have been explicitly expressed within the Netflix UI (Figure 34).

[74] OKCupid Admits To Purposely Giving Users Bad Matches In Site 'Experiment' https://www.businessinsider.com/okcupid-lied-to-its-users-to-help-them-2014-7?op=1&r=US&IR=T, Insider, Jul 29, 2014.



Profile Name	Show	Has Watched	Is Interested	Event Date
Filomena	Sherlock	TRUE	FALSE	2017-10-15
Filomena	How I Met Your Mother	TRUE	FALSE	2017-10-15
Filomena	Stranger Things	TRUE	FALSE	2017-10-15

Profile Name	Title Name	Country	Utc Title Add Date
[redacted]	Glitch Techs	Belgium	2020-08-25
[redacted]	Wizards: Tales of Arcadia	Belgium	2020-08-15
[redacted]	Glitch Techs	Belgium	2020-08-05
[redacted]	Wizards: Tales of Arcadia	Belgium	2020-08-05
Filomena	Boris: The Film	Italy	2017-11-01
Filomena	Piazza Fontana: The Italian Conspiracy	Italy	2017-11-01
Filomena	Suburra: Blood on Rome	Italy	2017-11-01
Filomena	Long Live Freedom	Italy	2017-11-01

Figure 34: Sample of Netflix's records of users' saved interests ("My List", bottom) and indicated preferences (top), from Filomena Chirico's SAR return.

We see similar interest records in other participants' SAR returned data from other media platforms that participants targeted, including Spotify, Bookbeat, BBC Sounds, Sanoma (the Ruutu and Nelonen channels), YLE and Telia (MTV3).

Sometimes, however, the intention is not revealed explicitly, and has to be deduced. In the previous section, when Mark Scott lingered at a location, multiple raw wifi traces could be used to deduce which store he exactly was at - he did not make this fact explicit. Similarly, when a user spends time on a piece of content, they are assumed to have some affinity for that content, which can then be used as input to build a model describing the reader's interest.

In his Sanoma Group SAR return, Atte Harjanne was able to see those deductions. It is clear that Sanoma has algorithms which attempt to infer what topics he is interested to read about, based on his past reading activity. In Figure 35 we show a graph illustrating the SAR-returned data they hold about his inferred reading interests.

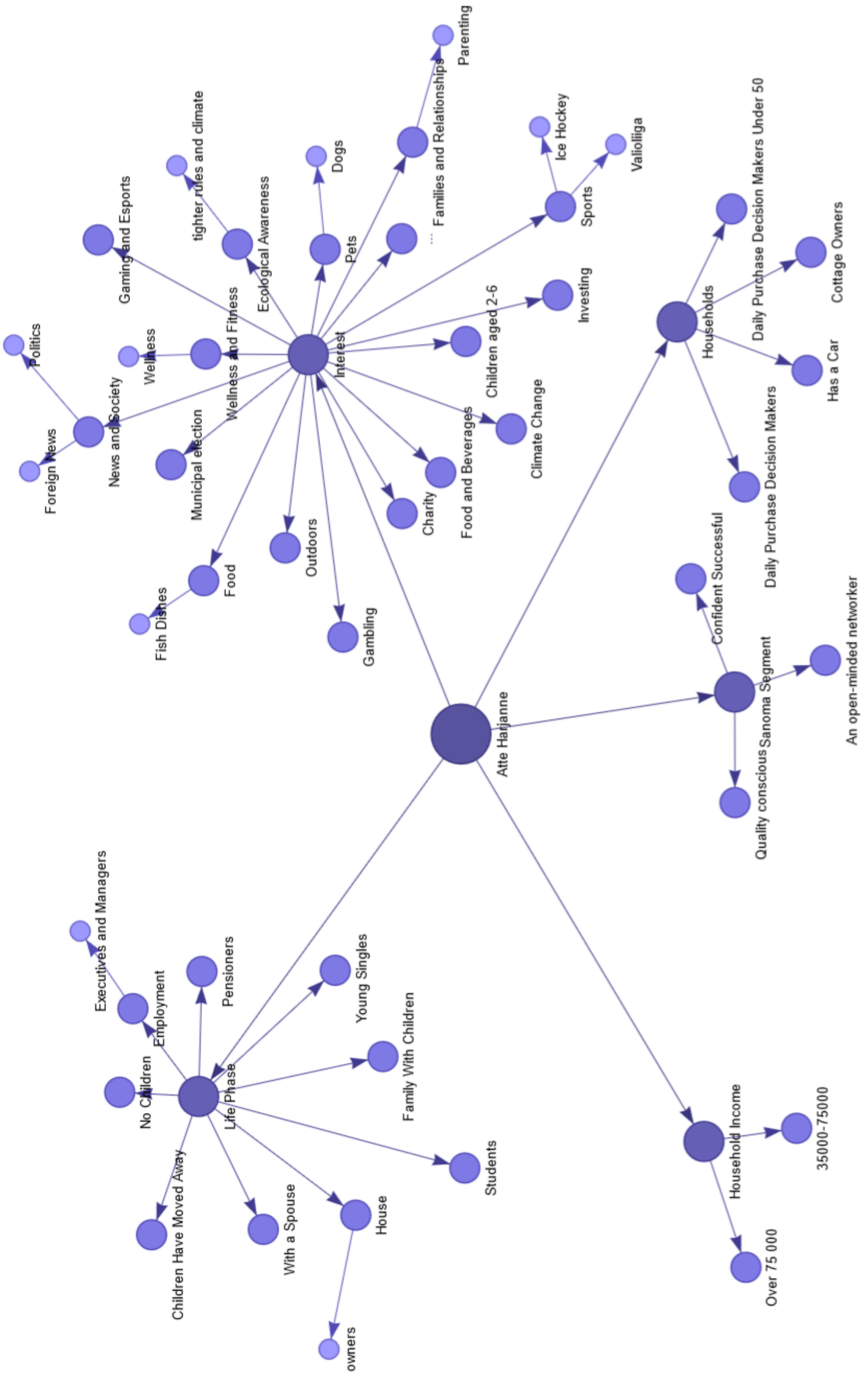


Figure 35: A graph of Atte Harjanne's inferred reading interests, as determined by the Sanoma group based on the articles he read, mostly on Helsingin Sanomat, their flagship newspaper.

Although those interests can sometimes be used to inform what content to recommend to the user on future visits (depending on the level of sophistication of the publisher)[75], these reading interests are typically used to decide which ads to show alongside content, and are sold to the highest bidder.

It is interesting that one of the profiling interests is "gambling". Gambling is an extremely problematic interest to profile on. Indeed, such an attribute can be used in a variety of ways to exploit people's psychological weaknesses[76]; imagine someone who has an addiction and is constantly reminded of the opportunity to indulge in it, the next bet always being only one click away.

Atte Harjanne is not personally interested in gambling, but that interest is nevertheless added to his Heslingin Sanomat profile. He does however have a professional interest in the topic, as a lawmaker: Harjanne wrote a blogpost[77] that was cited in an article on the topic in Helsingin Sanomat[78]. Presumably, he looked at the Helsingin Sanomat article after having consulted many other articles on the topic, and this led to Helsingin Sanomat adding "gambling" as an interest to his profile. It is debatable whether this was the correct move (it is an interest of his, after all), but without that additional context he will invariably be targeted by some advertisers who will naturally interpret "interest in gambling" as allowing the targeting of gamblers. Harjanne would then see gambling ads. He might then deduce that gambling is an even more serious problem than he thought (unlike his colleagues who do not share his legislative interest). The feedback effects here are very weak, because they require active targeting by advertisers, and the clear distinction between content and advertisements helps Harjanne keep a cognitive separation for the different purposes of the different pieces of content. Nonetheless, this example foreshadows what we will discuss later in this section.

During our investigation, multiple participants targeted the Sanoma group. While Atte Harjanne, Jyrki Katainen and Dan Koivulaakso targeted Helsingin Sanomat, Sari Tanus targeted Aamulehti, another newspaper in the Sanoma group. We saw from the returns that data seemed to flow freely between these different newspapers within the same business group, for the purpose of profiling reader interests. We built an aggregate picture in Figure 36, from which one can start to anticipate the power to influence subpopulations that result from building such an aggregate picture.

[75] This Converting behaviour is described further in the narrative Report, Section 3.6, lever 2

[76] See Cracked Labs, *Digital Profiling in the Online Gambling Industry*, https://crackedlabs.org/en/gambling-data, January 2022.

[77] https://atteharjanne.fi/2021/08/30/veikkauksen-tarina-alkaa-taputeltu-rahapelipolitiikan-remontti-kiistatta-tarpeen/

[78] https://www.hs.fi/visio/art-2000008262414.html

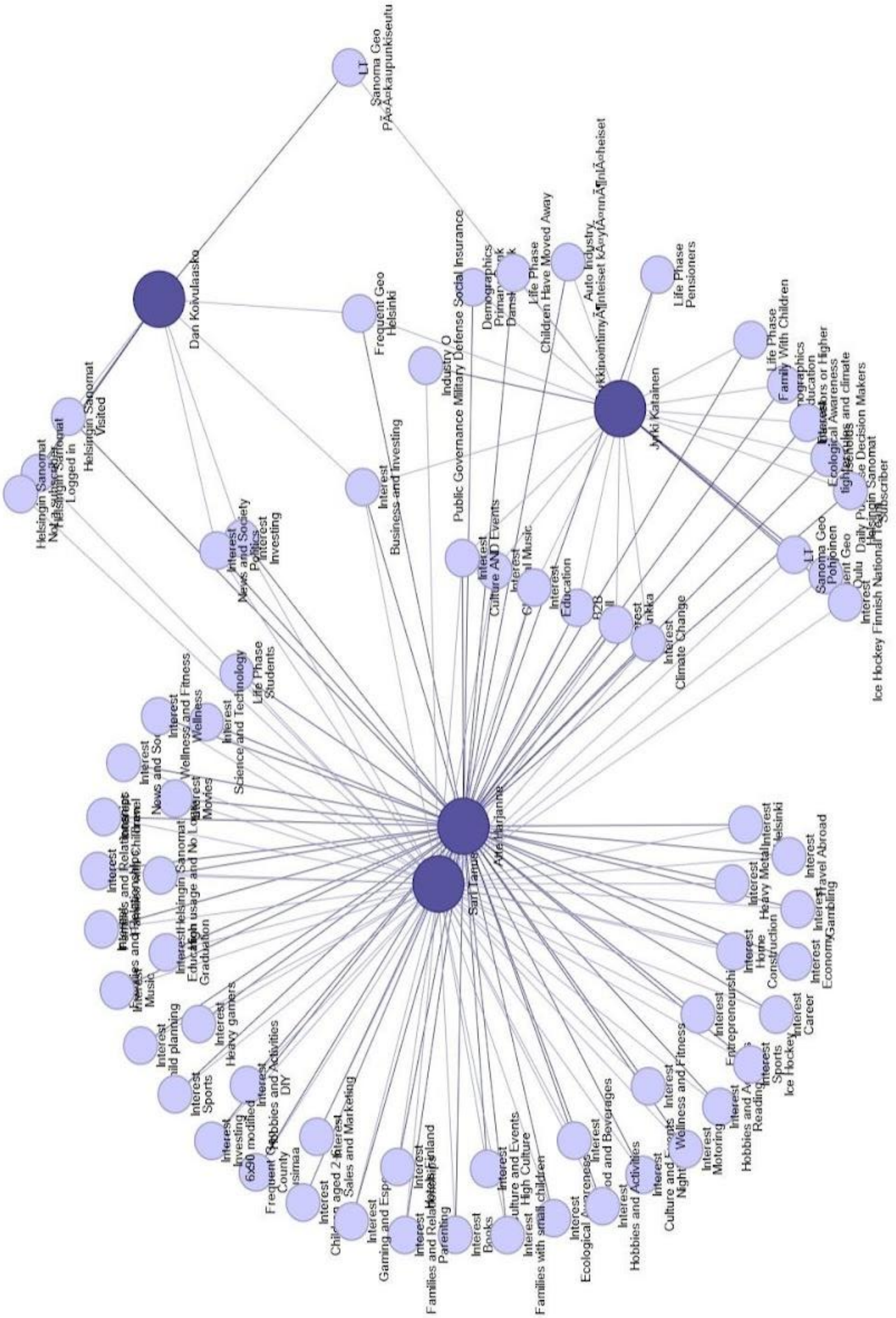

Dan Koivulaakso
Jyrki Katainen
Atte Harjanne
Sari Tanus
LT Sanoma Geo PÃ¤Ã¤kaupunkiseutu
Frequent Geo Helsinki
Helsingin Sanomat Visited
Helsingin Sanomat Logged in
Interest News and Society
Interest Politics
Interest Investing
Interest Business and Investing
Industry O
Public Governance Military Defense Social Insurance
Demographics Primary Bank Danske
Life Phase Children Have Moved Away
Auto Industry
Life Phase Pensioners
Life Phase Family With Children
Interest Culture AND Events
Interest Classical Music
Interest Education
B2B
Interest Climate Change
LT Sanoma Geo Pohjoinen
Frequent Geo Oulu
Helsingin Sanomat Subscriber
Interest Ice Hockey Finnish National Team
Life Phase Students
Interest Science and Technology
Interest Wellness and Fitness
Interest Wellness
Interest Movies
Interest Helsinki
Interest Travel Abroad
Interest Heavy Metal
Interest Gambling
Interest Home Construction
Interest Economy
Interest Career
Interest Entrepreneurship
Interest Sports Ice Hockey
Interest Hobbies and Activities Reading
Interest Motoring
Interest Food and Beverages
Interest Hobbies and Activities
Interest Ecological Awareness
Interest Families with small children
Interest High Culture
Interest Culture and Events
Interest Books
Interest Parenting
Interest Families and Relationships
Interest Gaming and Esports
Interest Sales and Marketing
Interest Hobbies and Activities DIY
Interest Heavy gamers
Interest Music
Interest Sports
Interest Child planning
Interest Education Graduation
6x90 modified
Interest Investing
Frequent Geo County

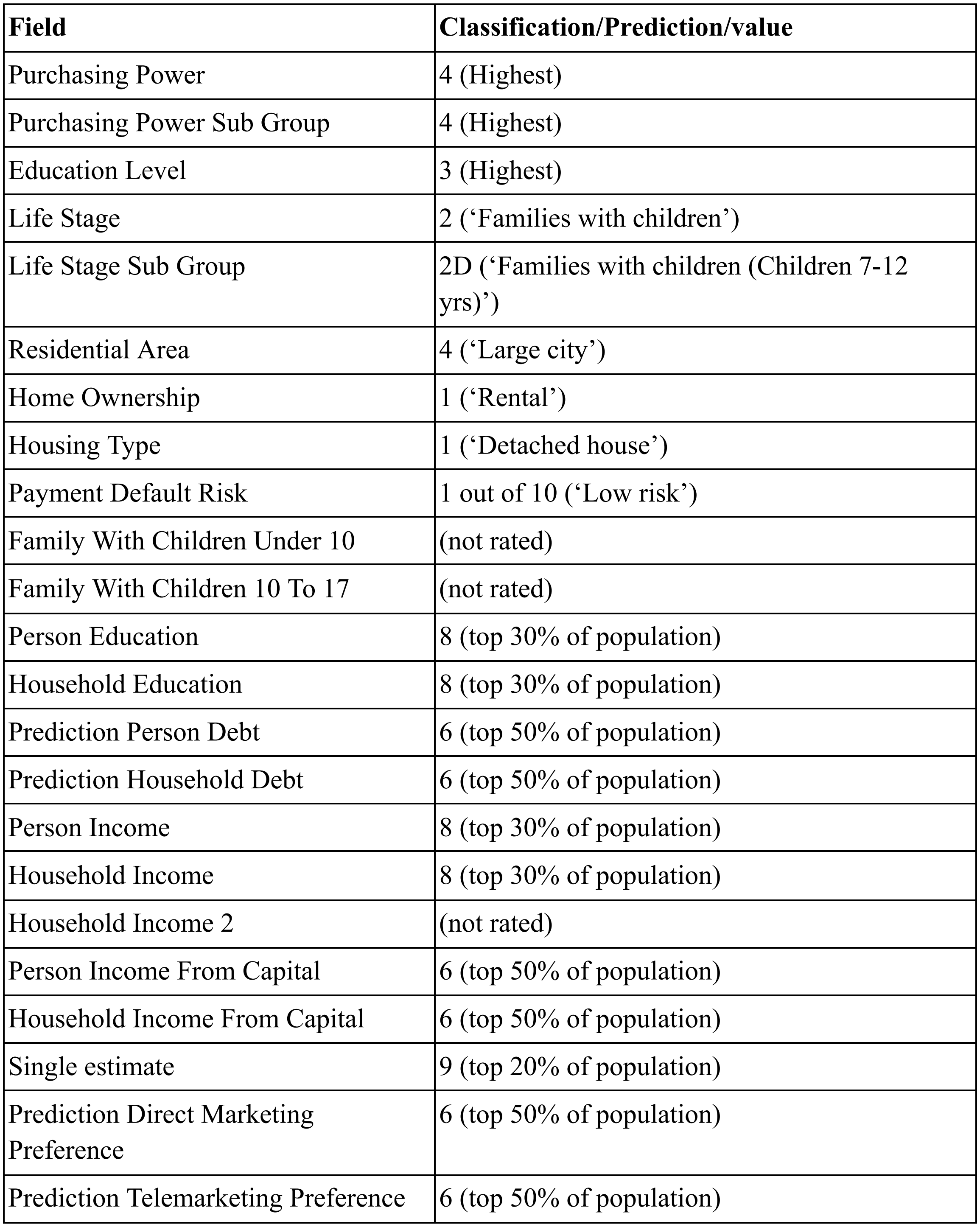
Figure 36: Interest profiles built by the Sanoma Group for Atte Harjanne, Jyrki Katainen, Dan Koivulaakso and Sari Tanus.

It is natural at this stage to turn to the financial incentives behind the content production. First off, not all revenue is tied to selling ads. Within the files that Sari Tanus got back from Aamulehti, she was told that her data included calculations of her predicted likelihood (‘purchase propensity’) to *buy different Sanoma group newspapers, magazine, comics and subscriptions*. For example, based on analysis of “demographics, order history, history of previous purchases and digital service usage data” they attribute a 4% likelihood that she might be inclined to purchase the HS Teema magazine but only a 0.1% likelihood that she would be interested in buying HS Digi. This is a good example of derived data (see Section 2.6).

Companies conduct analyses such as these in-house, but often make use of third parties to acquire additional data about their customers or prospective customers. Sari Tanus found evidence of the use of data broker Bisnode data in her output from Aamulehti. According to Sanoma’s SAR response:

> *“We use consumer analyses and forecasts prepared by our partner Dun & Bradstreet / Bisnode Finland Oy, which we use in particular to target our marketing. The projected classifications are calculated using mathematical modelling methods based on the statistical office’s area, age and sex data.”*

Looking into her data, we found that Bisnode had sent Sanoma data about Sari Tanus every three months. The most recent dataset of predictions at time of writing is shown in Figure 37.

