Exploring data sharing obligations in the technology sector

Appendix

Rotterdam, 30 November 2020

Exploring data sharing obligations in the technology sector

Appendix

Nicolai van Gorp
Paul de Bijl
Inge Graef
Gabor Molnar
Roel Peeters
David Regeczi

Rotterdam, 30 November 2020

ECORYS

Table of contents

Table of contents

1 About this appendix

The Ministry of Economic Affairs and Climate has commissioned Ecorys and its partners to investigate possibilities to realise the gains from data usage. The central question is: what is the role of data in the technology sector and what are the opportunities and risks of data sharing? The report explores this question in order to obtain insight into effective policy options that can help to speed up data sharing to the benefit of society and the economy. This document is a (separate) appendix to that report.

The analysis in the main report draws from five complementary case studies, on (i) online search, (ii) social networks, (iii) banking, (iv) business software, and (v) vertical exclusion. The last one does not zero in on a specific market or sector but is a topical case.

Chapter 1 in the main report describes the research methodology, including the approach to explore the cases. Note that in the case discussions that follow, we include observations from interviews. As discussed in chapter 1, for reasons of confidentiality, we cannot disclose the origin of such statements.

2 Questions in semi-structured interviews

Part of the approach to explore the cases was to carry out semi-structured interviews, in which we started with similar questions, to be followed up by case-specific discussions. The common questions were the following:

1. Which types of data do you use (which typology do you use in relationship to your business model, for instance when collecting, processing and storing different types of data)?
2. Which types of functionalities / services do you offer in connection to these types of data?
3. For each of these data types, how do you learn from them, and how does data-enabled learning translate into your value proposition to your customers?
4. Does the added value of extra data to improve your offerings eventually reduce (and if so, how fast), or is having more data always valuable? When and why? Does this differ between data types?
5. How important is (near) real-time data for your business model?
6. Where and how do you obtain these data types? Please describe to what extent you enter into agreements with data collectors, and in case of such agreements, what types of conditions typically apply. Which types of data are typically not included in data-sharing agreements and why?
7. To what extent are the data types you use proprietary and unique, that is, to what extent can it be obtained elsewhere, easily imitated, or reverse-engineered by competitors?
8. How hard is it for others to copy your functionalities or services without having access to the data that you use?
9. Which types of mandatory data access (at reasonable/viable tariffs) could have positive effects on competition and innovation?
10. Which types of mandatory data access would be most harmful to competition and innovation and why?

3 Case: online search

3.1 Case description

Search engines (such as Bing and Google search) try to match information seekers with information available on the Internet. In order to do so, these companies invest in indexing (the content on) websites and in developing algorithms for interpreting search queries entered by information seekers and matching these interpretations with entries in the index.

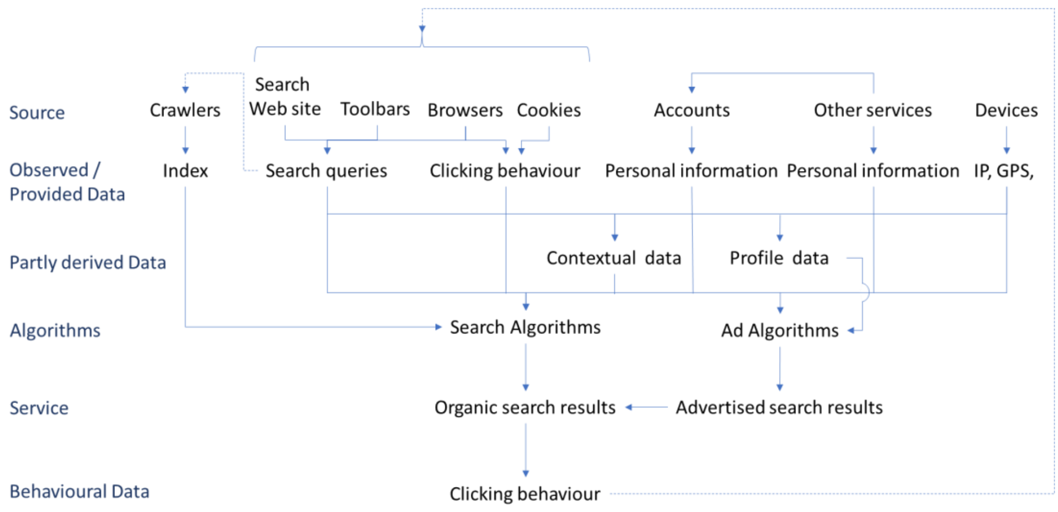
The development of search algorithms is a process which learns from past results. This would give rise to a first-mover advantage and contribute to a risk of tipping the market. However, the learning effect is not the only driver of growth which may explain market shares. Other drivers may be network effects[1], scope economies, innovative merrits and (possibly) anti-competitive behaviour.

The search market is highly concentrated with Google serving around 90% of the market. A hypothesis is that, in the absence of data sharing, rivals are not able to catch up with Google. If this hypothesis is confirmed, a central question is whether mandatory data sharing can facilitate entry and restore competition. In addition, what does this mean for privacy and innovation incentives?