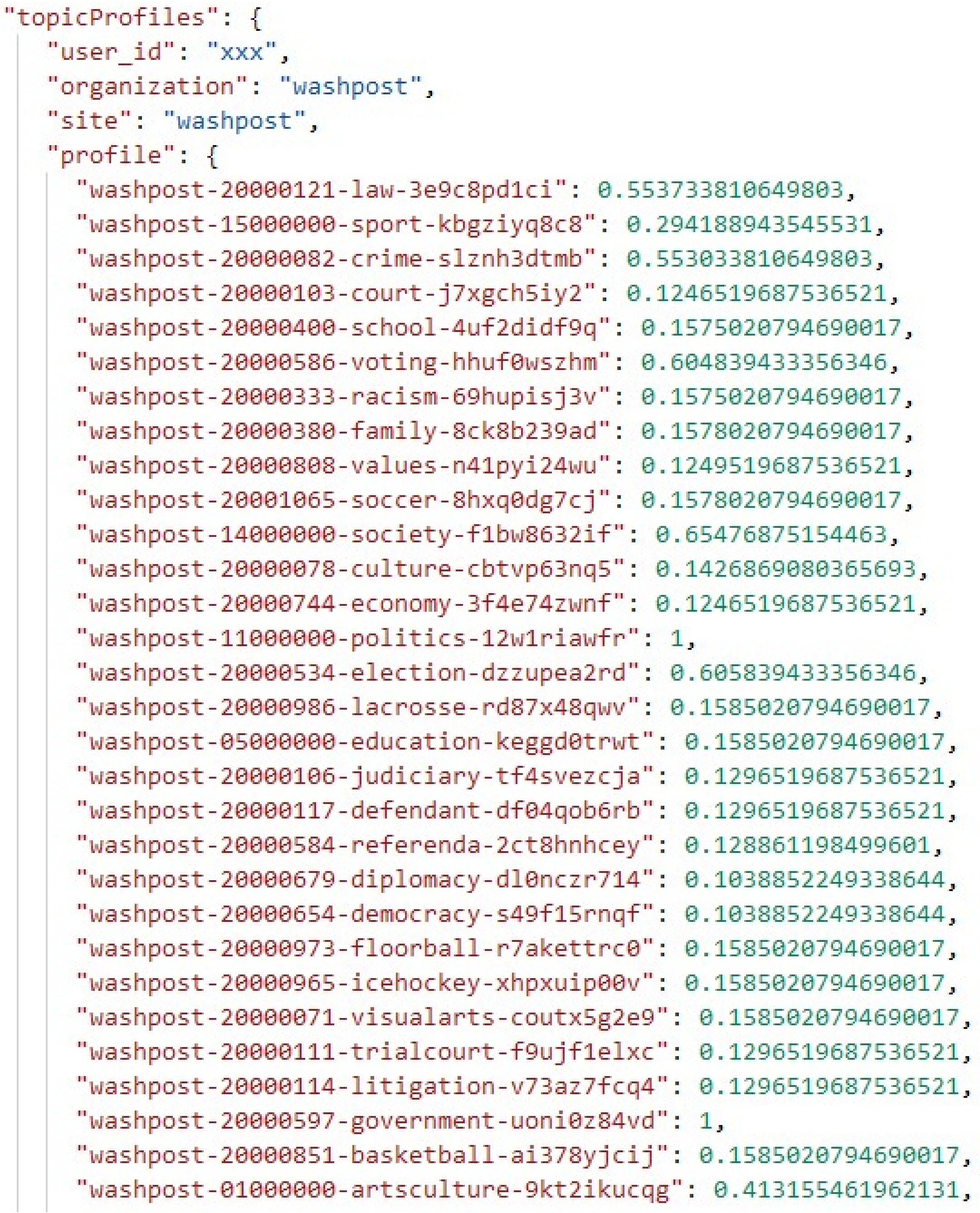


Field	Classification/Prediction/value
Purchasing Power	4 (Highest)
Purchasing Power Sub Group	4 (Highest)
Education Level	3 (Highest)
Life Stage	2 (‘Families with children’)
Life Stage Sub Group	2D (‘Families with children (Children 7-12 yrs)’)
Residential Area	4 (‘Large city’)
Home Ownership	1 (‘Rental’)
Housing Type	1 (‘Detached house’)
Payment Default Risk	1 out of 10 (‘Low risk’)
Family With Children Under 10	(not rated)
Family With Children 10 To 17	(not rated)
Person Education	8 (top 30% of population)
Household Education	8 (top 30% of population)
Prediction Person Debt	6 (top 50% of population)
Prediction Household Debt	6 (top 50% of population)
Person Income	8 (top 30% of population)
Household Income	8 (top 30% of population)
Household Income 2	(not rated)
Person Income From Capital	6 (top 50% of population)
Household Income From Capital	6 (top 50% of population)
Single estimate	9 (top 20% of population)
Prediction Direct Marketing Preference	6 (top 50% of population)
Prediction Telemarketing Preference	6 (top 50% of population)

Figure 37. Sari Tanus’ most recent profile held by Bisnode, as supplied to the Sanoma group.

Even after making a followup SAR to Bisnode, the picture of exactly what data points were used by Bisnode to make these calculations, and from which companies they were obtained, remains unclear.

Of course another revenue generating stream for newspapers consists of advertising, to which we now return, bearing in mind Atte Harjanne's *legislative* interest in gambling. Indeed, Anders Adlercreutz also obtained his interest profile from a newspaper, this time from *The Washington Post* (Figure 38).



```
"topicProfiles": {
   "user_id": "xxx",
   "organization": "washpost",
   "site": "washpost",
   "profile": {
     "washpost-20000121-law-3e9c8pd1ci": 0.553733810649803,
     "washpost-15000000-sport-kbgziyq8c8": 0.294188943545531,
     "washpost-20000082-crime-slznh3dtmb": 0.553033810649803,
     "washpost-20000103-court-j7xgch5iy2": 0.1246519687536521,
     "washpost-20000400-school-4uf2didf9q": 0.1575020794690017,
     "washpost-20000586-voting-hhuf0wszhm": 0.604839433356346,
     "washpost-20000333-racism-69hupisj3v": 0.1575020794690017,
     "washpost-20000380-family-8ck8b239ad": 0.1578020794690017,
     "washpost-20000808-values-n41pyi24wu": 0.1249519687536521,
     "washpost-20001065-soccer-8hxq0dg7cj": 0.1578020794690017,
     "washpost-14000000-society-f1bw8632if": 0.65476875154463,
     "washpost-20000078-culture-cbtvp63nq5": 0.1426869080365693,
     "washpost-20000744-economy-3f4e74zwnf": 0.1246519687536521,
     "washpost-11000000-politics-12w1riawfr": 1,
     "washpost-20000534-election-dzzupea2rd": 0.605839433356346,
     "washpost-20000986-lacrosse-rd87x48qwv": 0.1585020794690017,
     "washpost-05000000-education-keggd0trwt": 0.1585020794690017,
     "washpost-20000106-judiciary-tf4svezcja": 0.1296519687536521,
     "washpost-20000117-defendant-df04qob6rb": 0.1296519687536521,
     "washpost-20000584-referenda-2ct8hnhcey": 0.128861198499601,
     "washpost-20000679-diplomacy-dl0nczr714": 0.1038852249338644,
     "washpost-20000654-democracy-s49f15rnqf": 0.1038852249338644,
     "washpost-20000973-floorball-r7akettrc0": 0.1585020794690017,
     "washpost-20000965-icehockey-xhpxuip00v": 0.1585020794690017,
     "washpost-20000071-visualarts-coutx5g2e9": 0.1585020794690017,
     "washpost-20000111-trialcourt-f9ujf1elxc": 0.1296519687536521,
     "washpost-20000114-litigation-v73az7fcq4": 0.1296519687536521,
     "washpost-20000597-government-uoni0z84vd": 1,
     "washpost-20000851-basketball-ai378yjcij": 0.1585020794690017,
     "washpost-01000000-artsculture-9kt2ikucqg": 0.413155461962131,
```

Figure 38: Anders Adlercreutz' Washington Post profiling information (slightly obfuscated and modified for privacy reasons)

We understood his profile to represent a set of article topics, with a percentage value for each one. So for example, politics and government show a value of 1 suggesting 100% likelihood that he would read those topics, whereas soccer has only 0.157 or 15%. Upon investigation, we learned that these numbers are produced by a recommendation engine called Clavis, which was built by The Washington Post to suggest articles to readers. It is inspired by Amazon's product recommendation engine (Amazon CEO Jeff Bezos' acquired The Washington Post in 2013).

From the Clavis dataset, we see that *The Washington Post* predicts Anders Adlercreutz's preferred news topics are Politics, Society, Elections, Law, Crime and Arts and Culture. This profiling feeds into another product the *Washington Post* has created: the Zeus Platform. This is a machine for converting article reads into actionable advertisement targeting. In this way, the articles you read directly contribute to a re-shaping of the user experience. Put simply, advertisements will be tailored to the set of topics you most often read. This product is now sold to other news organisations and content producers[79].

3.2.2. Structuring Your Content Choices to Influence Your Decisions

Yet another actor engaging in interest profiling is Twitter. In Figure 39, we present a joint view of Anders Adlercreutz, Filomena Chirico, Jyrki Katainen and Mark Scott's inferred Twitter interests. This feels, and objectively is, very similar to Figure 36, which presented the interests deduced by Helsingin Sanomat.

[79] https://www.washingtonpost.com/pr/2019/05/29/washington-posts-jarrod-dicker-opening-fastest-smartest-most-performant-monetization-stack-all-publishers/

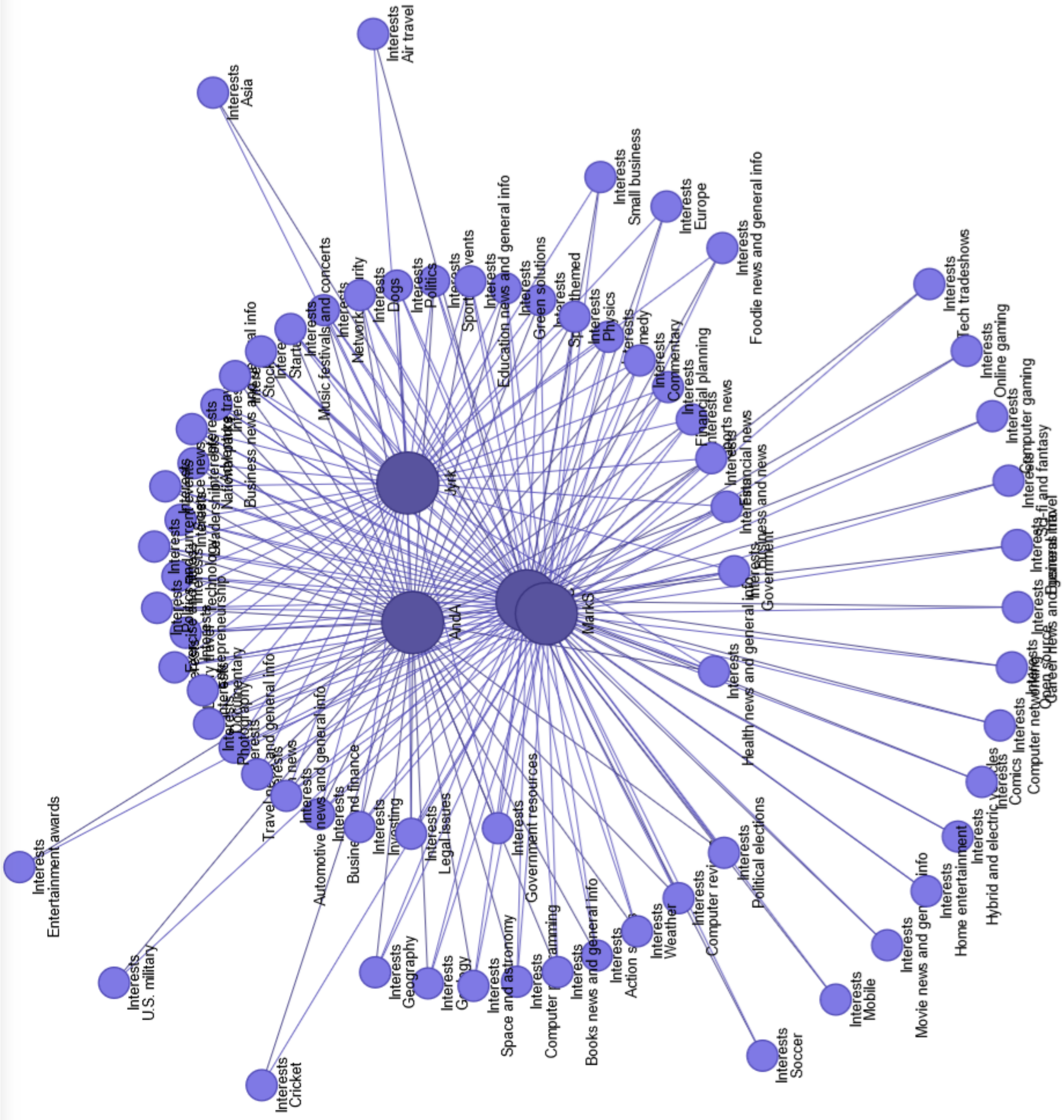


Figure 39: Twitter interests for Anders Adlercreutz, Filomena Chirico, Jyrki Katainen & Mark Scott.

However, there is a difference between Twitter's and Helsingin Sanomat's interest profiling. On Twitter, ads and original content are blended and compete in the same space, the stream. The entire context can be reshaped to maximise some carefully crafted objective, some blend of showing us ads and engaging us more on the platform. This space that was originally meant as a place of human connection becomes **a permanent place of experimentation *on* us**. This is the main breakdown of the analogy with the previous section: the **associative leap** between moving in the physical world and moving between content, which has been identified by many people before, starts to become misleading. Once we move to a feed system, where the content of the feed we can consume, and indeed the totality of the content we can see and choose from, is chosen for us[80]. Thanks to the feedback loops associated with profiling (as described in 3.2.1 above), it becomes more accurate to think of the content as moving to us, and our entire environment being crafted just for us. The situation is much more like FarmVille's optimization in the Free-to-Play format, to get us to keep on playing in order for us to spend money later.

It is important to understand this experimentation is conducted through many intermediate objectives, and does not rely on user interests exclusively. Indeed, beyond getting people to keep on viewing ads or clicking on them, the intermediate goals of a platform like Twitter extend to:

- keeping you reading on the platform, so there are more opportunities to show ads in your Twitter stream,
- keeping you writing on the platform, so there is more content alongside which to show ads in others' Twitter streams,
- keeping you engaged (retweeting, commenting, pausing in your scrolling to read certain content), so Twitter can learn more about you,
- keeping you returning to the platform, to maximise opportunities to do all of the above.

This leads to slightly different forms of profiling. For instance Twitter *Conversation Topics* (see Figure 40) are different from Twitter *Interests* (Figure 39). Indeed the latter could be engaged with passively, while the former entails a more active role by definition. For each of us, depending on the revenue we generate for Twitter based on the ads already shown to us, Twitter will seek to refine our profiling more precisely, and will seek to engage us in conversations that are more revealing. In simple terms, **Twitter can choose what topics we should talk about**. This again evokes the images of the FarmVille 'dictator' and the Big Brother totalitarian society, and can lead to group dynamics around the global conversation on certain issues.

[80] This Structuring behaviour is further described in the narrative report, Section 3.6, lever 3.

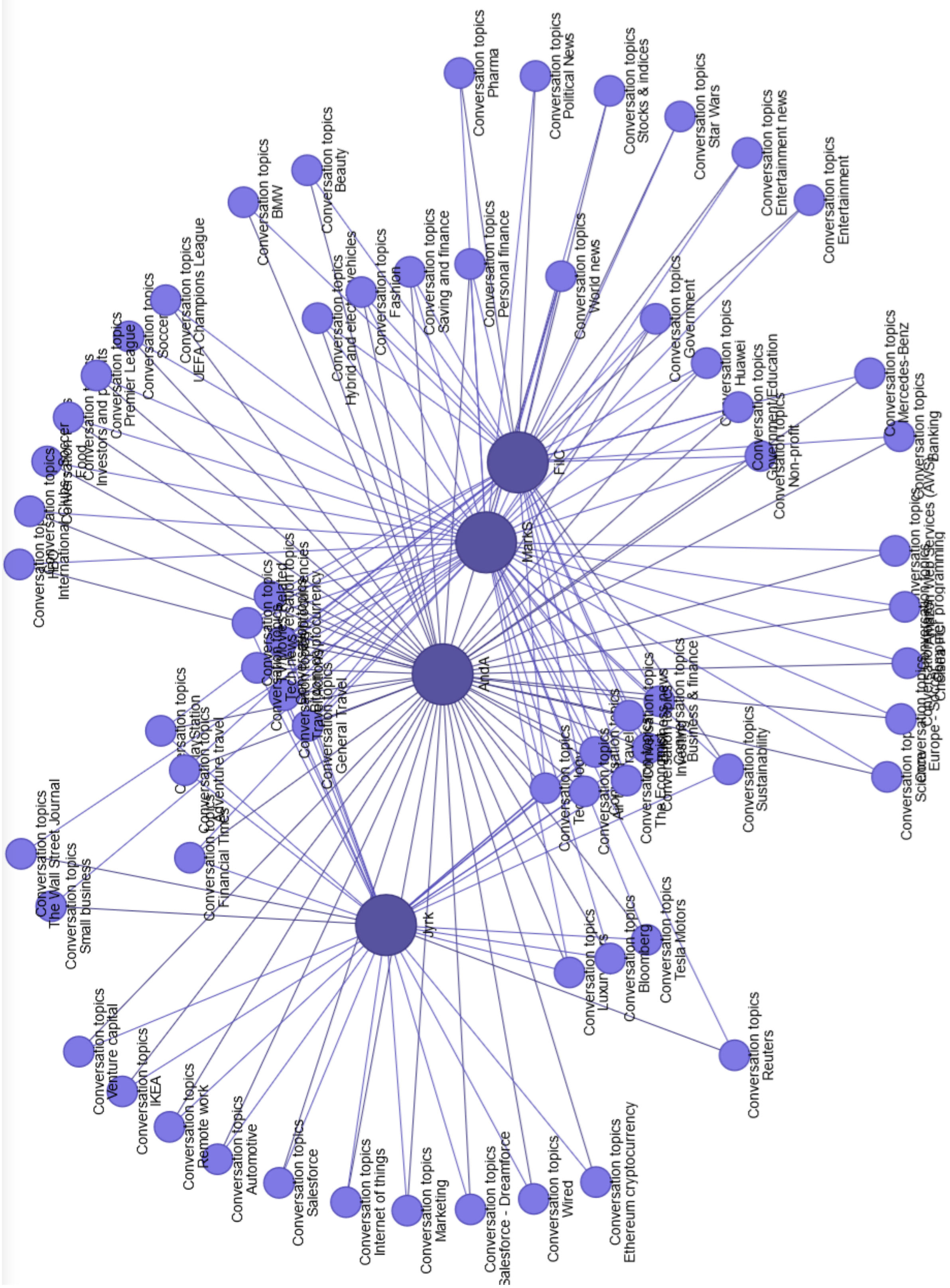
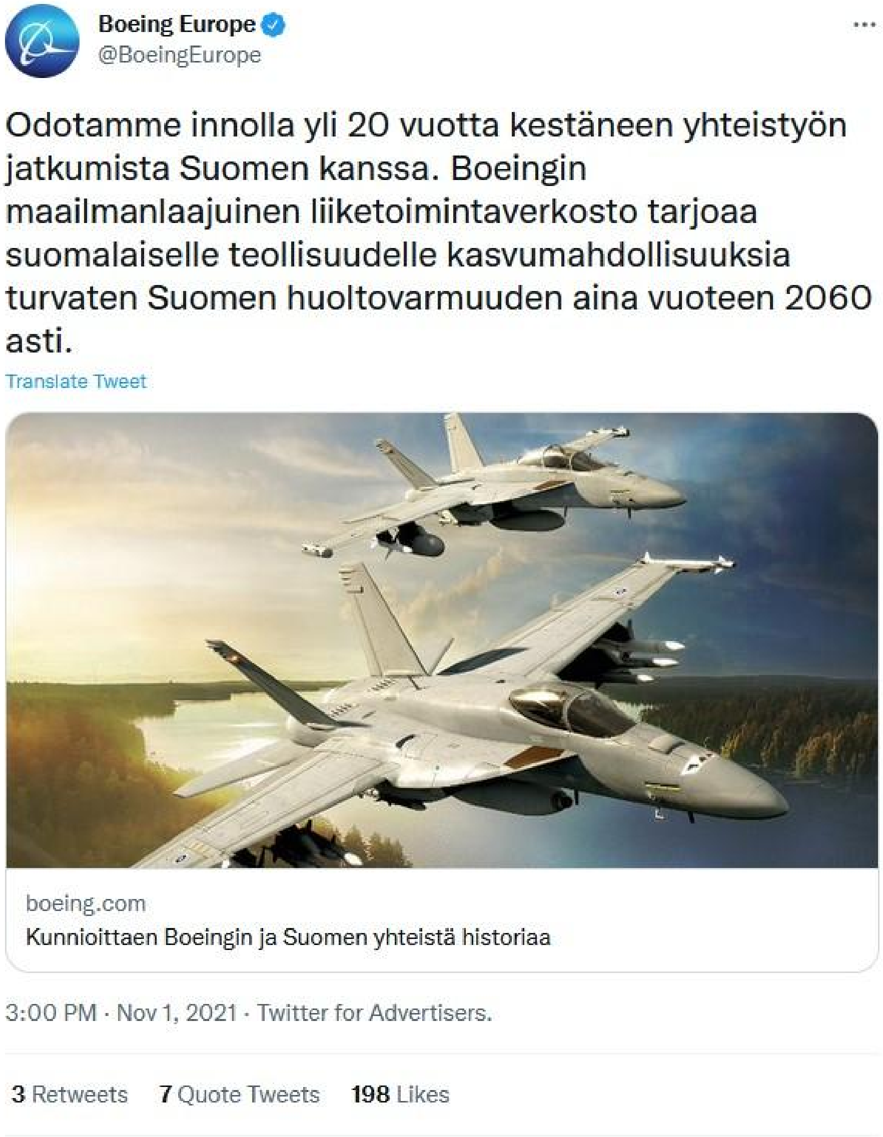

FiiC
MarkS
AndA
Jyrk
Conversation topics Pharma
Conversation topics Political News
Conversation topics Stocks & indices
Conversation topics Star Wars
Conversation topics Entertainment news
Conversation topics Entertainment
Conversation topics BMW
Conversation topics Beauty
Conversation topics Hybrid and electric vehicles
Conversation topics Fashion
Conversation topics Saving and finance
Conversation topics Personal finance
Conversation topics World news
Conversation topics Government
Conversation topics Huawei
Conversation topics Government/Education
Conversation topics Non-profit
Conversation topics Mercedes-Benz
Conversation topics Banking
Conversation topics Amazon Web Services (AWS)
Conversation topics UEFA Champions League
Conversation topics Soccer
Conversation topics Premier League
Conversation topics Investors and patents
Conversation topics Food
Conversation topics HBO
Conversation topics Tech news
Conversation topics Cryptocurrency
Conversation topics Travel
Conversation topics General Travel
Conversation topics PlayStation
Conversation topics Adventure travel
Conversation topics Financial Times
Conversation topics The Wall Street Journal
Conversation topics Small business
Conversation topics Venture capital
Conversation topics IKEA
Conversation topics Remote work
Conversation topics Automotive
Conversation topics Salesforce
Conversation topics Internet of things
Conversation topics Marketing
Conversation topics Salesforce - Dreamforce
Conversation topics Wired
Conversation topics Ethereum cryptocurrency
Conversation topics Reuters
Conversation topics Tesla Motors
Conversation topics Bloomberg
Conversation topics Luxury
Conversation topics Sustainability
Conversation topics Business & finance
Conversation topics Investing
Conversation topics The Economist
Conversation topics Business news
Conversation topics Science
Conversation topics Europe - Soccer
Conversation topics Chelsea
Conversation topics Apple
Conversation topics programming

Figure 40: Aggregate view of Conversation Topics deduced by Twitter for multiple participants.

Beyond *Interests* and *Conversation Topics*, Twitter also has the concept of *Follower Look-alikes*[81], where it selects for a user the Twitter accounts whose followers that user most resembles. This offers some interesting opportunities for targeting and influencing small populations.



Boeing Europe
@BoeingEurope

Odotamme innolla yli 20 vuotta kestäneen yhteistyön jatkumista Suomen kanssa. Boeingin maailmanlaajuinen liiketoimintaverkosto tarjoaa suomalaiselle teollisuudelle kasvumahdollisuuksia turvaten Suomen huoltovarmuuden aina vuoteen 2060 asti.

Translate Tweet

boeing.com
Kunnioittaen Boeingin ja Suomen yhteistä historiaa

3:00 PM · Nov 1, 2021 · Twitter for Advertisers.

3 Retweets 7 Quote Tweets 198 Likes

Figure 41: Boeing advertising fighter jets on Twitter. This is an example advert that was shown to Anders Adlercreutz based on who Twitter thinks he looks like.

For instance, in the context of Finland looking to purchase new fighter planes, Anders Adlercreutz and Jyrki Katainen both saw ads from Boeing. What was most interesting however

[81] https://business.twitter.com/en/help/campaign-setup/campaign-targeting/interest-and-follower-targeting.html

were the criteria used. In Jyrki Katainen's case, he saw these ads because he acts on Twitter similarly to followers of Hjallis Harkimo[82]. Hjallis Harkimo is a Finnish businessman, sportsperson, YouTuber and a member of the Finnish Parliament, with nearly 170k followers on Twitter. Concerning Anders Adlercreutz, the story is a bit more complex: he has been targeted on the exact same four criteria as Jyrki Katainen but also because he follows (or acts like users following) the Finnish politicians Marisanna Jarva, Tiina Elovaara, Antero Vartia (32K followers), Ville Skinnari (Minister for Development Cooperation and Foreign Trade of Finland), Pauli Kiuru, Markku Rossi, Antti Kaikkonen (Minister of Defence, 33.3K followers), Antti Rinne (Deputy Speaker of Parliament, 41.7K followers) and the journalists Riikka Suominen and Reijo Ruokanen. We learned later that another bidder in this fighter plane procurement battle has used the same techniques[83].

This poses interesting questions: are those people aware they have been used as targeting criteria? Should they be able to object to it? Are those people – including the Minister of Defence – not likely to also be influenced, if all their followers or people like them see those ads? How do journalists feel about being used as targeting criteria, did they really build an audience on Twitter for the exclusive benefit of Twitter?

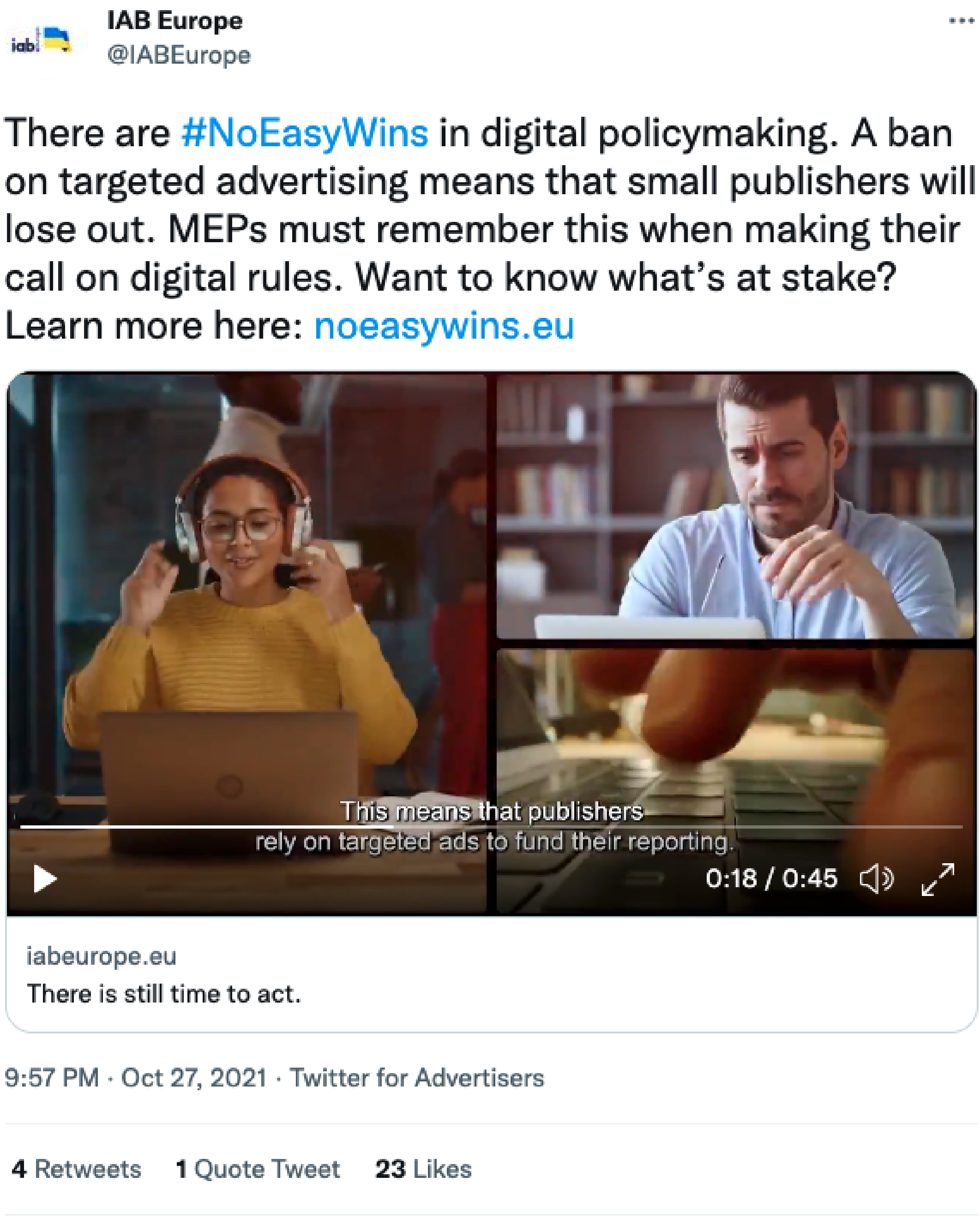
Of course, none of this is limited to Finland. As part of the #digipower investigation, we helped journalist Mark Scott investigate[84] similar lobbying by Facebook/Meta in Brussels (on Twitter!).

Filomena Chirico uncovered more of this type of targeting, ironically by the IAB, a lobbying organisation for internet advertising. In Figure 42, we see the IAB attempting to discourage her from supporting a ban on targeting advertising. The targeting criteria used for tweets like these included looking like followers of @EUCouncil, @EURACTIV, @EU_Commission, @Europarl_EN, @FT, @PoliticoEUROPE, @FinancialTimes, and @politico. This again highlighted the fact that organisations were spending money on Twitter based on audiences built by journalistic organisations.

[82] https://twitter.com/hjallisharkimo

[83] Pratt & Whitney, engine suppliers for the Lockheed Martin bid, also seem to have engaged in such targeting. In their case they used politician Jussi Halla-aho and defence expert C Salonius-Pasternak as targeting criteria. See https://twitter.com/kallemaatta/status/1511609507763400706

[84] https://www.politico.eu/newsletter/digital-bridge/meta-lobbying-google-vs-apple-russian-disinformation/



IAB Europe
@IABEurope

There are #NoEasyWins in digital policymaking. A ban on targeted advertising means that small publishers will lose out. MEPs must remember this when making their call on digital rules. Want to know what's at stake? Learn more here: noeasywins.eu

This means that publishers
rely on targeted ads to fund their reporting.
0:18 / 0:45

iabeurope.eu
There is still time to act.

9:57 PM · Oct 27, 2021 · Twitter for Advertisers

4 Retweets **1** Quote Tweet **23** Likes

Figure 42: Screenshot from a tweet[85] that IAB Europe targeted Filomena Chirico with, using her age, location and the Twitter users she resembles.

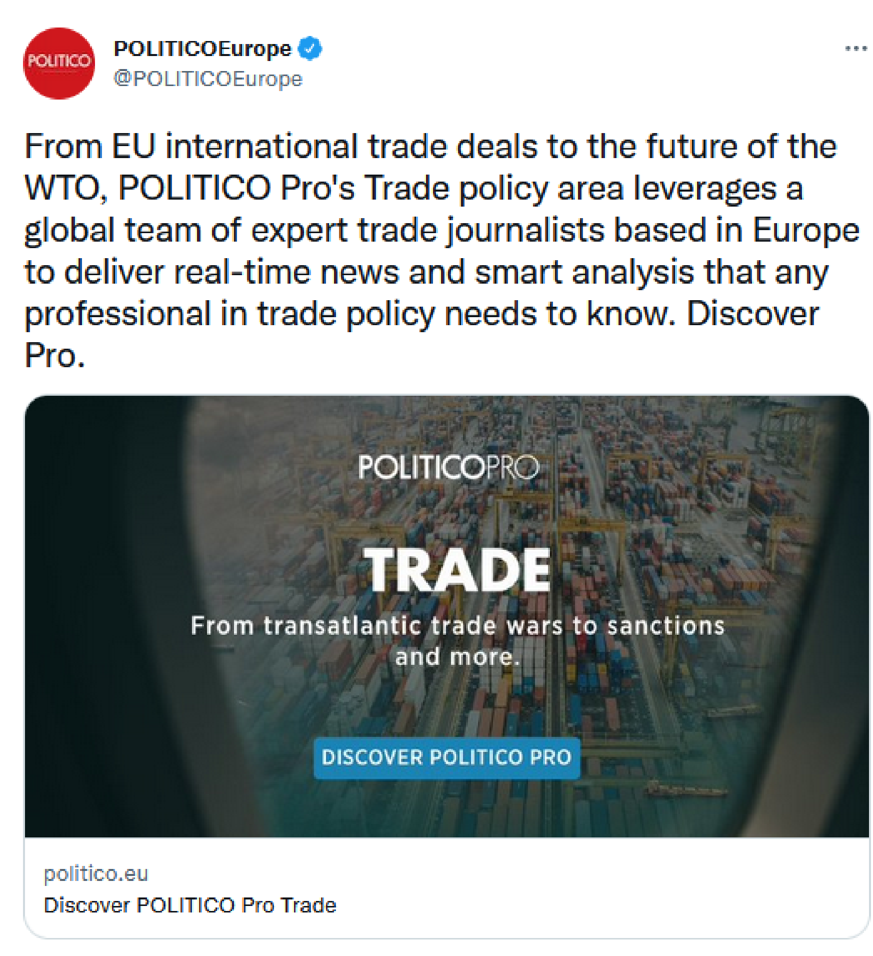
The account @PoliticoEUROPE (belonging to Mark Scott's employer) itself targeted Filomena Chirico with the following two criteria:

- Retargeting user engager: <some identifier removed for privacy reasons>

[85] https://twitter.com/IABEurope/status/1453465979879673859

- List: *ALL Events Attendants - May 2021*, *ALL PROs - May 2021*, *Brussels Playbook - 10-04-21*, *EU Confidential - 05-19-2021*, *EU Influence - 05-19-2021*.

Both of those criteria are tied to customer lists. The first is tied to retargeting of users who have engaged previously with content (by clicking a tweet), while the second criterion is due to explicit customer lists (most likely event attendance and newsletter subscription lists) that get passed on to Twitter for subsequent targeting[86].



POLITICOEurope
@POLITICOEurope

From EU international trade deals to the future of the WTO, POLITICO Pro's Trade policy area leverages a global team of expert trade journalists based in Europe to deliver real-time news and smart analysis that any professional in trade policy needs to know. Discover Pro.

politico.eu
Discover POLITICO Pro Trade

Figure 43: A tweet targeted to Filomena Chirico.

It was also interesting to observe that Jyrki Katainen had been targeted by @POLITICOEurope, because he looks like Twitter followers of Jennifer Baker (@BrusselsGeek), who is a freelance journalist covering mostly digital issues with Brussels-based groups – compounding the questions around the value of audience-building on Twitter for journalists[87].

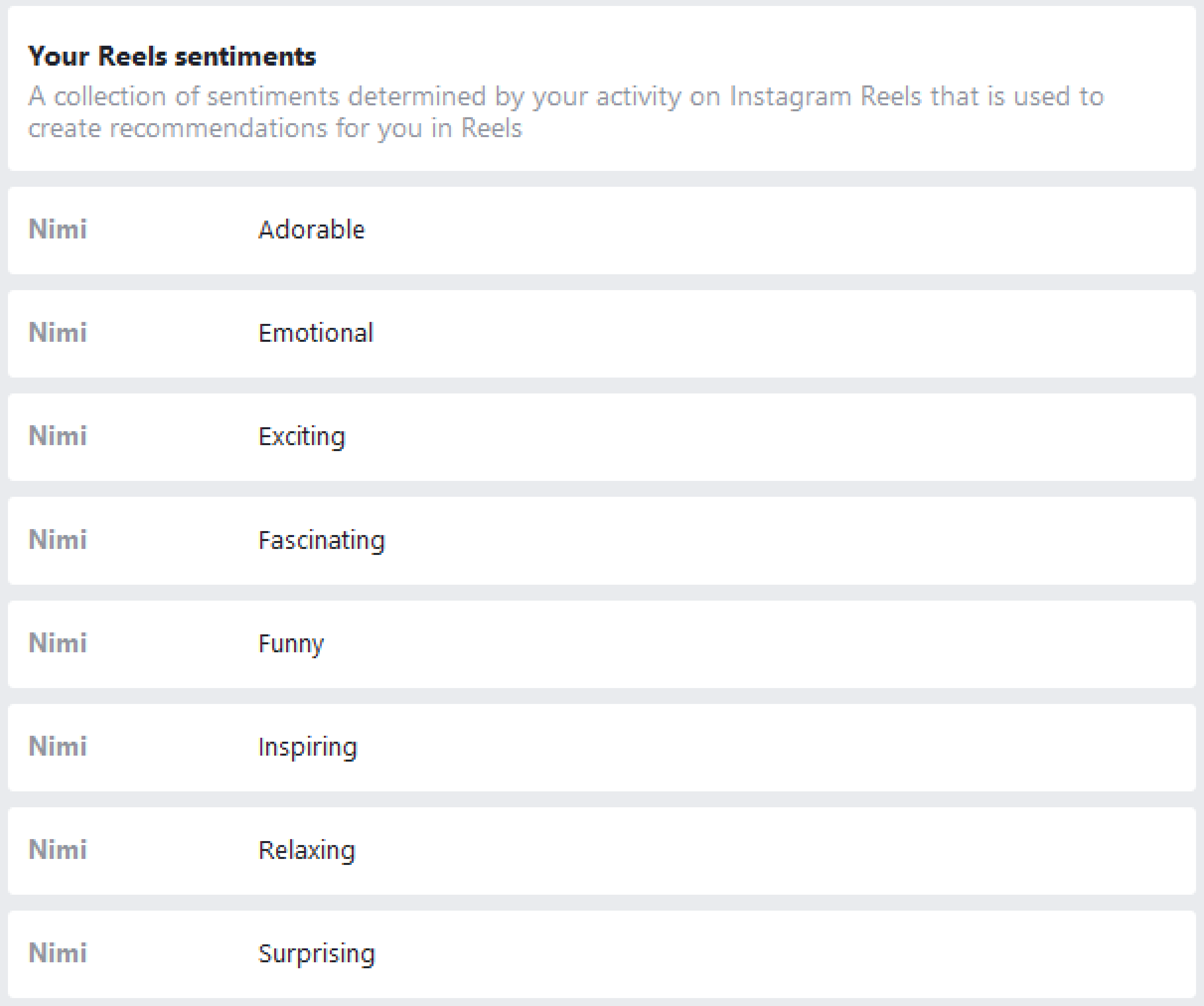
[86] We will look in more detail at this mechanism in the next section.

[87] Imagine you are a journalist who is asked to spend some time building an audience on Twitter as part of their newsroom work. You then leave that newsroom, but the publisher still targets advertisements for their content at "your" lookalike followers.

Certainly this shows a lively ecosystem of Twitter influence affecting the Brussels bubble. Again, beyond the direct ads displayed, we should think of the set of incentives built in for Twitter: Twitter wants people to engage with the topics selected as criteria by advertisers, and that will lead users to reveal more and more of their preferences, interests, and topics they are willing to engage with. This is structural to the conversation that emerges on Twitter.

None of this is unique to Twitter. Facebook, Instagram or TikTok, for instance, would be structured in the same way and have similar feedback loops. However Twitter is the most transparent on the criteria used so we can see the process and describe it more easily.

We did find evidence of some form of sentiment profiling in Instagram, but it was not exactly clear what it was or how it was used (see Figure 44).



Your Reels sentiments

A collection of sentiments determined by your activity on Instagram Reels that is used to create recommendations for you in Reels

Nimi	Adorable
Nimi	Emotional
Nimi	Exciting
Nimi	Fascinating
Nimi	Funny
Nimi	Inspiring
Nimi	Relaxing
Nimi	Surprising

Figure 44: Sari Tanus's sentiments, as used by Instagram to create recommendations in the Reels subproduct of Instagram.

The mechanisms described here have multiple negative side effects. Those tools can be used **to engineer new realities**. For instance as mentioned in the introduction to Section 3.2, OKCupid has engineered matches based on purposefully false matching metrics. Other instances are more consequential. It has been shown for instance that Russia organised[88] both a protest and a counter protest from scratch around an Islamic centre in Texas during the 2016 election. This was not done through ads, but by building audiences progressively through two Facebook groups they controlled: *United Muslims of America* and *Heart of Texas*, a group promoting secession. The audiences for those groups were built mostly organically, by sharing content that led to high engagement into those two subcommunities. Beyond veracity of the content, long term interest of those communities, or authenticity of the posters, Facebook was happy to build the audiences for those groups as they kept many people engaged on the platform. Similarly, African-American groups have been extensively targeted in the past with foreign-controlled groups. Another consequence of those tools is the micro-influence they can lead to. It has been argued for instance that this micro-influence was a national security risk[89]. This risk has materialised very concretely in the context of the 2017 UK general election: Labour leader Jeremy Corbyn (and friends) was deceived by his own party operatives into believing a particular digital campaign strategy had been adopted, while in fact it had only been rolled out for his entourage[90]. Researchers have shown that it is even possible to do influence operations on one person at a time[91], what they called nano-targeting. In so doing, they formalised and extended what had been done before by others[92]. Finally, there are health concerns around the long term use of those products which do not present a balanced view of content. For instance, Instagram is suspected to cause mental health problems and body image issues[93].

[88] Claire Allbright, *A Russian Facebook page organized a protest in Texas. A different Russian page launched the counterprotest*, The Texas Tribune, November 2017 https://www.texastribune.org/2017/11/01/russian-facebook-page-organized-protest-texas-different-russian-page-l/

[89]Jessica Dawson. "Microtargeting as Information Warfare." *The Cyber Defense Review* 6, no. 1 (2021): 63-80, https://cyberdefensereview.army.mil/Portals/6/Documents/2021_winter_cdr/CDR_Winter_2021.pdf

[90] Tim Shipman, *Labour HQ used Facebook ads to deceive Jeremy Corbyn during election campaign*, The Sunday Times, July 2018 https://www.thetimes.co.uk/article/labour-hq-used-facebook-ads-to-deceive-jeremy-corbyn-during-election-campaign-grlx75c27

[91] Gonzalez-Cabañas, Cuevas, Cuevas, Lopez-Fernandez and Garcia, *Unique on Facebook: Formulation and evidence of (nano)targeting individual users with non-PII data*,IMC 2021 - Proceedings of the 2021 ACM Internet Measurement Conference, November 2021 https://dl.acm.org/doi/10.1145/3487552.3487861

[92] for instance Michael Harf, *Sniper Targeting on Facebook: How to Target ONE specific person with super targeted ads*, December 2017, https://medium.com/@MichaelH_3009/sniper-targeting-on-facebook-how-to-target-one-specific-person-with-super-targeted-ads-515ba6e068f6

[93] *Instagram Worsens Body Image Issues And Erodes Mental Health*, NPR Public Radio, 2021 https://www.npr.org/2021/09/26/1040756541/instagram-worsens-body-image-issues-and-erodes-mental-health?t=1649637305992

In this section, we have started by describing situations where the agency of a content consumer online was comparable to that of someone navigating a city. We then transitioned to the situation of the newsfeed/stream, picking the particular example of Twitter. The capabilities of Twitter amount in the offline world to structuring entire streets to their liking, promoting particular types of shops because they attract better revenue, or to creating new intersections to facilitate mobility but also to increase traffic in front of particular billboards. In comparison to the physical world of Section 3.1, it is important to note one key difference for the world of digital content we have explored in Section 3.2: Unlike in the physical world, we cannot objectively **judge what information and processes exist** in our world, as we only see the tips of the icebergs that those with the power to shape the content landscape wish us to see. So too is our ability to accurately **judge what is real or objectively true** limited, as every 'building' (website) or 'billboard' (advert or promoted post) is presented and framed to us through interfaces built by actors who have a strong bias and a **commercially-motivated desire to influence how we think and what we will do**. The impacts of this are playing out in the world today, through filter bubbles, radicalisation, 'fake news' and the rise of populism and nationalism, although each of those is hotly debated. It is essential to highlight however that enabling such impact is precisely the point of those technologies - provided there is a customer for it. We hope that the reader will find these parallels (and differences) between the two situations helpful to guide them thinking about both the digitisation of our physical and digital contexts as well as the presentation of digital content online.

3.3. MOVE FAST AND CAPTURE ALL SIGNALS, EVERYWHERE

As we have seen in the previous sections, some key commonalities throughout these situations are:

- the capability to observe an end user performing an action;
- the capability through data to influence the choices made by the end user;
- the capability to structure the choices made available to the end user.

Throughout we have not focused however on *who* had those capabilities. The FarmVille example was useful as a reference scenario, because in that case all that power is clearly concentrated into Zynga's hands, and the quantitative aspects in particular into the hands of the data analyst presented in the introduction to Chapter 3.

We will now focus on *who* is doing what. In Section 3.3.1, we will discuss general advertising technology players (adtech), in what is a clear continuation of what had been done for the digitrail study (see Section 2.8). The *who* there, will be very diffuse, involving an entire ecosystem. However in Section 3.3.2, we will highlight the role of some players in that ecosystem, and in particular Facebook. We focus on Facebook because it has been a subject of greater scrutiny in courts, which has given us more public evidence from internal documents.

3.3.1. The Data Vortex: Adtech

As we will see in this section, as a reader browses content (on an app or on a website), multiple intermediaries are informed within milliseconds of the general interests of the reader, of additional information about the content on the page, and the ad slots available around that piece of content. A system of bidding then decides which ad to show, and that gets implemented within 100s of milliseconds.

This system is present in websites and in apps, and in fact some of those intermediaries are there precisely to link profiles on different devices together (e.g. recover that the person currently browsing a news app was previously shopping for shoes on their laptop).

We can see some glimpses of the sensors they install ("trackers") in Figure 12, which shows how Aamulehti used companies such as Google, Amazon, AT&T, PubMatic, Adform, Facebook, OpenX, Adobe, Krux, FreeWheel, Criteo, etc as intermediaries to facilitate that bidding, while Sari Tanus was using their app on her #digipower loan phone. In Figure 45, we compare the trackers across three different publishers in the Sanoma group: Aamulehti, Helsingin Sanomat and Ilta-Sanomat.

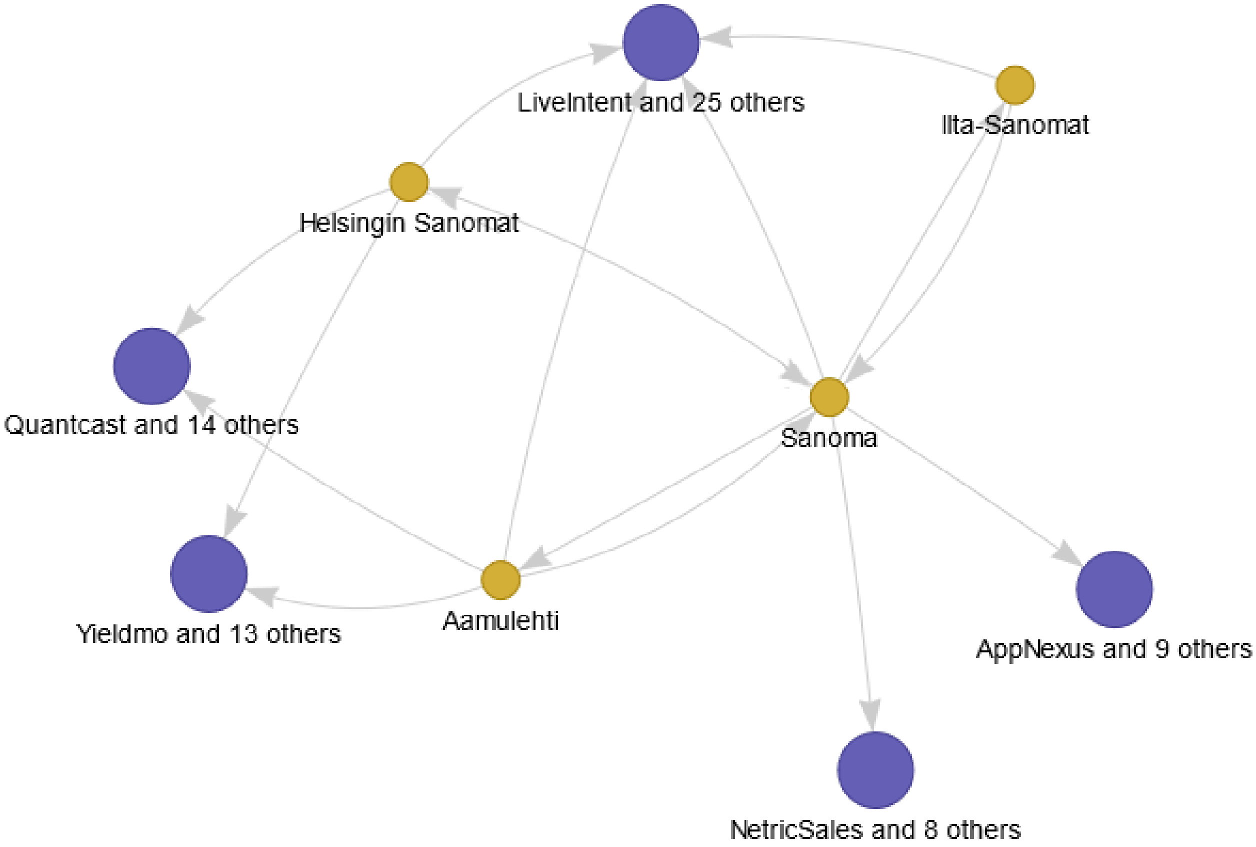


Figure 45: Flows of data we have uncovered for three media properties of the Sanoma group: Helsingin Sanomat, Aamulehti and Ilta-Sanomat. This has been evidenced either through

TrackerControl audits or by looking at the ads.txt[94] protocol files for each publication. The trackers are coalesced into one node for each distinct context, with a count. There might be biassing effects due to unequal use amongst participants.

There is nothing exceptional about Aamulehti or the Sanoma group in this; many newspapers use trackers in this way, but data gathered from Dan Koivulaakso's participation is interesting here. He decided to target different media companies (Telia/MTV3, Helsingin Sanomat and public broadcaster Yle), which showed that trackers were not a foregone conclusion[95] (see Figure 46), since only Adobe's were detected across all three, but especially since there was a vast majority of the trackers were present with Helsingin Sanomat[96].

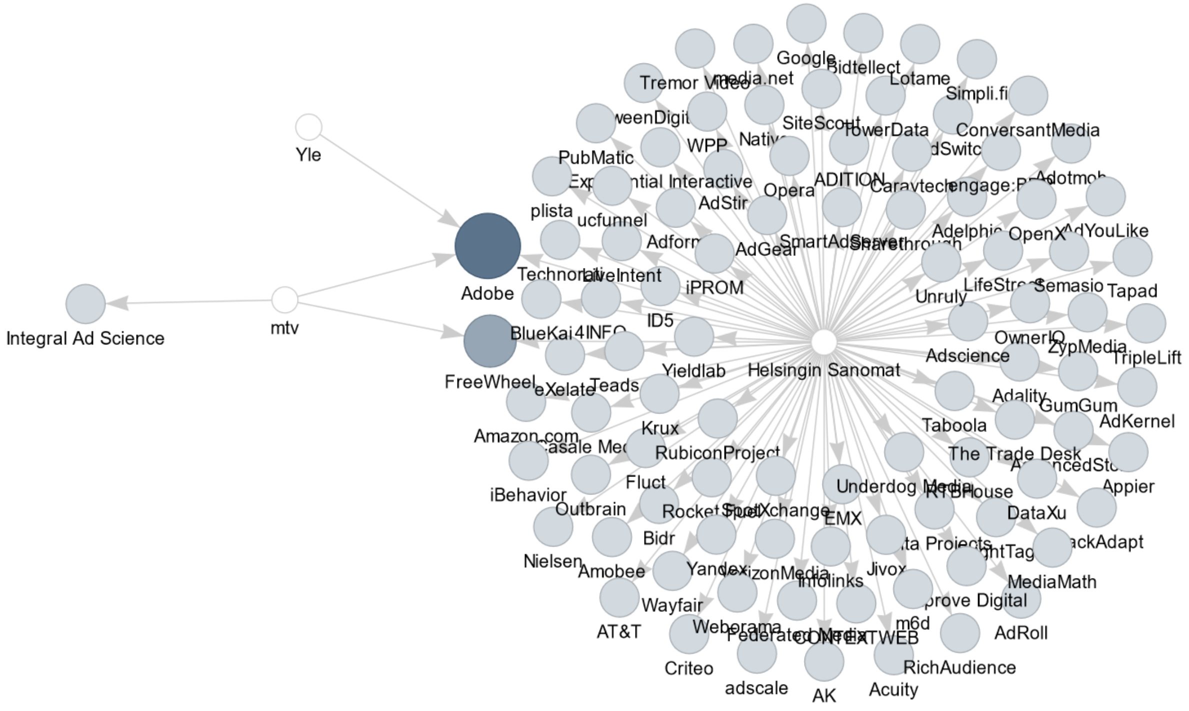


Figure 46: A technical audit of Dan Koivulaakso's loan phone showed that Sanoma Group use dozens of third party trackers, compared to minimal tracking by Yle and MTV3/Telia.

We also present an aggregate picture across participants in the study in Figure 47, in which we had to coalesce multiple trackers into single nodes due to the presence of numerous trackers. Observe that trackers are able to observe participants across websites.

[94] The ads.txt protocol is a protocol that allows publishers to broadcast their valid brokering intermediaries. See https://en.wikipedia.org/wiki/Ads.txt for an explanation, and https://www.hs.fi/ads.txt for an example listing.
[95] Some Yle apps include substantially more trackers though.
[96] This finding is subject to bias that would have been introduced through different usage he would have made of each app, for instance loading more pages for the HS app.

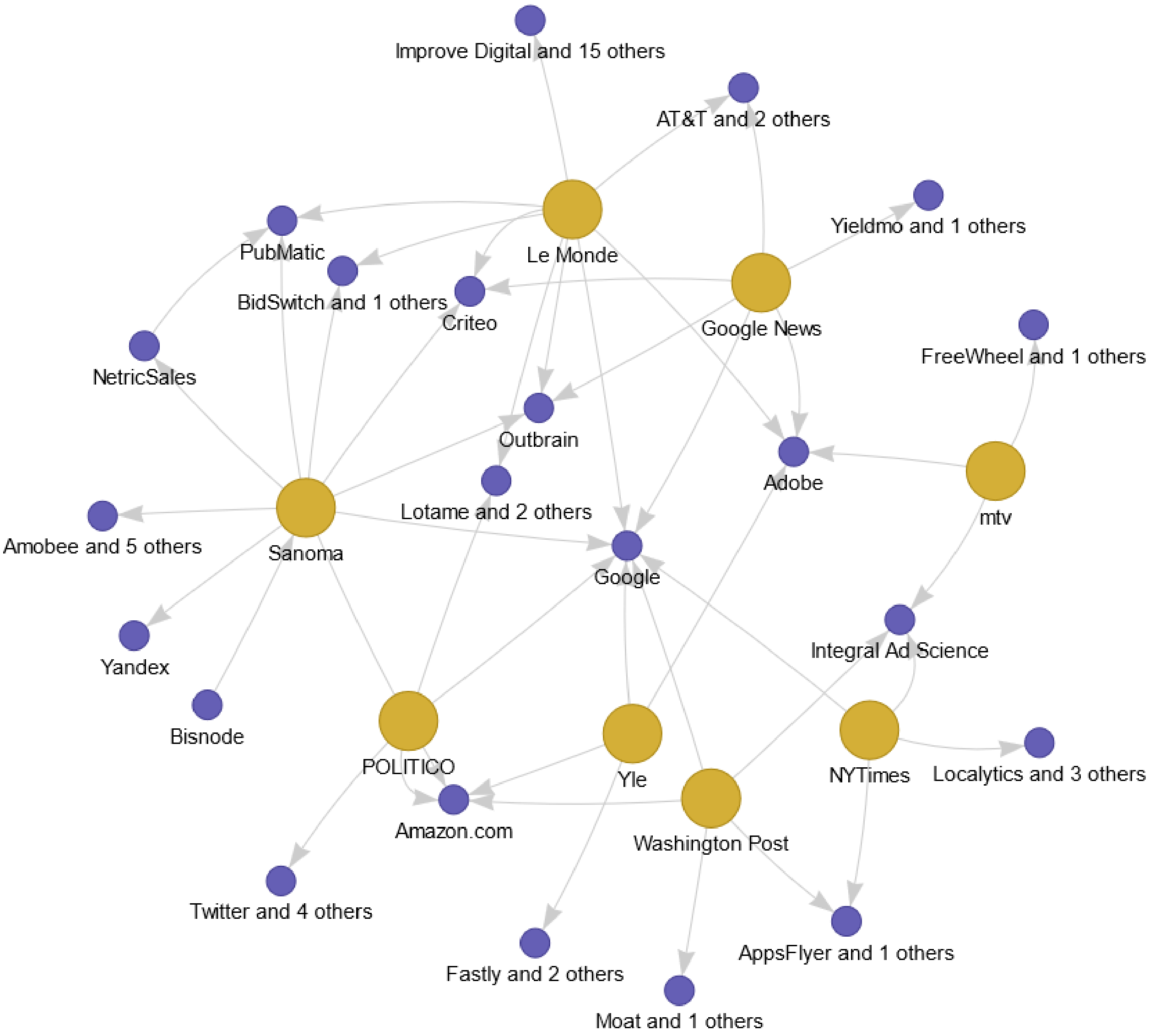


Figure 47: Aggregate view of tracker presence across publishers for the media apps used by the participants. There might be biassing effects due to unequal use of the apps by different participants.

The ecosystem as a whole is called the Lumascape, and has been linked to numerous negative consequences on individual readers, such as identity fraud, blackmail or social engineering attacks.

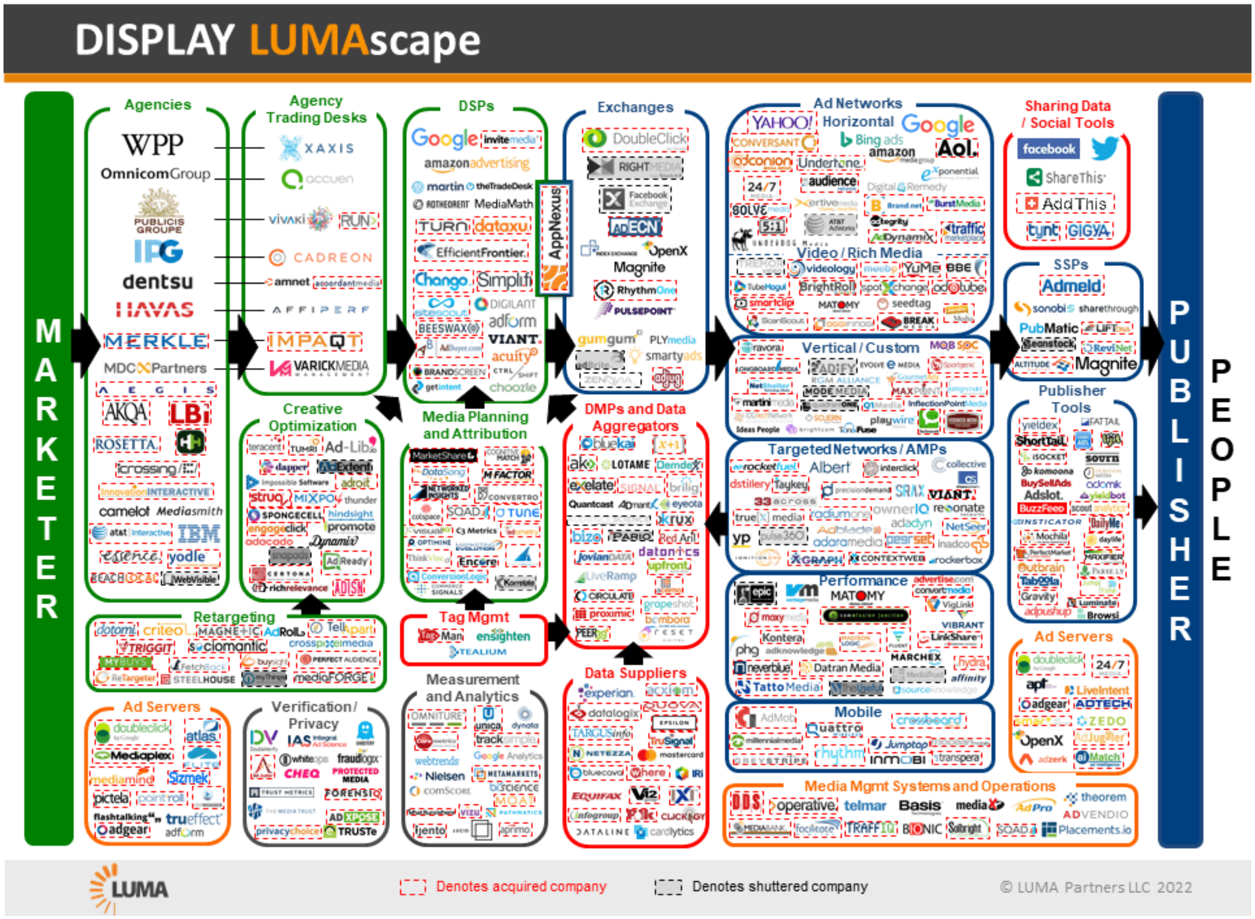


Figure 48: A (very simplified) visualisation of today's adtech ecosystem of trackers, data brokers and other third parties exchanging personal data and information about users[97]

This ecosystem can lead to numerous negative consequences beyond individual ones. It has been argued for instance that it was a national security risk[98], due to its influence potential. In a weird development, this ecosystem has also been highlighted as a security threat against the Catholic church while being used to out a gay priest[99].

We have found within the #digipower investigation some traces of trackers that are concerning. For instance we have found through TrackerControl that service provider Yandex (often introduced as the "Russian Google") was sent data in the following circumstances:

[97] Copyright LUMA Partners LLC, 2022, https://lumapartners.com/content/lumascapes/display-ad-tech-lumascape/

[98] Jessica Dawson. "Microtargeting as Information Warfare." *The Cyber Defense Review* 6, no. 1 (2021): 63-80, https://cyberdefensereview.army.mil/Portals/6/Documents/2021_winter_cdr/CDR_Winter_2021.pdf

[99] Journalists made the point that embedded trackers in apps could be used to spy on the use of gay dating apps by priests as a destabilization tactic against the Vatican, and then used that threat vector to themselves out a high profile bishop. See Joseph Cox, *The Inevitable Weaponization of App Data Is Here*, VICE Moterboard https://www.vice.com/en/article/pkbxp8/grindr-location-data-priest-weaponization-app

- when Sari Tanus used the Aamulehti app;
- when Atte Harjanne and Dan Koivulaakso used the Helsingin Sanomat app;
- when many participants used one of the web browsers installed on the loan phones (such as Chrome, Firefox or Samsung's browser).

The latter is concerning but since it occurs through a web browser, Tracker Control affords us less traceability on which website had been instrumented (equipped) with trackers, so we are not able to attribute the data leakage to Yandex to a specific source.

In each case, one can expect that pseudonyms tied to the specific device would have been exchanged with Yandex, along with information about the content visited. Yandex would have the ability to use these device identifiers to retrieve more information (such as geolocation) through data exchange with other ecosystem players. There are of course numerous American players and Chinese players (e.g. Tencent) involved in such tracking as well.

Sanoma Group (the media group behind both Aamulehti and Helsingin Sanomat) were asked for comment about the link to Yandex, and informed Sitra on 9th June, 2022 that Yandex Taxi had purchased a small amount of advertising from Sanoma programmatically via Google's ad purchasing tools, and that following the Russian invasion of Ukraine, Yandex systems have been prevented from interacting with Sanoma's systems since 3rd March, 2022.

News sites and apps are among the most prevalent in their use of third party trackers, because they want to monetize their audience. However this is a two-sided game. Potential advertisers also want to include those trackers on their website (such as an ecommerce site) because they need to know who they are targeting, and to match this up with data they hold about them (see for instance Gigantti's use of third parties in Figures 31 and 32). To illustrate this, consider Gigantti's response to Miapetra Kumpula-Natri and Sari Tanus' Subject Access Requests. To questions about cookies, they responded: "Joint controllers: Joint controlling applies for third party cookies on our webpage. Please see our cookie policy for detailed information on third party cookies." and included a personalised link to their cookie policy[100]. "**Joint controllers**" is a legal relationship defined in the GDPR's Article 26 that implies a joint responsibility for the processing operations undertaken by either of the two, but the article is quite unclear on the limitations, if any, that can be or should be put by one party upon the processing activities of the other party. Gigantti lists over 200 cookies in their privacy policy, which is a lot of liability to take on.

We found evidence (Figure 49) that Gigantti was using Yandex as a tracker as well. When informed of our preliminary findings, Gigantti stated:

"We do not send our own or our customers' data from Gigantti to Yandex.

[100] available directly at https://www.gigantti.fi/asiakaspalvelu/evasteet

Our advertising network partner Criteo's advertising platform has used a cookie that has not passed on any personal information. Criteo has used anonymous, non-personally identifiable pixels that have allowed advertising on a few Russian publishers' websites. These pixels have been removed at our request and by Criteo since the start of the war in Ukraine on March 8, 2022. All other functions related to Yandex have been removed from the Gigantti.fi website."
– Gigantti response dated April 5th, 2022.

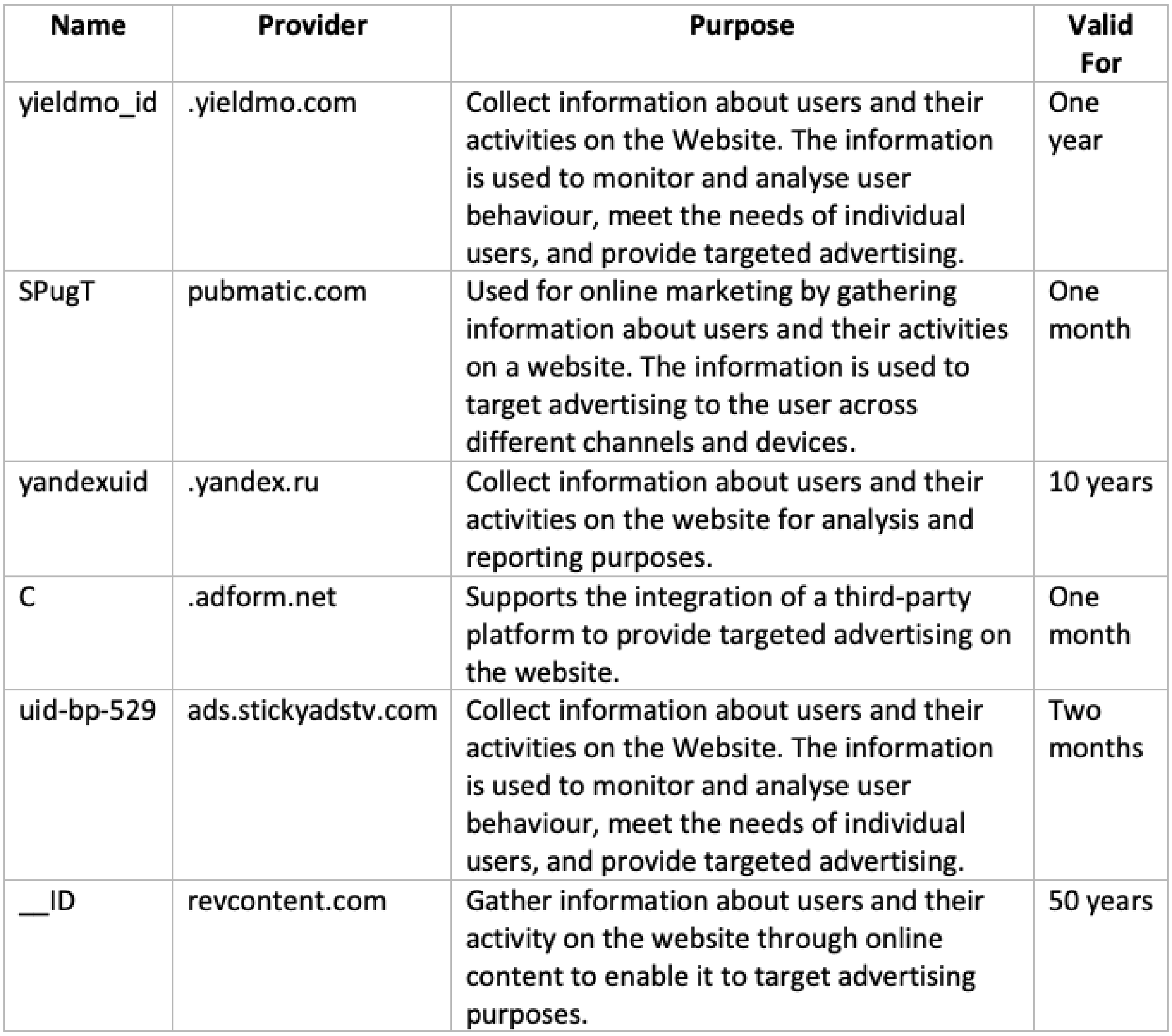
We congratulate Gigantti for taking proactive measures *before* being contacted by us in late March[101,102]. We do not think such a cookie is truly anonymous. Generally such cookies are pseudonymous - especially if Criteo, a known retargeter, is involved. Pseudonymous data, i.e. personal data referenced through an identifier, is still personal data in the GDPR. Based on services generally offered by Criteo, we interpret this response as follows: Gigantti (through Criteo) wants to be able to track past customers or visitors to their site when these customers visit Russian publisher's websites, in order to show them advertisements for Gigantti products that they have shown an interest in but not bought. Yandex has an agreement with Criteo to lift website visit data across a range of websites and visitors, which Criteo tries to match up with Gigantti customers. *Only when there is a match with a Gigantti customer*, Criteo bids and in case it wins a Gigantti ad flows back to the Russian publisher's visitor. This sends money from Gigantti to the publisher (and the intermediary Yandex), and also some personal data back (at minimum that this pseudonymous user is a Gigantti customer). What data gets sent to which intermediary is impossible to determine within the #digipower context, short of engaging in long exchanges with Gigantti[103].

We additionally observe that Gigantti's cookie policy includes a Vietnamese intermediary and a South Korean one, and that Gigantti also used mail.ru cookies. The mail.ru cookies were also removed.

[101] and remind the reader we focus on Gigantti because they have been more transparent in the first place, although this particular analysis only involves reading the privacy policy.

[102] For a comparison of their privacy policies, with changes on April 5th, see https://web.archive.org/web/diff/20220402055911/20220408174315/https://www.gigantti.fi/asiakaspalvelu/evasteet

[103] Article 26 of the GDPR states: *"Where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers. They shall in a transparent manner determine their respective responsibilities for compliance with the obligations under this Regulation, in particular as regards the exercising of the rights of the data subject and their respective duties to provide the information referred to in Articles 13 and 14, by means of an arrangement between them unless, and in so far as, the respective responsibilities of the controllers are determined by Union or Member State law to which the controllers are subject. [..] The essence of the arrangement [referred to in paragraph 1] shall be made available to the data subject."* In other words, if pushed Gigantti should provide more information about what is exchanged with whom.



Name	Provider	Purpose	Valid For
yieldmo_id	.yieldmo.com	Collect information about users and their activities on the Website. The information is used to monitor and analyse user behaviour, meet the needs of individual users, and provide targeted advertising.	One year
SPugT	pubmatic.com	Used for online marketing by gathering information about users and their activities on a website. The information is used to target advertising to the user across different channels and devices.	One month
yandexuid	.yandex.ru	Collect information about users and their activities on the website for analysis and reporting purposes.	10 years
C	.adform.net	Supports the integration of a third-party platform to provide targeted advertising on the website.	One month
uid-bp-529	ads.stickyadstv.com	Collect information about users and their activities on the Website. The information is used to monitor and analyse user behaviour, meet the needs of individual users, and provide targeted advertising.	Two months
__ID	revcontent.com	Gather information about users and their activity on the website through online content to enable it to target advertising purposes.	50 years

Figure 49: In Gigantti's SAR responses to Miapetra Kumpula-Natri and Sari Tanus, a link is provided to a part of Gigantti's website where joint controller data sharing relationships are declared. The table above shows a sample. The list[104] seems to be updated regularly. At the time of writing, over 230 different companies were listed.

3.3.2. Sitting at the Top: the Example of Facebook

So far in this chapter, we have leveraged the example of FarmVille as a reference scenario, to compare the capabilities exerted over individuals in different contexts (physical, digitised version of the physical, purely content-driven). We have considered FarmVille as a closed world, a convenient allegory we could use to highlight particular practices and situational aspects. However, some online actors try very much not to behave like a closed world, and instead to

[104] https://www.gigantti.fi/asiakaspalvelu/evasteet

constantly expand into what users would consider other parts of their life. In their paper *Getting Under Your Skin(s): A Legal-Ethical Exploration of Fortnite's Transformation Into a Content Delivery Platform and Its Manipulative Potential*[105], Sax and Ausloos argue that the capabilities within games are also fast-evolving. They pick as their example the game Fortnite, where indeed the freemium model of the game requires, just as in FarmVille, that users be engaged and influenced to engage in micro-transactions. This compares to social networks that require us to keep using the app/site for a long time, so that we might click on ads once in a while. However, they also observe that Fortnite is turning into something more than a game, into what they see as a content delivery platform itself, "where the game itself becomes a means to deliver other non-game related content and services to users by integrating them natively into the engaging video game experience offered by Fortnite". These offerings seem to be expanding fast, for instance including concerts by real-life artists, or special movie premieres. In a sense, the game is starting to engulf the outside world into its orbit, and the interactions are mediated through game mechanics.

We have seen in the previous section that many players are involved in the advertising technology business, giving the impression that whatever control or influence there was, it would be exerted by a large variety of actors. Our goal in this section is to show that there is actually a lot of consolidation in this control and to hint at why this consolidation took place. We will also show that those actors continue to engage in outward-facing dynamics, always trying to expand further and digitise more around them.

An example of this was observed where Gigantti sends Facebook brick-and-mortar purchase data (see Figures 31 & 32). In fact, in order to reduce the cognitive complexity for the reader, we want to focus on Facebook as the prototypical example of such an ever-expanding data collection company, and on some of the tools they use that incentivize others to enter its orbit: Facebook Custom Audiences and Facebook Pixels. Other advertising companies such as Google or Amazon offer similar services[106].

Facebook has additional relevance to the public interest given that it is the subject of multiple ongoing US lawsuits (customer protection, antitrust, privacy), which lead to judicial discovery

[105] Marijn Sax & Jef Ausloos, Getting Under Your Skin(s): A Legal-Ethical Exploration of Fortnite's Transformation Into a Content Delivery Platform and Its Manipulative Potential, Interactive Entertainment Law Review, jan 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3764489

[106] and indeed in the previous section we saw that Twitter did offer similar capabilities, and this is part of the move described earlier for the *Washington Post* through privileged partnerships with Amazon.

and the publication of internal documents[107], showing how Facebook employees conceive of their relationships with other companies.

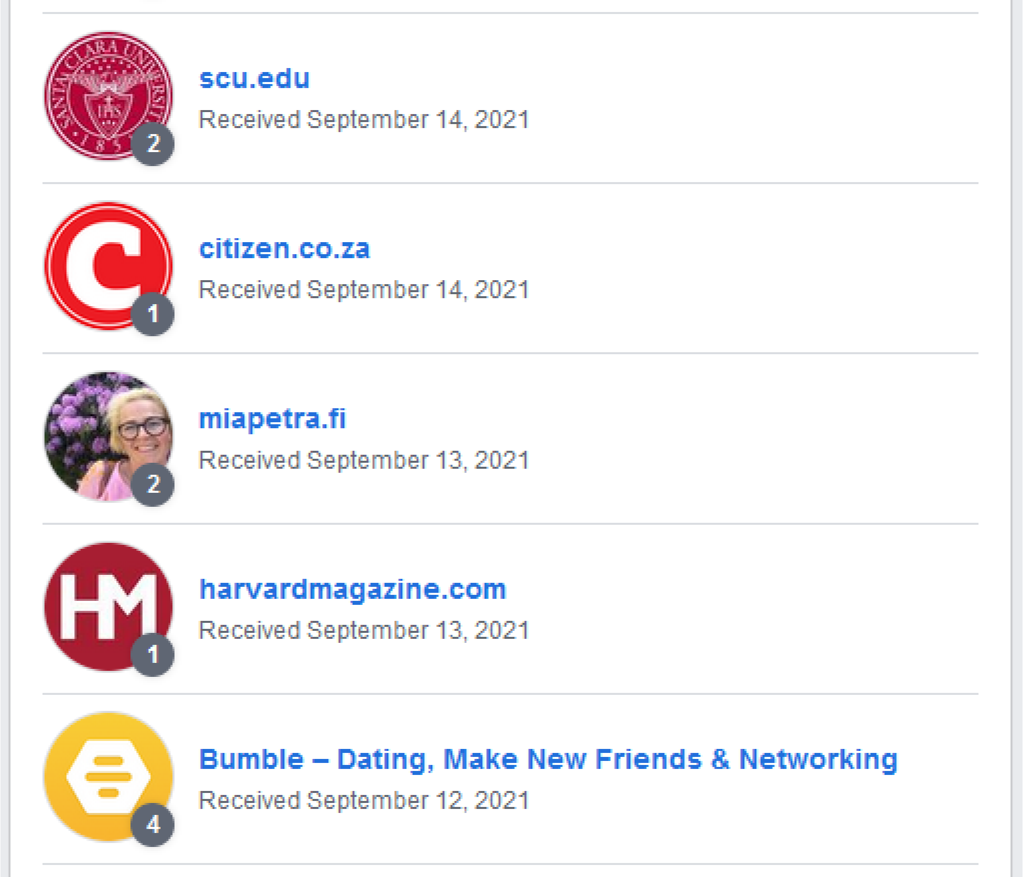
The first tool we look at is **Facebook Pixel[108]**. Facebook Pixel embeds a 1x1 pixel on the website of the webmaster implementing the Pixel, which will require any browser accessing that website to fetch data from Facebook. Once this is done, the Pixel is best conceptualised as a sensor controlled by Facebook. Every time the page is loaded, Facebook will then fingerprint[109] the browser (i.e. construct a unique identifier) and associate a bit more data with that fingerprint. Through triangulation between multiple websites, combined with additional information collected through the direct use of Facebook products such as their Facebook and Instagram apps or facebook.com, this will allow Facebook to build a transversal profile of the user behind that browser, regardless of which device or browser the individual uses. Tracking pixels are not unique to Facebook, but their efficacy is greater for large companies like Facebook because of Facebook's ability to combine Pixel data with their other sources of data. Additionally, Pixels, like cookies under the same-site origin policy, are more effective when part of a vast network across sites, and the incentives for businesses to adopt Facebook Pixels are great[110] because of Facebook's powerful role in online profiling.

Among our participants, we found evidence of this ecosystem dependence at the level of a small website owner. Participant Miapetra Kumpula-Natri's own website indeed includes a Facebook Pixel (Figure 50). The usual reason to use this service in such a situation (she is not running Facebook ads) is to get more precise demographics on website visitors, which Facebook is able to deliver because they can match visitors to a specific website to their known identity profile[111].

[107] Lawsuit. Court Filing: MAXIMILIAN KLEIN, et al. Plaintiffs, v. FACEBOOK, INC., Defendant., Court:United States District Court, Northern District of California, legal reference20-CV-08570-LHK (N.D. Cal. Jul. 20, 2021), Document 244-3, https://www.courtlistener.com/docket/18714274/244/3/klein-v-meta-platforms-inc/

Lawsuit. Court Filing: In re: Facebook, Inc. Consumer Privacy User Profile Litig., Court: United States District Court, Northern District of California, legal reference 18-md-02843-VCJSC) (N.D. Cal. Dec. 30, 2021), Document 491, https://cand.uscourts.gov/wp-content/uploads/cases-of-interest/in-re-Facebook-consumer-privacy-VC/Second-Amended-Consolidated-Complaint-Dkt-491.pdf

[108] https://www.facebook.com/business/learn/facebook-ads-pixel
[109] https://coveryourtracks.eff.org/learn
[110] This orchestrating behaviour is explored more in the narrative report, Section 3.6, lever 4.
[111] In this regard, the Pixel is very similar to Google Analytics.



scu.edu
Received September 14, 2021

citizen.co.za
Received September 14, 2021

miapetra.fi
Received September 13, 2021

harvardmagazine.com
Received September 13, 2021

Bumble – Dating, Make New Friends & Networking
Received September 12, 2021

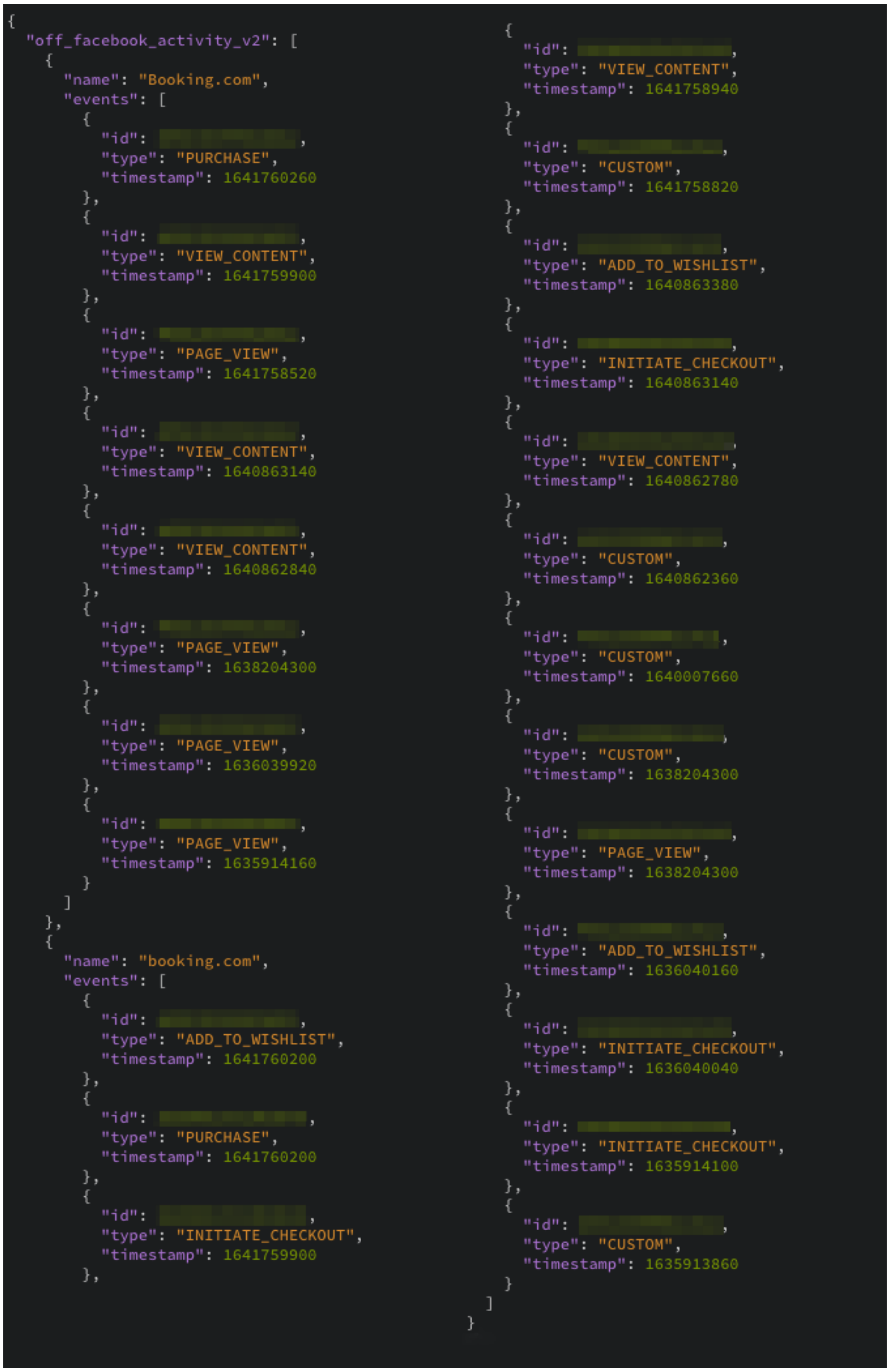
Figure 50: Miapetra Kumpula-Natri's website miapetra.fi shows up in investigator Paul-Oliver Dehaye's Facebook Offsite Activity Data, showing that her website's use of Facebook Pixel contributed to collection by Facebook of personal data.

We can see that the Facebook Pixel is a symbiosis between the website and Facebook. Facebook gets information about the website's visitors, and the website is better able to benefit from Facebook's profiling capabilities.

Of course, that data can also serve direct advertising benefits, as illustrated quite vividly in Facebook's own documentation[112]:

"Say you're an online florist that wants to reach people similar to those that made purchases on your website. Now you can use data from your Facebook pixels (Facebook Conversion Pixel or the Custom Audiences for Websites pixel) to reach people who are most similar to people who previously made purchases on your website."

[112] https://www.facebook.com/business/news/Expanded-Capabilities-for-Lookalike-Audiences

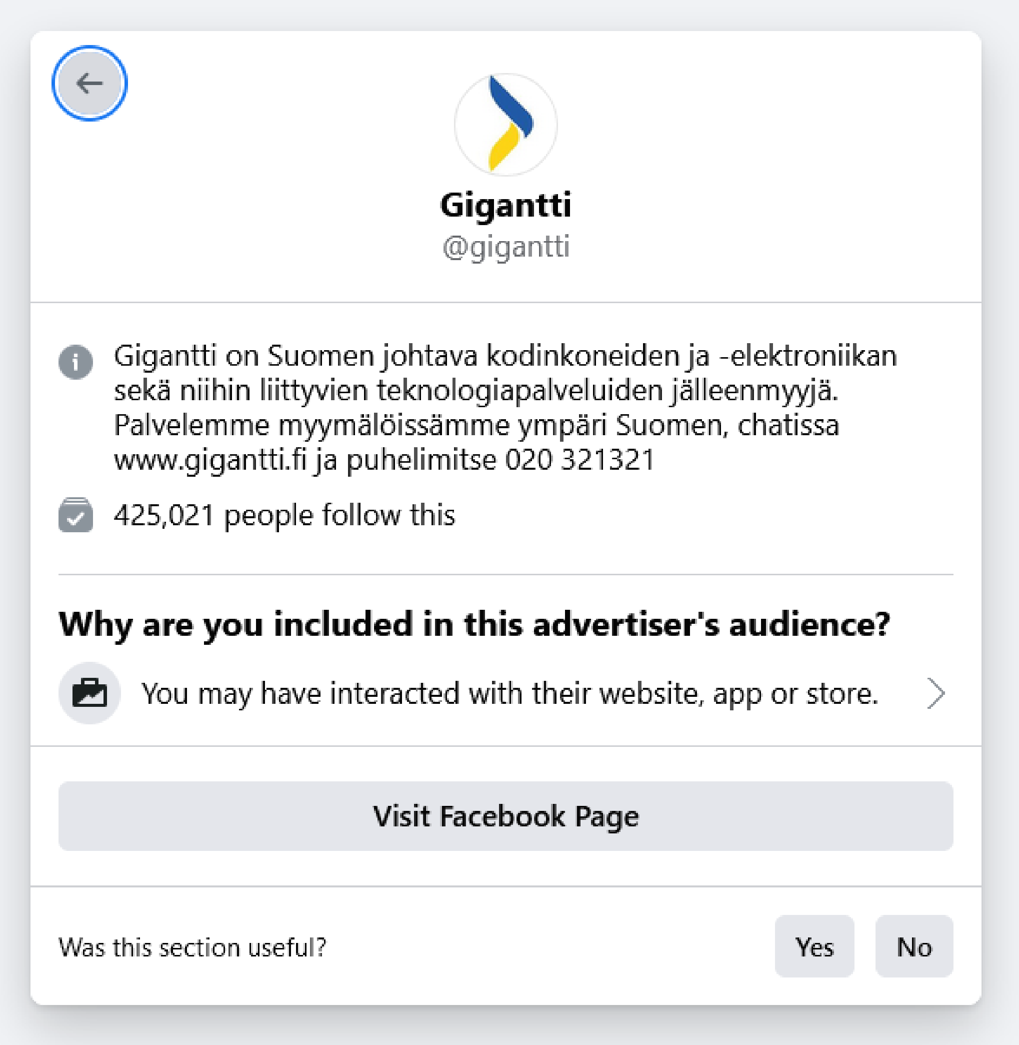


```
{
  "off_facebook_activity_v2": [
    {
      "name": "Booking.com",
      "events": [
        {
          "id": [redacted],
          "type": "PURCHASE",
          "timestamp": 1641760260
        },
        {
          "id": [redacted],
          "type": "VIEW_CONTENT",
          "timestamp": 1641759900
        },
        {
          "id": [redacted],
          "type": "PAGE_VIEW",
          "timestamp": 1641758520
        },
        {
          "id": [redacted],
          "type": "VIEW_CONTENT",
          "timestamp": 1640863140
        },
        {
          "id": [redacted],
          "type": "VIEW_CONTENT",
          "timestamp": 1640862840
        },
        {
          "id": [redacted],
          "type": "PAGE_VIEW",
          "timestamp": 1638204300
        },
        {
          "id": [redacted],
          "type": "PAGE_VIEW",
          "timestamp": 1636039920
        },
        {
          "id": [redacted],
          "type": "PAGE_VIEW",
          "timestamp": 1635914160
        }
      ]
    },
    {
      "name": "booking.com",
      "events": [
        {
          "id": [redacted],
          "type": "ADD_TO_WISHLIST",
          "timestamp": 1641760200
        },
        {
          "id": [redacted],
          "type": "PURCHASE",
          "timestamp": 1641760200
        },
        {
          "id": [redacted],
          "type": "INITIATE_CHECKOUT",
          "timestamp": 1641759900
        },
        {
          "id": [redacted],
          "type": "VIEW_CONTENT",
          "timestamp": 1641758940
        },
        {
          "id": [redacted],
          "type": "CUSTOM",
          "timestamp": 1641758820
        },
        {
          "id": [redacted],
          "type": "ADD_TO_WISHLIST",
          "timestamp": 1640863380
        },
        {
          "id": [redacted],
          "type": "INITIATE_CHECKOUT",
          "timestamp": 1640863140
        },
        {
          "id": [redacted],
          "type": "VIEW_CONTENT",
          "timestamp": 1640862780
        },
        {
          "id": [redacted],
          "type": "CUSTOM",
          "timestamp": 1640862360
        },
        {
          "id": [redacted],
          "type": "CUSTOM",
          "timestamp": 1640007660
        },
        {
          "id": [redacted],
          "type": "CUSTOM",
          "timestamp": 1638204300
        },
        {
          "id": [redacted],
          "type": "PAGE_VIEW",
          "timestamp": 1638204300
        },
        {
          "id": [redacted],
          "type": "ADD_TO_WISHLIST",
          "timestamp": 1636040160
        },
        {
          "id": [redacted],
          "type": "INITIATE_CHECKOUT",
          "timestamp": 1636040040
        },
        {
          "id": [redacted],
          "type": "INITIATE_CHECKOUT",
          "timestamp": 1635914100
        },
        {
          "id": [redacted],
          "type": "CUSTOM",
          "timestamp": 1635913860
        }
      ]
    }
```

Figure 51: "Off-site Facebook Activity" data from Christian D'Cunha's downloaded Facebook data shows that Booking.com share details of his purchase and wishlist (i.e. intent) activity on their website and/or app (and most likely more detailed information we cannot see). This information collecting is enabled via the placement of a Facebook Pixel onto the Booking.com website by Booking's developers.

Another example of the use of Facebook Pixel comes through Christian D'Cunha's Facebook data, which reveals the very granular view that Facebook acquires of his behaviour on the booking.com site (Figure 51): as he browsed that site, granular events were logged and passed to Facebook, such as PURCHASE, VIEW_CONTENT, PAGE_VIEW, ADD_TO_WISHLIST, INITIATE_CHECKOUT, or CUSTOM. This was by no means exceptional, as dozens of other websites sent similar data (and in particular – across participants – travel services such as KLM, TUI, AirFrance, LetsTravel, TripAdvisor, Trainline, Hyatt, SBB, Uber, SNCF, AirBNB, etc).

Investigator Paul-Olivier Dehaye has also found evidence of Helsingin Sanomat or Gigantti embedding trackers on their website sending *pseudonymous* data to Facebook (see Figure 52).



Gigantti
@gigantti

Gigantti on Suomen johtava kodinkoneiden ja -elektroniikan sekä niihin liittyvien teknologiapalveluiden jälleenmyyjä. Palvelemme myymälöissämme ympäri Suomen, chatissa www.gigantti.fi ja puhelimitse 020 321321

425,021 people follow this

Why are you included in this advertiser's audience?

You may have interacted with their website, app or store.

Figure 52: Disclosure by Facebook of the use by Gigantti of a Facebook Pixel for Paul-Olivier Dehaye's account (for instance possibly as he was browsing the privacy policy of Gigantti, while preparing this report).

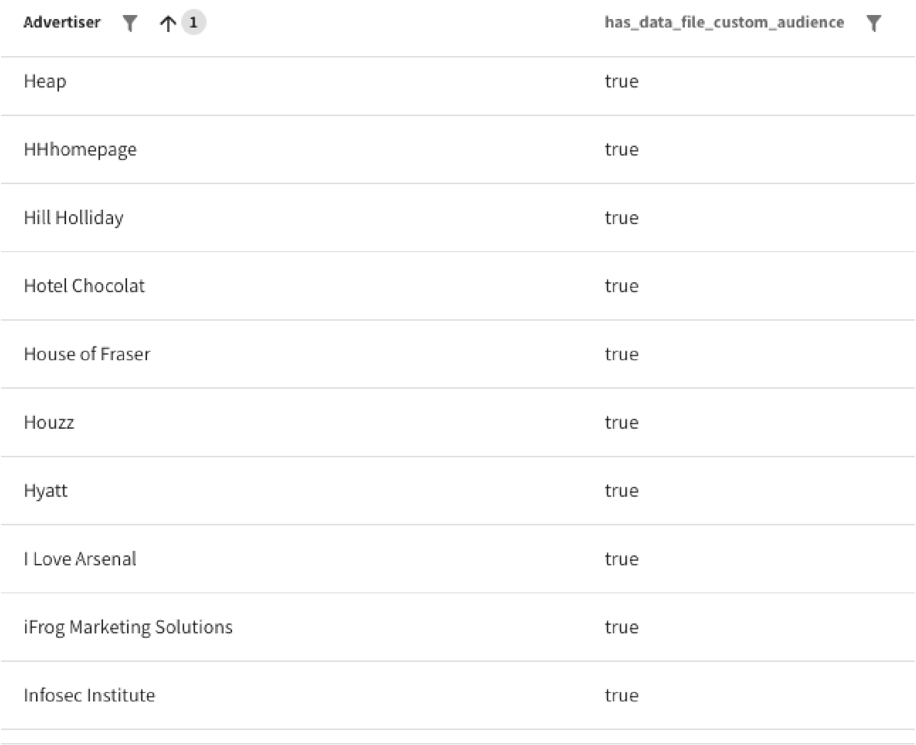
This idea of embedding a sensor on someone else's area of presumed exclusive control (a practice which users are largely unaware of, incidentally) is also present within apps, not just websites. Consider for instance Figure 12, which was the TrackerControl view offered on Sari Tanus' use of the Aamulehti app. This includes a mention of Facebook as a tracker (and many others), because Facebook has developed a Software Development Kit (SDK), i.e. a toolbox that developers like Aamulehti can freely integrate to their app but which ends up communicating device-identifying information to Facebook. Unlike the florist in Facebook's documentation example, Aamulehti does not intend to buy ads, they want to *show* ads. These ads could be for florists, for political candidates or parties (where allowed) or for other advertisers who will better target readers within the Aamulehti app thanks to the multitude of signals collected. This shows Facebook gaining a stranglehold on the entire ecosystem, since Facebook is now a gateway for others' core business activities.

However Pixels are only one way to specify a **Facebook Custom Audience**[113], which is a group of Facebook profiles that an advertiser has identified as desired targets. This could be for instance the list of "All the Facebook users who have visited my website", for which the advertiser does not have profile information but where Facebook can nevertheless associate an audience through that website's use of a Facebook Pixel. Custom Audiences can be built in multiple other ways, such as importing data from a company's own store, or from a third party or a **Matched Custom Audience**. A Matched Custom Audience is a custom audience built by matching from an extensive list actually presented by the advertiser, including (partial) information such as email, phone number, physical address, device identifier, etc. which Facebook then compares against its own comprehensive bank of user profiles.[114] These audiences can be expanded through Facebook to identify other similar people in a process leveraging machine learning called Lookalike Audiences.

In Christian D'Cunha's Facebook data one file shows a list of advertisers who acquired Custom Audience functionality from Facebook in order to target him (Figure 53).

[113] https://www.facebook.com/business/help/744354708981227?id=2469097953376494

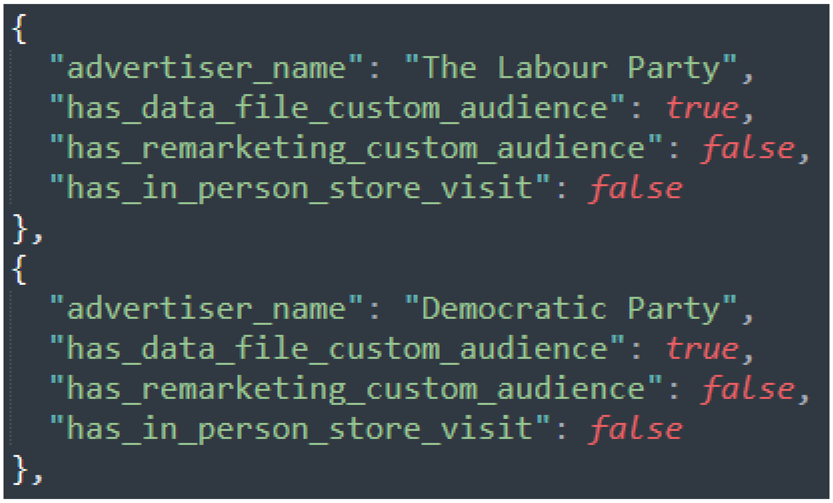
[114] Facebook includes an intermediate step of hashing, which is purely cosmetic given the end goal of matching with individual profiles, but whose only effective purpose is to obscure Facebook's and the advertiser's liabilities under data protection laws that recognize pseudonymous data as still personal data (such as the GDPR).



Advertiser	has_data_file_custom_audience
Heap	true
HHhomepage	true
Hill Holliday	true
Hotel Chocolat	true
House of Fraser	true
Houzz	true
Hyatt	true
I Love Arsenal	true
iFrog Marketing Solutions	true
Infosec Institute	true

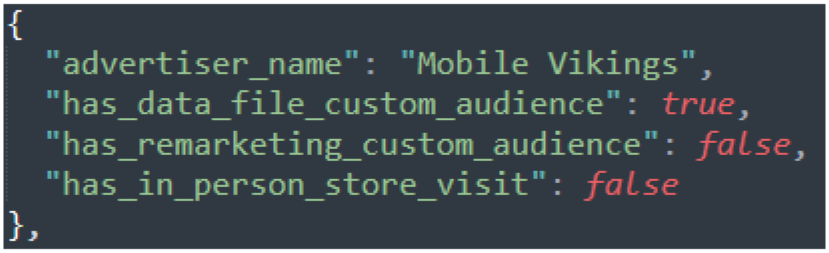
Figure 53: An extract from an example seen in Facebook SAR-returned data for Christian D'Cunha, listing companies that have acquired Custom Audience functionality from Facebook in order to target him

These might look innocuous in their vast majority: it includes "Arsenal", "I Love Arsenal", "ebay.co.uk", "Netflix", "The New York Times", and "POLITICO Europe". A closer look does raise some alarm bells however (beyond listing "Fulham FC Official"). For instance in Figure 54, we see that "The Labour Party" and the "Democratic Party" have listed him in one of their Custom Audiences. In Figure 55, we see that his mobile phone provider Mobile Vikings has done the same.



```
{
  "advertiser_name": "The Labour Party",
  "has_data_file_custom_audience": true,
  "has_remarketing_custom_audience": false,
  "has_in_person_store_visit": false
},
{
  "advertiser_name": "Democratic Party",
  "has_data_file_custom_audience": true,
  "has_remarketing_custom_audience": false,
  "has_in_person_store_visit": false
},
```

Figure 54: An extract from Christian D'Cunha's downloaded Facebook data, showing that the Labour Party (UK) and the Democratic Party (USA) have added him to a Facebook Custom Audience (edited so the two entries are contiguous).



```
{
  "advertiser_name": "Mobile Vikings",
  "has_data_file_custom_audience": true,
  "has_remarketing_custom_audience": false,
  "has_in_person_store_visit": false
},
```

Figure 55: An extract from Christian D'Cunha's downloaded Facebook data, showing that his mobile phone provider Mobile Vikings have added him to a Facebook Custom Audience.

We have already seen signs of Facebook's Custom Audiences tool in Gigantti's response to Miapetra Kumpula-Natri and Sari Tanus (Figure 32) informing them data about brick-and-mortar store purchases could be sent to Facebook: *"Third parties we have shared personal information with: Facebook Inc. and Google Inc. The data has been shared for the purpose of analyzing the effect of marketing campaigns, and Google and Facebook are considered to be processors when processing data for this purpose."* Note that Gigantti does not claim this data had been shared for the purpose of analysing *their* marketing campaigns. Knowingly or unknowingly, they are contributing to everyone's targeting accuracy, and first and foremost Facebook's[115].

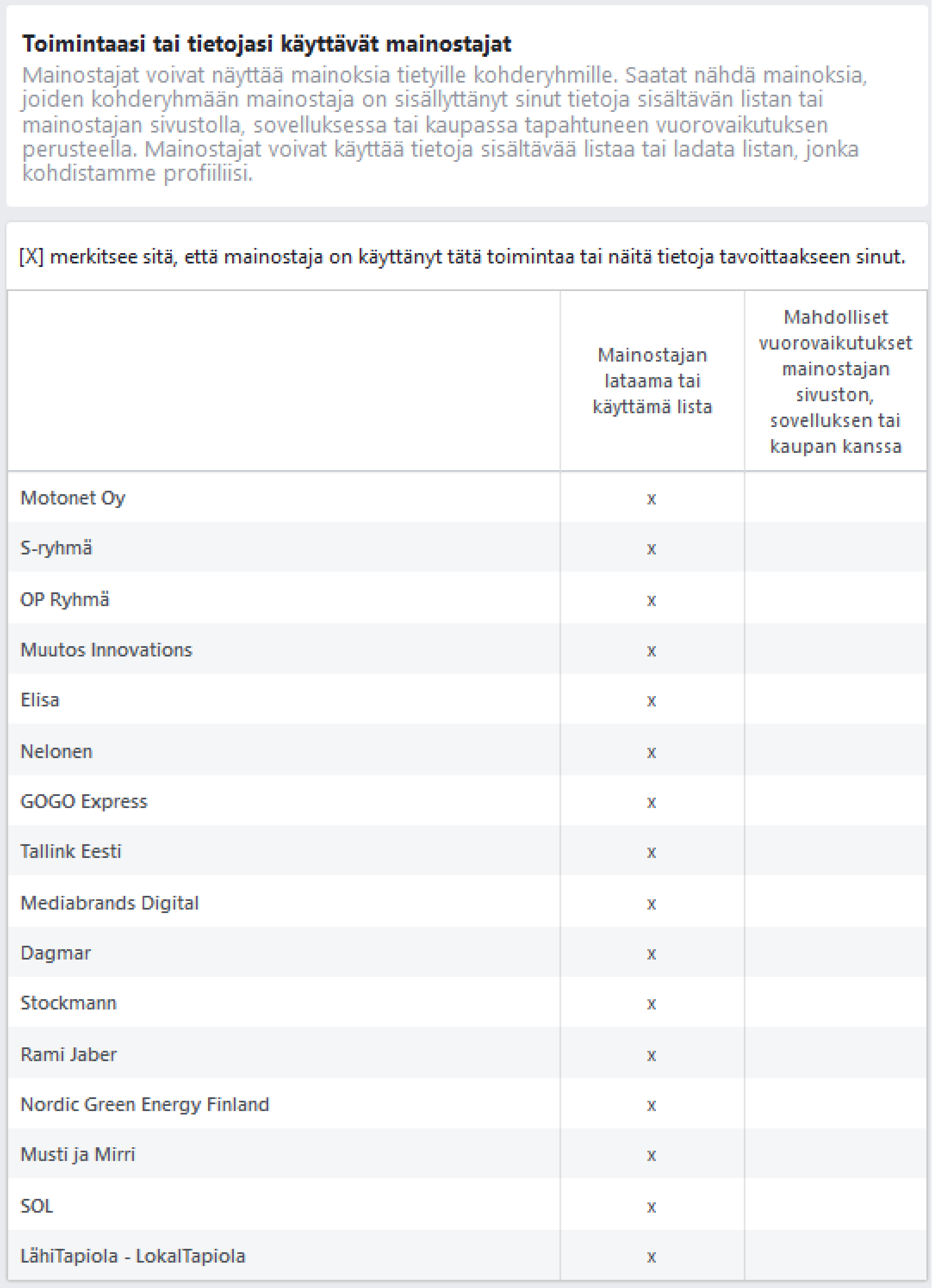
[115] In a response to those particular findings, Gigantti has stated: "As stated in the Gigantti.com Privacy Statement, we send anonymous in-store receipt information from our club members to Google and Facebook. All information is encrypted before it is transmitted, and we do not send any data in an identifiable form to third parties. Facebook and Google will destroy the data transmitted to them 14 days after receipt. We use club member data to analyze the effectiveness of digital advertising and not for any other purpose. Upon joining, our club members consent to the use of their information in accordance with Gigantti's Privacy Statement."

We believe this data is *not* anonymized but instead pseudonymized, and as such still personal data. Indeed the whole point would be for Facebook to match those hashed identifiers with its own collection.

All of those actors indeed intend to use Facebook's advertising capabilities to target their ads. With Facebook Like buttons, sign ins, and pixels reporting data back to a centralised source that also has a method of targeting ads based on extrapolated audiences, other companies and organisations are dependent on their infrastructure and left with few options but to continue paying for ads. From that position, once again, Facebook is able to observe the user's behaviour across the Internet and from multiple devices. Data is scattered across an ecosystem of actors and platforms, all benefiting from the breadth and spread and multiplicity of ways to see and understand people - but with the most capability to convert (or enable others to convert) and to attribute conversions - being held by the infrastructure providers and data brokers, those with the most or broadest data.

This reach even extends across platforms: indeed Sari Tanus's list of Instagram Custom Audiences (Figure 56) is targetable on both Facebook and Instagram.

The fact that the information is "encrypted while transmitted" is irrelevant, since this is proper IT practice and the recipient (Facebook) has the decryption key.



Toimintaasi tai tietojasi käyttävät mainostajat

Mainostajat voivat näyttää mainoksia tietyille kohderyhmille. Saatat nähdä mainoksia, joiden kohderyhmään mainostaja on sisällyttänyt sinut tietoja sisältävän listan tai mainostajan sivustolla, sovelluksessa tai kaupassa tapahtuneen vuorovaikutuksen perusteella. Mainostajat voivat käyttää tietoja sisältävää listaa tai ladata listan, jonka kohdistamme profiiliisi.

[X] merkitsee sitä, että mainostaja on käyttänyt tätä toimintaa tai näitä tietoja tavoittaakseen sinut.

	Mainostajan lataama tai käyttämä lista	Mahdolliset vuorovaikutukset mainostajan sivuston, sovelluksen tai kaupan kanssa
Motonet Oy	x	
S-ryhmä	x	
OP Ryhmä	x	
Muutos Innovations	x	
Elisa	x	
Nelonen	x	
GOGO Express	x	
Tallink Eesti	x	
Mediabrands Digital	x	
Dagmar	x	
Stockmann	x	
Rami Jaber	x	
Nordic Green Energy Finland	x	
Musti ja Mirri	x	
SOL	x	
LähiTapiola - LokalTapiola	x	

Figure 56: Sari Tanus' Instagram data, showing a list of advertisers with her information.

Another very interesting angle seen from Christian D'Cunha's data is that the following actors have also included him in their Custom Audiences: Amobee EMEA, Klaviyo, Clearbit, Adsmurai Spain, adverity, Fifty.io, Aidata.me, Ramp, ClearScore, LiveRamp, AdRoll, **Microsoft Customer Insights Center**, **Experian Marketing Services - Audiences**, TargetSmart, Acxiom, **Neustar FB Syndication**, Adobe, and Experian. As is more obvious from the name of some of those (bold), these are **data brokers**. They are using a functionality little known to the general public, enabling advertisers to share Custom Audience Lists[116]. This functionality allows data brokers to build audiences on Facebook and then monetize the ability to target them there[117], with little transparency[118]. This is an exceptionally powerful position to be in, since Facebook is able to intermediate the business of data brokers themselves, and not only profit financially from this but also learn more accurately similarities between profiles – thanks to the bread-and-butter operations of the data brokers themselves!

We can see that Facebook has managed to engineer a system that looks at first glance like it is a win for all stakeholders. However:

- it is debatable if an individual using Facebook wins here, given the multitude of consequences outlined in previous sections;
- the data of individuals not using Facebook also gets uploaded by all of those actors, enabling Facebook to build **shadow profiles**[119];
- the relation can sometimes be adversarial, as outlined in numerous ongoing lawsuits against Facebook, resting on a variety of legal arguments (consumer - i.e. advertiser - protection, antitrust, etc.)

We now illustrate this last point. In a recent antitrust lawsuit against Facebook, some internal emails were revealed that showed internal efforts to exploit the data shared with them by companies, in order to build tools that will compete directly with those companies, while keeping those companies' ad spending on the Facebook platform. In 2015, a product manager with experience at ebay reported the following to Mark Zuckerberg[120] :

[116] How to Share Custom Audience Lists, Facebook Business https://www.facebook.com/business/help/499290663823687?id=2469097953376494
[117] or to use them as components in the building of a new custom audience, by doing basic arithmetic on the audiences ("give me all profiles that are in my customer list AND in this list obtained through a data broker")
[118] It is noteworthy that for a little while after the introduction of the GDPR, Facebook did disclose who reused those brokered lists. They have however backpedaled on that transparency.
[119] to be concrete: Facebook gets to see what a Gigantti club member buys, and then match data with some data broker data. It just has to delete the raw data within 14 days to honor the contractual obligations Gigantti has mentioned.
[120] Lawsuit. Court Filing: MAXIMILIAN KLEIN, et al. Plaintiffs, v. FACEBOOK, INC., Defendant., Court:United States District Court, Northern District of California, legal reference20-CV-08570-LHK (N.D. Cal. Jul. 20, 2021), Document 244-3, https://www.courtlistener.com/docket/18714274/244/3/klein-v-meta-platforms-inc/

> *While we are building Marketplace for the long term, if we are not careful, we can have a short term negative impact on the ads business before we build out sustainable value. Several large advertisers are marketplaces and multi-channel retailers who may find our launch threatening to the extent that they may decide to pull ad spend or investment in key strategic ad products (e.g., dynamic product ads). The Facebook marketplace is good for partners who themselves are not marketplaces but clear messaging and value exchange will be needed to help them understand our intentions and value proposition. This situation is particularly risky during Q4 holiday season.*

This quote demonstrates a concerted effort to undermine their customers while keeping them paying and dependent on their ecosystem for as long as possible. This is the power of controlling the data infrastructure.

3.4. PARTICIPANTS CHASING THEIR DATA

In this case study, we look at what individuals can do to **see and understand** the complex hybrid physical and digital **world they inhabit**. The first step to taking action is being better informed. Through our participant experiences we highlight how the GDPR can help give us a view of what goes on behind those murky, ever-changing feeds and ads that make up the infoscape, but also how it falls short of giving us **the full transparency we need** to properly become autonomous free-thinking and unconstrained agents as we exist in the digital world.

The #digipower investigation explored the data ecosystem from the standpoint of individuals having a practical, grounded experience of accessing, making sense of and interpreting their data. In this section, we draw out some patterns and observations from the participants' experiences that give a picture of how effectively people are able to access their data, the usefulness of this access, and in particular how effectively GDPR access requests are handled by digital service providers.

3.4.1. Data Access is Hard Work, Rarely Meets Expectations, and May Not Succeed

In general, the experience of accessing one's own data was a time-consuming and sometimes quite difficult task for participants. Even with the coaches doing everything they could to make the process easy, such as by providing detailed instructions and templates and providing advice on how to respond to emails received, it took a lot of effort from participants to see their data requests through. This was exacerbated by the nature of our participant sample, being as there are, high-profile, busy people with extremely packed agendas and not a lot of time. In some cases, the procedures were so hard or time-consuming that target service providers had to be dropped to reduce workload. Several participants could not manage the workload and had to withdraw their participation in the data access part of the study.

Typically the process for sending a subject access request began with the sending of a template email, drafted by hestia.ai, to the Data Protection Officer or privacy e-mail address identified in each company's privacy policy. This email needed to be tweaked slightly by the participant before sending to fill in key facts such as account details or identifying information. In one case, Stephane Duguin's request to Swiss train company SBB, the request could not be executed because the listed email address in the privacy policy was invalid. In another, Jyrki Katainen's request to retailer Zalando, the request failed to be processed because the company could find no data associated with the address he emailed from. This appeared as a dead end, and it was only in retrospect that our analysis revealed this could perhaps have been circumvented by sending an email from a different address. The account holder problem necessitates yet more work by participants to see their responses through, e-mailing from the account-holding address and subsequently checking that account for responses (which caused further difficulty in the case of participants wishing to delegate tasks to their assistants, but without having to grant access to personal accounts). The account-holder problem served as a barrier too to some participants, who were unable or unwilling to submit a request to certain providers once they realised the account was actually in the name of another household member. In Leila Chaibi's case, she was told no data about her Lime scooter use was available, as that service had been used within the Uber app (Uber bought Lime) and not tied to a retrievable user account in the normal way.

Even where requests were successfully submitted, there were problems in some cases in getting companies to engage. In several cases participants resorted to using their connections to named individuals at companies to help 'unblock' requests. It was notable that journalists had more influence than politicians in this regard, both Mark Scott and Niclas Storås were able to expedite responses from companies (FullContact and Stockmann respectively) by contacting communications or press offices. The bar is very low, but this pattern of privileged access to journalists has been observed in other similar efforts[121] and might hint at increased sensitivity of corporate actors to bad publicity compared to standard GDPR enforcement processes.

In seeking an ideal SAR response, it appears to us that it matters not just who you know, as in the journalists' case, but also what your role is: As detailed in 3.4.2 below, Google (like Facebook, Twitter and others) routinely do not engage in the substance of Subject Access Requests, instead referring users to their download portal, Google Takeout. Drawing on experiences beyond this investigation, this is the uniform response from Google. However, something unusual happened in the case of Filomena Chirico and her SAR to Google. First, Google invoked the "3 month delay" clause claiming complexity - as is their right (GDPR Art.

[121] See for instance *L'Amour sous algorithme*, Judith Duportail, Mars 2019, La Goutte D'Or.

12.3). Then after that time had elapsed, Google did respond to the SAR, sending her a bespoke set of additional files, which came closer to addressing the substance of her request than what is available on Google Takeout, something we have never previously observed within or outside of this investigation. It is the personal and unsubstantiated belief of at least one of our co-investigators that Google's Data Protection Office recognised her role in the Breton cabinet in charge of platform regulation, and chose to provide a higher quality of response than they do to the average citizen, knowing that failure to do so might have greater consequences for them.

In general, companies responded with data files, or provided access to a download portal, within the 30 days mandated by the GDPR. Figure 57 shows an overview of participant success rates.

	Desired, Discussed or Considered as Target	**Successfully Targeted**	**Successfully obtained some data from target/company**
Number of GDPR targets	102	81 (79% of desired)	74 (91% of targeted)
Number of distinct companies	42	41 (97% of desired)	35 (85% of targeted)

Figure 57: Participant success rates in data access

However, the fact that most targets returned data hides the fact that not all these responses can be considered complete or adequate in terms of responding to participant inquiries or satisfying GDPR rights. Our email template collected together a number of different GDPR rights and requested all of them, as illustrated in Figure9 in Section 2.4.2.

There were two key problems with data that was returned. First, where data was returned directly, it did not always cover all categories. Consistent with prior studies[122,123] the return of derived and profiling information, low level metadata observations, and information about sharing, exchange and acquisition of data with third parties, was often absent or only minimally disclosed.

As a specific example, some of the participants requested some specific datapoints which are not available in the Download Your Information portal. To the best of our knowledge, Facebook should be conserving these long-term as statically associated to a profile. This includes:

[122] Alex Bowyer *et. al.*, Human-GDPR Interaction: Practical Experiences of Accessing Personal Data, CHI Conference on Human Fa§g Systems (CHI '22), 2022, https://doi.org/10.1145/3491102.3501947
[123] Michael Veale *et. al.*, When data protection by design and data subject rights clash, International Data Privacy Law, 2018, https://doi.org/10.1093/idpl/ipy002

- Xcheck data;
- Civic amplification score;
- Close-friendness data;
- meaningful people data;
- world2vec vector[124] (i.e. all static data on an individual, the contexts they can be situated in, as well as their personal data associated with any concept they can interact with or through - such as a post they can read or share, a group they can join, a friendship they are part of, etc);
- "Feed Unified Scoring System" data;
- Five user interest segments

Four participants (Filomena Chirico, Christian D'Cunha, Anders Adlercreutz and Nicolas Storås) specifically asked Facebook in their SAR request to provide these specific data points in addition to the standard data types requested of other companies, but this was not provided to anyone who asked, which represents a clear violation of the GDPR. Where participants tried to challenge this gap, they were stonewalled [see below]. It should be possible for journalists or regulators to pursue this matter further than we have been able to, but it might require extended explanations of why these data points are extremely meaningful for transparency - a much harder job than need be without access to underlying data[125].

[124] See CS 4803 / 7643: Deep Learning Guest Lecture: Embeddings and world2vec, a guest lecture at Georgia Tech by Facebook AI research engineer Ledell Wu, given Feb 18th 2020. https://www.cc.gatech.edu/classes/AY2020/cs7643_spring/slides/L13_Embedding_world2vec_final_version.pdf , archived at https://web.archive.org/web/20211018015836/https://www.cc.gatech.edu/classes/AY2020/cs7643_spring/slides/L13_Embedding_world2vec_final_version.pdf

[125] Meanwhile, in the absence of raw data to illustrate what would be the most meaningful data to extract from Facebook, we can rely on the insights of a variety of disciplines for structured approaches to the question of what is the most meaningful data to an observer situated within an environment. See:

- from a legal perspective: *Meaningful information and the right to explanation* by Selbst and Powles https://academic.oup.com/idpl/article/7/4/233/4762325 ;
- from an evolutionary biology perspective, *The information theory of individuality* by Krakauer, Bertschinger, Olbrich, Flack and Ay https://link.springer.com/article/10.1007/s12064-020-00313-7
- from a statistical physics perspective, *Observers as Systems that Acquire Information to Stay out of Equilibrium* by David Wolpert, https://www.youtube.com/watch?v=zVpSAjAe-tE as well as *Semantic information, autonomous agency, and nonequilibrium statistical physics* by Kolchinsky and Wolpert https://arxiv.org/abs/1806.08053 ;
- from a perspective straddling biology and physics, *Meaning = Information + Evolution* by Rovelli https://arxiv.org/abs/1611.02420 ;
- from a perspective closest to algorithmic systems and deep neural networks, see the articles by Bennequin and Belfiore: *Mathematics for AI: Categories, Toposes and Types* in the book *Mathematics for Future Computing and Communications* and their article *Topos and Stacks of Deep Neural Networks* https://arxiv.org/abs/2106.14587 .

The last approach has the advantage of making explicit the compositionality of neural networks in order to approach their explainability – think of a "divide and conquer" strategy through the mathematical theory of category theory. While that last approach is ongoing work at Huawei's research centre in Paris, focused on structuring "top-down" approaches to explainability of neural networks, the compositionality is relevant across the perspectives coming from different domains. Keeping this compositionality top-of-the-mind as an aligning backbone, and with the physical and biological instantiations as support

The nature of the responses (or lack of responses) has a clear impact on perceptions of companies: Detailed analysis data is not available at this time, but participants' scores of trust in service providers (collected before and after obtaining and reviewing data) generally decreased following the experience of GDPR, and similar changes in scores revealed that most target companies were found to be less transparent in their responses than participants had initially expected. A deeper analysis of the trust impacts of GDPR responses for lay people has been carried out in the prior study by Bowyer et. al.[8].

3.4.2. Many SARs Involved Blockages and Delays

The second problem in accessing data was experienced where companies provided a data portal (see Section 2.4.1) for user convenience, but then (by discouragement or blocking) made it hard or impossible to pursue the subject access request further. These 'SAR block' emails (explored further in Section 3.4.3) are problematic because, while the use of a download portal is certainly more convenient for users, it represents a standard set of the data that the company has voluntarily made available, and therefore cannot be guaranteed to have provided all data to which users are entitled, nor to have addressed specific points or questions posed by individuals within access request emails[126]. From a company perspective, it is clear that the use of download portals is seen as a cost-saving measure, but we contend that it can never entirely replace the need for an effective communication channel for dealing with aspects of access requests that are not satisfied through the download portal.

A number of participants experienced delays in getting a response to their SARs, despite the 30 day response obligation. Some companies, including Booking, KLM, and Google (in Filomena Chirico's case only) immediately invoked the 3 month delay on the grounds of complexity, but in some other cases target companies did not respond within the 30 days they are legally allowed to take. The nature of this investigation makes it impractical to produce precise figures on response times, due to the excessive effort burden it would place upon participants to obtain this timing information. We also saw evidence of target companies in some cases starting the 30 day counter from a date other than the user's initial email; it is not supposed to be allowed that a company can wait 3 weeks to acknowledge a request and then take 30 days from the date of that acknowledgement.

for intuition, we can thread back from the neural network perspective onto the legal perspective to inform the raw data that should be requested through SAR requests for maximal impact. One of the co-authors sees the effort of #digipower participants towards requesting these particular data points as a natural continuation of his previous efforts on transparency of Custom Audiences, as outlined in Section 3.3. In fact, both approaches are indeed rooted in compositionality. The role of compositionality is further explained in the narrative report.

[126] In fact, in the consumer protection *Re Facebook Privacy Litigation, 791 F. Supp. 2d 705 (N.D. Cal. 2011)* case in North California courts, the existence of *Download Your Information* is quite clearly used by Facebook to try to avoid transparency obligations demanded by the plaintiffs, in an effort to map the US judicial discovery process to the European data protection obligation of access.

Other delays and obstacles encountered by participants included requirements to fill, scan and email in forms (despite the fact that this is not a lawful practice[127]), or to complete additional ID requirements before their request could be processed.

3.4.3. Each of the Three Lenses had Limitations

With each of the three lenses (see Section 2.4) unique insights were obtained, but also limitations were encountered.

Subject Access Requests (SARs) offer the broadest and also the most specific data access, one can ask any question or for any datapoint, and if it falls within the constraints of what the GDPR allows, a response is supposed to be provided (and even if it does not, it may still be given, as opinions differ on exactly what data points and questions are in scope of GDPR compliance). However, the process was found to be time-consuming, involved a lot of effort, and was often disappointing as it may not answer a specific question or may result in data that is hard to use or understand, or may yield no answer at all. Any data obtained is a copy and might be out-of-date.

Download Portals were generally found to be powerful tools offering users the ability to access a lot of their data, generally in standard and portable data formats, in a matter of minutes or hours (or sometimes a few days). The data returned was generally very up-to-date and extensive. However, the scope of data returned is determined by the voluntary choices of the company as to what to share, which may not answer specific questions or address all GDPR data access rights. In this light, it can be seen that companies referring SAR requesters to download portals does not satisfy GDPR requirements.

TrackerControl, and similar audit-based approaches offer a different angle, looking at what apps actually are doing. These have the advantage of identifying specific facts about provider-to-third-party data relationships that may not be visible in SARs or Download Portals, but do not provide any information about what data is being transferred, or for what purpose[128].

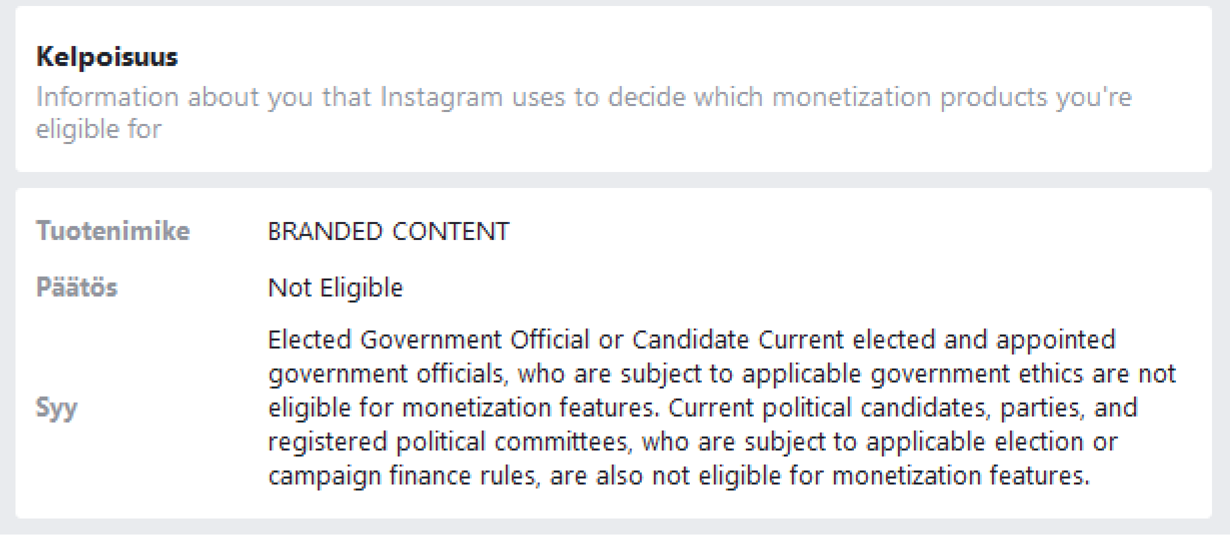
3.4.4. Response quality was variable

Across the data returns that participants received, we saw a wide variation in response quality, both in terms of how detailed or well-formatted the files were, and in terms of how understandable they were. In many cases, explanatory texts and lists of abbreviations and field names were provided, which aided understanding. With some companies, such as HSL, Gigantti Apple and Google, there was a difference in the breadth of data returned to different

[127] UK Information Commissioner's Office, Subject Access Code of Practice (9 June 2017) p13; Information Commissioner's Office, 'Guide to the GDPR: Right to access' (22 May 2019)

[128] Note that this limitation is due to gatekeeping imposed by OS or hardware manufacturers, often justified through the argument of security.

participants. In some cases participants saw data that uniquely recognised their status as public servants, such as in Sari Tanus' Instagram return. Figure 58 shows that Meta are aware of her status as an MP, and treat her differently for advertising purposes accordingly.



Kelpoisuus

Information about you that Instagram uses to decide which monetization products you're eligible for

Tuotenimike	BRANDED CONTENT
Päätös	Not Eligible
Syy	Elected Government Official or Candidate Current elected and appointed government officials, who are subject to applicable government ethics are not eligible for monetization features. Current political candidates, parties, and registered political committees, who are subject to applicable election or campaign finance rules, are also not eligible for monetization features.

Figure 58: A screenshot from Sari Tanus' Instagram return, showing how her status as an elected representative to the Finnish Parliament is recognised and recorded in data.

Some of the lowest quality responses were obtained from French train company SNCF and French newspaper Le Monde, targeted by Leïla Chaibi, both of whom provided only brief PDFs (2 and 9 pages respectively) with screenshot/image based content. Thuisbezorgd (a.k.a. Takeaway/Just Eat) provided a very limited response with a single page PDF (although this was at least textual and hence copy/pasteable into a spreadsheet for analysis, rather than being image based). European guidelines state that PDFs are not to be considered as machine readable responses[129], which is relevant because all participants' requests were clearly identified as being both a Subject Access Request *and* a Data Portability Request, the latter of which requires data to be returned in machine-readable formats.

In terms of volume of data, the greatest breadth of data was available from Google, from their Google Takeout download portal, which amounts to several gigabytes if you download everything (however a lot of this volume comes from user data such as photos, videos, Drive files and Gmail). Facebook's portal download was also extensive in breadth, though (with the exception of their Offsite Activity Data) largely contained user-volunteered data. Uber, K-Ryhmä, Spotify, Sanoma,Twitter, and Deliveroo all returned detailed and extensive sets of files.

By far the best of the SAR responses, not just in this investigation but among all the hundreds of SAR responses the lead investigators have ever seen, was provided by British/Nordic retail giant

[129] Article 29 Working Party, Guidelines on Transparency under Regulation 2016/679 WP260 rev.01, 11 April 2018.

Gigantti, to both Sari Tanus and Miapetra Kumpula-Natri. This consisted of a 5 page PDF with detailed, transparent and understandable explanations of what data was included in the return, why they keep it, and how it has been used or shared. What was exceptional about this response was that it used the terminology and framing expressed in our participants' email requests, and had clearly been written bespokely with great care having been taken to address the questions and requests we outlined in the email, and with extraordinary candour, providing a window beyond the immediate return of data and into some of the data exchange relationships that retailers have with Facebook, Google, and other third parties. This has led however to disproportionate visibility for that company, which is in certain ways unfair compared to some companies simply ignoring participants' requests, which in turn led us to not include it in this report (since this was a distraction from our core findings). The co-investigators would like to insist that within an entire ecosystem where the norm is to not be transparent (which is currently the case), the only stakeholder incentivized to publicise non-compliance is the regulator – and that this is simply their job.

3.4.5. Challenging Responses and Gaps was Ineffective

A number of our participants, upon experiencing resistance to the broadness and specificity of their SAR requests, replied to data protection officers at companies including Facebook, Google, Gigantti and Uber with follow-up queries or challenges. Some responses to queries were helpful, such as when Sari Tanus asked Gigantti to clarify why she had been given less data than Miapetra Kumpula-Natri - the reason being that customer club members are tracked in more detail. However in some cases, queries and challenges were not dealt with satisfactorily, if at all.

Meta (Facebook), after multiple back-and-forth messages asking for data and information (the data points described in Section 3.4.1) that were not available on Facebook's Download Your Information portal (the problem described in Section 3.4.2), appear to have a practice of outright refusing to take the individuals' subject access request any further. As illustration, please see Figure 59, which aggregates a pattern of responses experienced by multiple participants. It shows how Meta systematically avoids having to engage individual bespoke subject access requests and refuses to provide information they are known to keep and hence legally bound to supply. This endurance test was documented by having Riitta Vänskä, #digipower project manager at SITRA, pursue the same requests to Meta, in parallel to the participants.

Some of the #digipower participants asked for the same data points, and simply had to give up earlier given their extremely busy schedules. She went as far as filing a complaint with the Finnish data protection authority (see Appendix), with the conclusion that the Finnish data protection authority would have to rely on the Irish Data Protection Authority to investigate, which would take at least a year. This is for data points that are held by Facebook, yet not available in Facebook's Download Your Information portal - a determination that should be very quick to make and that indeed all involved participants have been able to make for themselves. In the context of the #digipower investigation, this lack of access was very unfortunate, as this data is of utmost relevance for understanding in depth how Facebook enriches raw data, and even architectures their systems - sometimes to evade regulatory action as alleged in ongoing US lawsuits[130].

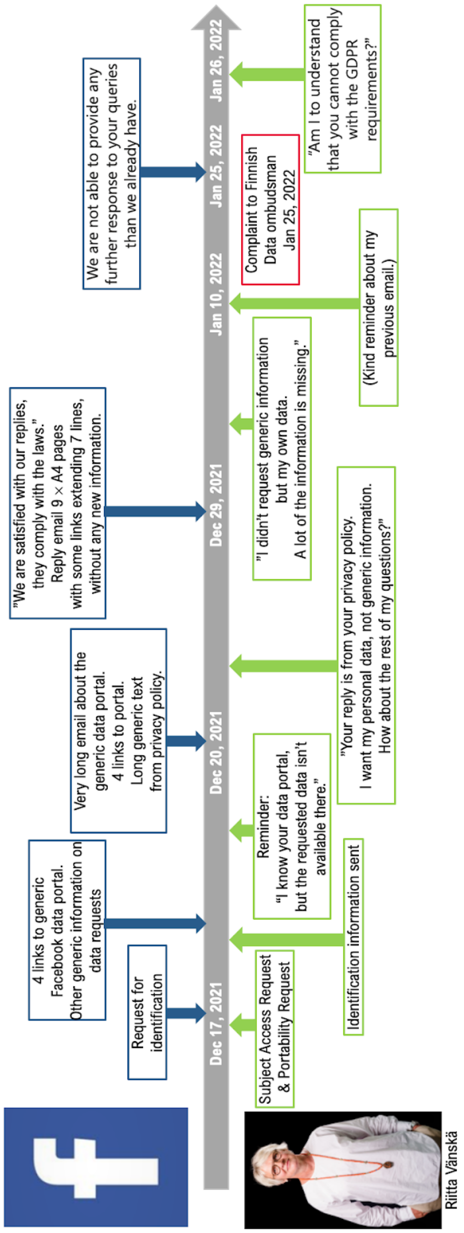

Dec 17, 2021
Dec 20, 2021
Dec 29, 2021
Jan 10, 2022
Jan 25, 2022
Jan 26, 2022
Subject Access Request & Portability Request
Request for identification
Identification information sent
4 links to generic Facebook data portal. Other generic information on data requests
Reminder: "I know your data portal, but the requested data isn't available there."
Very long email about the generic data portal. 4 links to portal. Long generic text from privacy policy.
"Your reply is from your privacy policy. I want my personal data, not generic information. How about the rest of my questions?"
"I didn't request generic information but my own data. A lot of the information is missing."
"We are satisfied with our replies, they comply with the laws." Reply email 9 × A4 pages with some links extending 7 lines, without any new information.
(Kind reminder about my previous email.)
Complaint to Finnish Data ombudsman Jan 25, 2022
We are not able to provide any further response to your queries than we already have.
"Am I to understand that you cannot comply with the GDPR requirements?"
Riitta Vänskä

Figure 59: Facebook's endurance test for users, as experienced by Riitta Vänskä, #digipower project manager at SITRA, when trying to demand data points held by Facebook which are not available in Facebook's download portal.

[130] See for instance *Klein v Meta*, an antitrust action, where it is alleged that Facebook/Meta engineered their machine learning systems to intermesh data obtained through Facebook (Blue), Instagram and WhatsApp in order to make it impossible to disaggregate the company's social data by source or line of business.

3.4.6. Following Up on a SAR Response

One positive outcome from Subject Access Requests, that shows that they are delivering actionable information in some cases, is where participants felt empowered to take further action and make further subject access requests as a result, in order that they could peer further into the data ecosystem based upon what they learned. This happened in two cases, the first where Sari Tanus learned about data broker Bisnode through her Aamulehti SAR, and subsequently sent a SAR to Bisnode to find out more about what data they have and where they got it from, and also in Mark Scott's case, where a SAR to FullContact (another data broker) allowed him to find out where they had got their data about him, which turned out to be from a contacts app called Contact Plus (a service which was itself a spinoff from FullContact)

.

4. Coaches’ Reflections

4.1. #DIGIPOWER PROCESS

Overall, the investigative process was time-consuming for the participants. Scheduling meetings was extremely difficult due to their busy schedules. They needed constant help with the Subject Access Requests, mostly due to non-standardised systems in the companies they were facing. The combination of three lenses (SAR, data download portals, and technical audit with TrackerControl) in our study sometimes contributed to this burden, but also made the investigation more resilient. The SARs proved helpful in communicating to participants that this is a process they could be active in, and to expand the range of service providers' targeted beyond those that had a data download portal or an app that could be audited with TrackerControl. Portals were efficient for their predictability in terms of quick access to data, and Hestia.ai's visualisation tools were powerful in terms of digital literacy to read files received, analyse the data and make it understandable. The use of TrackerControl was insightful as a baseline, across apps and across participants, to provide predictability in the output beyond what SARs could deliver. While each lens had its strengths and drawbacks, they were definitely complementary. A significant observation made when auditing an app through TrackerControl, was to see an app sending data to multiple third parties, and not see a trace of this data flow in the Subject Access Request results[131].

Throughout the investigative process we found that the participants had different reactions according to every data type returned by the service providers:

- participants found volunteered data such as a name and email not so consequential – they clearly knew what they had provided;
- they sometimes found observed data surprising, such as geolocation collected in circumstances they would not expect, or data that was kept for a very long time;
- acquired data like the segmentation categories of residents' postal code area identified from the customer's address bothered them more, and it felt invasive in the few instances where we found evidence of how a service provider directly acquired participants' personal data from third-party providers;
- derived data like inferred Twitter interests from user activity provoked a lot of curiosity and also amusement whenever it was wrong;
- participants had a lot of trouble understanding the relevance of the metadata to the data economy (e.g. how one could build a business model when merely observing someone else's business processes through data), and needed lots of assistance with this data type, but they could understand the significance of onward data flows from a privacy perspective (for instance with Gigantti transferring purchase data to Facebook, which was immediately understood as surprising and concerning).

During the coaching sessions with the participants, it was invaluable to have concrete evidence in front of them. Having their own data available to view as facts focused the participants on

[131] This happened, for instance, for any participant who had used the Helsingin Sanomat app.

understanding what they saw – and the motivations to collect it – without the need for any theoretical discussions as motivation. We are grateful for the participants' willingness to trust the coaches in showing and sharing their personal data with us. Without this, the discussions in the deep dives would have been much shallower and harder to facilitate, and without our ability to examine and analyse returned data files, most of the insights and visualisations presented in this and the narrative report would not have been possible.

We conducted the investigation through fifteen separate individual tracks, one for each participant, which was a necessity due to their extremely busy schedules. In hindsight, we should have held workshop-style sessions to enable sharing knowledge amongst participants without repeating ourselves, as well as to facilitate a fluid and "horizontal" consent process from the participants that enables sharing the evidence received between participants or groups of participants, so they could benefit more and faster from each others' learnings without the coaches being a bottleneck as a central hub of expertise and data flows. Such a workshop-based approach would also have enabled more participant-led discussions about issues arising, as these would have been discussed among participants earlier, and would be ready to discuss in public at an earlier stage.

The targets and goals for participants in this study were quite open-ended, and necessarily so. The findings of this study have revealed many specific areas where transparency and compliance is weakest (for example, the lack of information provided about data sharing and data exchange, or the "SAR block" practice of large companies described in 3.4.2). These areas could be targeted more efficiently in a future investigation, which would reduce the workload upon participants, and potentially have a strong chance of impact, especially if combined with the influence some of our participants - journalists and EU civil servants in particular - had over data holding organisations to press for better compliance.

4.2. MAXIMISING PEDAGOGICAL VALUE, IN THE CONTEXT OF THE NARRATIVE REPORT

While this methodology and case studies report is focused on illustrating the power dynamics in the data economy with participants' own data, in the accompanying narrative report (entitled *"Understanding Influence and Power in the Data Economy"*) we focused on systematically deconstructing the mechanisms of power acquisition. In particular we highlighted two feedback loops:

- *Accumulating Information and Knowledge to Act* loop
- *Composing Complex Infrastructures for a Dominating Position* loop

The first loop refers to the power's holder capacity of accumulating information and knowledge to act over stakeholders, i.e., a service provider, a user, a group of service providers, and users.

The power's holder can then control information flows, the veracity of information, guide population behaviour and manipulate their mindset.

This loop is illustrated by both case studies 1 and 2: *Who cares about my geolocation, and why?* and *When you view the web, the web views you*. These cases are concerned respectively with navigating physical space and navigating content but a comparison is fruitful as justified in chapter 3. By situating the participant's agency in a purely online context (picking content) or purely offline context (moving around), we were able to better contrast how their actions and mindset were structured between the two contexts. This made more explicit the influence of the data economy over their content and movement choices.

This is not just an analogy though, and we struggled for a very long time in finding the right way to present our argument so it would be consistent. We found that going through the notion of offline and online *worlds* was too restrictive. This distinction is largely debated in the literature.

According to our evidence, the two worlds are intermeshed and digital power produces more harmful effects on society than traditional power in how they are interlocked and interacting with each other when they are digitised. So we found it more helpful for understanding how digital power works to talk about *contexts* and not about offline and online worlds.

The second loop refers to the mechanisms through which a service provider composes, or builds, a complex technical architecture made of multiple simpler architectures. It enables a power's holder to organise relationships between architectures to its benefit, this way knowing more about the population's behaviour across services. The most impactful decisions are about the protocols, the rules in which every composition is made, and the communication that is possible between services.

This loop corresponds to case study 3 in this report, *Move fast and capture all signals, everywhere*. We start there by introducing the entire ad tech ecosystem in Section 3.3.1. This ecosystem is known as highly problematic, and indeed we introduce more evidence of this – partly thanks to its fragmented nature. However it merely serves as a background to Section 3.3.2 where we focus especially on one actor, Facebook, and show how they are able to orchestrate the entire ecosystem of advertisers, publishers or even data brokers to their benefit. While through the #digipower lenses we would have been able to provide evidence of many Facebook data collection channels, we would have struggled to contextualise them since what happens inside Facebook is opaque. Fortunately ongoing lawsuits have revealed lots of internal documents, which helps us thread everything together a bit more throughout Section 3.

4.3. FINALLY, A MAP

Much of what we have described is of course depressing. We take pleasure in presenting (a simplified view) of the aggregate work of all the participants and the investigators in Figure 60, as we feel these are first steps to build a coherent view across the entire industry of data flows.

Note that unlike most such maps, this is based on evidenced individual data flows[132] and involves multiple complementary lenses.

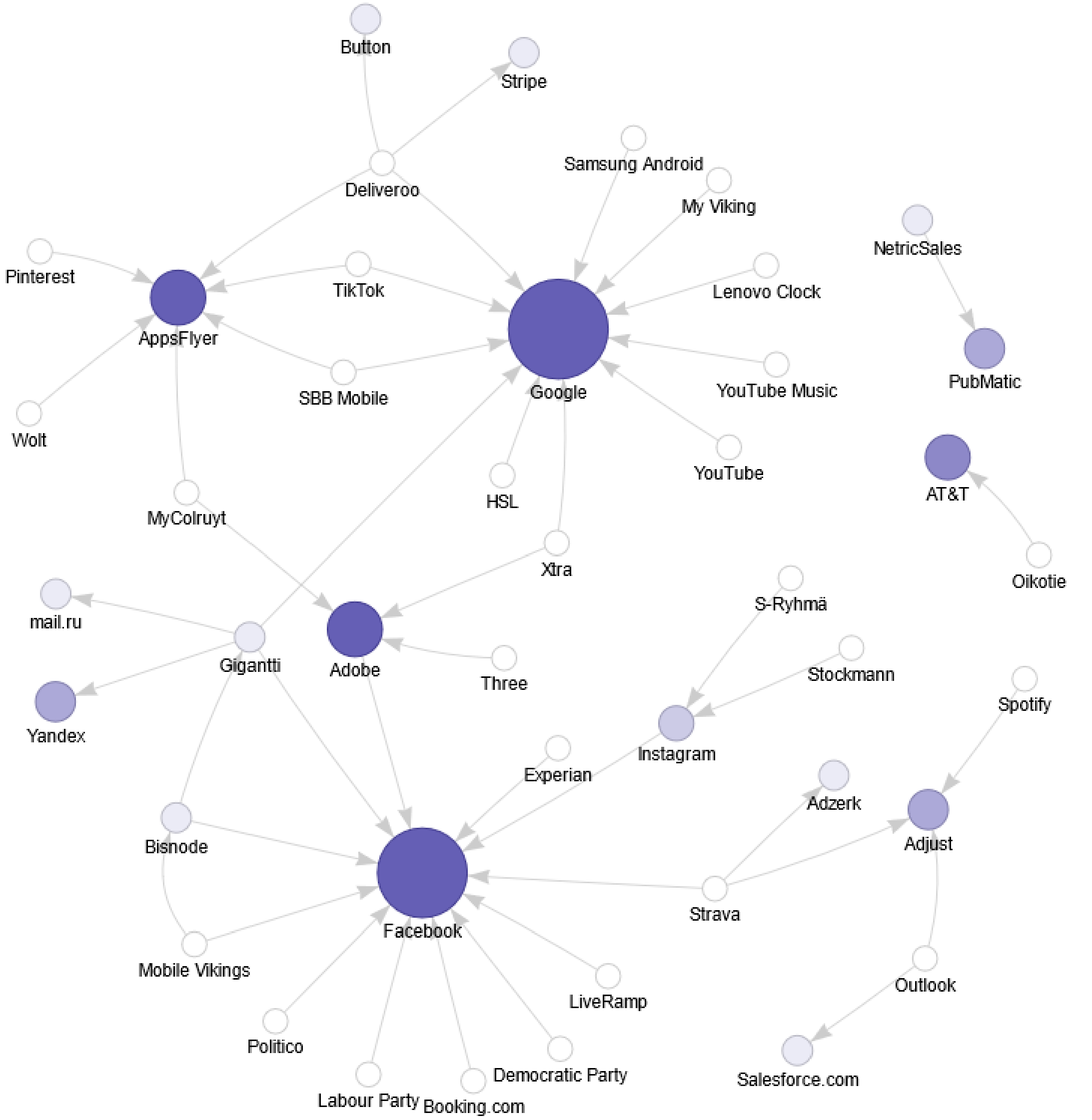


Figure 60: A simplified aggregate view of some of the data flows uncovered by the participants, excluding media companies as origins of flows (because they would dominate). The size of the inbound nodes is determined by the number of inbound flows (considering this time also those media companies).

[132] as opposed to information obtained by reading privacy policies, or through automated visits.

Appendix: a response from the Finnish data protection authority

As explained in Section 3.4, one of the #digipower project managers, Riitta Vänskä, has pursued her Facebook data extensively. After lots of unsuccessful efforts, she has resorted to complaining to the Finnish data protection authority (Office of the Data Protection Ombudsman, Tietosuojavaltuutetun toimisto), which has led to the following response:

Hyvä Riitta Vänskä,

kiitos yhteydenotostanne tietosuojavaltuutetun toimistoon.

Olette tiedustelleet asianne käsittelyn tilannetta. Pahoittelen vastauksemme viipymistä.

Valituksenne odottaa vielä käsittelyvuoroaan tietosuojavaltuutetun toimistossa.
Tietosuojavaltuutetun toimisto on valitettavasti ollut hyvin ruuhkautunut, mutta pyrin edistämään asianne käsittelyä mahdollisimman pian.

Tässä vaiheessa voin todeta asianne käsittelystä sen verran, että Facebook on sijoittautunut ETA-alueella Irlantiin ja luultavasti tietosuojavaltuutetun toimiston tulee näin ollen käsitellä asianne yhteistyössä muiden jäsenvaltioiden valvontaviranomaisten kanssa ns. One Stop Shop -yhteistyömenettelyssä. Johtavana valvontaviranomaisena toimii siten todennäköisesti Irlannin tietosuojaviranomainen. Tietosuojaviranomaisten välinen yhteistyö perustuu yleisen tietosuoja-asetuksen VII lukuun. Asioiden käsittely tapahtuu Euroopan komission tarjoaman sisämarkkinoiden tietojenvaihtojärjestelmän (IMI) kautta. Yhteistyömenettelyyn vietävien asioiden käsittelyaika vaihtelee asian laadusta ja sen edellyttämistä lisätoimenpiteistä riippuen, mutta käsittelyaika-arvio on tällä hetkellä valitettavasti yli yksi (1) vuosi.

Lue tarkemmin valvontaviranomaisten rajatylittävästä yhteistyöstä.
Lue tarkemmin toteuttamastamme henkilötietojen käsittelystä.

Toimin jatkossa asian käsittelijänä tietosuojavaltuutetun toimistossa ja pidän teidät ajan tasalla prosessin etenemisestä!

In English, the most relevant parts translate to (manually improved automatic translation):

Dear Riitta Vänskä,

Thank you for contacting the Office of the Data Protection Ombudsman.

You have inquired about the state of play of your case. I apologize for the delay in our response.

Your complaint is still pending before the Office of the Data Protection Ombudsman. Unfortunately, the Office of the Data Protection Ombudsman has been very congested, but I will try to take your case forward as soon as possible.

At this stage, I can say to the extent that Facebook is based in the EEA in Ireland we will probably have to deal with your case in cooperation with the supervisory authorities of other Member States in the so-called In the One Stop Shop co-operation procedure. The lead supervisory authority is therefore likely to be the Irish Data Protection Authority. Cooperation between data protection authorities is based on Chapter VII of the General Data Protection Regulation. Cases are dealt with through the Internal Market Information System (IMI) provided by the European Commission. The time taken to deal with cases subject to the co-operation procedure will vary depending on the nature of the case and the additional measures required, but unfortunately the estimated time for processing is currently more than one (1) year.

Read more about cross-border cooperation between supervisory authorities.[133]
[..]

I will continue to deal with the matter in the Office of the Data Protection Ombudsman and will keep you informed of the progress of the process!

[133] With link pointing to https://tietosuoja.fi/en/processing-involving-several-eu-countries