3.2 Preliminary analysis

Internet Search Building Blocks

Internet search services comprise of the following building blocks:



Sources and data

When using a search engine, one is not searching the Internet itself, but an index or library of webpages stored by a search company like Google or Microsoft. These indexes are created by **web crawlers**, which are automated processes that record content of webpages and follows links from those webpages to other webpages to record their content. While websites change on a daily basis, web crawling never stops. All the information which is gathered through web crawling is then processed into an **index** like an index at the end of a book, with an entry for every word on every webpage that is indexed. Besides websites, an index may also include other information, such as

[1] In this case, indirect network effects from searchers to advertisers.

information about locations, videos, books, pictures, art, buildings, public transport schedules, and programmes from sport leagues.[2] Note that search queries may signal crawlers to start indexing particular content because the users appear to gain interest in that content.

An algorithm goes through the index to link websites to search queries. It needs to interpret the search query in relation to the information in the index. This involves language analysis (like interpreting spelling mistakes and homonyms[3]), analysis of backlinks (how many other websites refer to a website in relation to a search query), analysis of comparable **search queries** by other users in the past and their subsequent **clicking behaviour**[4,5,6]. Search queries and clicking behaviour is either provided by the end-user when using a search engine or it is observed by cookies or the web browser (EC-Council, 2016, p. 87). Note that clicking behaviour is behavioural data which results from interaction with the search engine's user interface and feeds back into the algortihm. This feedback loop results in learning effects.

Next to search queries, algorithms also incorporate **contextual information** (such as location, device used, et cetera) to better interpret search queries.[7] For example, the algorithm may differentiate between search queries entered on mobile devices and those entered on desktop devices or between people's modalities[8]. This type of information is typically observed by logging IP-addresses, tracking locations, finger printing, cookies, et cetera. Sometimes, **personal information** like recent search activities[9] may serve as contextual information. However, the other personal information (age, gender, nationality, occupation, or hobbies) feed into **profile data** which generally does not feed into the search algorithm to maintain objectiveness of organic search results. Profile data does (along with the other data) feed into the algorithm that matches advertised search results to be displayed among the organic search results (often with a disclaimer). Personal information may be provided by the individual when subscribing to an account, but it may also be observed or derived from data obtained through cookies or browser.[10]

The process of developing the search algorithm

The previous discussion shows that internet search is data intensive. The volume of data is growing with the volumes of websites and internet users, as well as the number of explanatory variables entering the equation. This makes the problem which the algorithms must solve more complex. In order to solve those problems, technologies are employed such as **data mining, machine learning and data fusion** (Kathuria, 2019). Without going into the details of the difference between these technologies, the basic idea is that data are merged and combined at a large scale which allows for obtaining more consistent and accurate information than the original raw data and for finding new correlations that would otherwise not have been identified. For example, Google's RankBrain is an AI technology that interprets what people are looking for based on their (often) incomplete, inaccurate and misspelled search queries. This algorithm learns from behavioural data (such as search queries and clicking behaviour) as well as contextual data (mother tongue of an individual, recurring spelling mistakes, past search queries, locations, device used, et cetera).

---

2 https://www.google.com/search/howsearchworks/crawling-indexing/

3 such as I *left* my phone on the *left* side of the room

4 https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines

5 https://backlinko.com/google-rankbrain-seo

6 Clicking behaviour includes (amongst others) ***click-through-rates*** (how many times has a person with similar search query clicked on a particular website), ***dwell time*** (how much time users spend on a website once they click it) and ***pogo sticking*** (the degree to which users keep going back and forth till they find what they look for). See https://onlineidealab.com/google-rankbrain-what-is-it-how-can-you-benefit-from-it/

7 Case AT. 40099 (Google Android)

8 such as whether they are walking, driving in public transport, or driving in a car.

9 https://www.google.com/search/howsearchworks/algorithms/

10 https://edu.gcfglobal.org/en/internetsafety/understanding-browser-tracking/1/

In addition, the algorithms require creativity from engineers[11], sandbox testing[12], lab-panel testing by (over thousands of) 'search quality raters'[13], and live traffic experiments using a small fraction (0.1% according to Google) of the users[14]. Note that the development of a search service is language dependent. In other words, many of the investments in innovation and testing need to be duplicated for every new language in which the search service is provided.

3.3 Analysis of research questions

3.3.1 What role does data play in the business model?

Data is a key input (next to other inputs and innovative activities) for generating relevant search results and for improving the quality of these search results.

Indexes

As is clear from the previous, the Index is an essential data base for search providers. Interview partners indicated that there are four indexes with sufficient scale and quality to provide meaningful search results (those of Google, Bing, Yandex, and Baidu). In addition, there are open-source crawlers such as Common Crawl[15], providing free access to copies of significant parts of the web. Although this index is much smaller in size[16], this does not say much about the quality of the index. One of our interview partners indicated that indexing the most relevant or useful webpages quickly leads to a better outcome than trying to index as many webpages as possible.

'Optimising' a web index is a matter of technical analysis, insights, and judgment calls. However, there are also feedbacks from search queries to the web crawling process: people reveal interests by entering search queries, these revelations instruct crawlers to target these interests in the indexing process. For example, in case of an unexpected event of global relevance (say a meltdown of a nuclear power plant in Fukushima), news items about this event are popping up all over the internet. These news items need to be indexed quickly. The crawlers must immediately go to work and target relevant news items. The fact that people suddenly start searching for specific terms triggers crawlers to get moving and to target keywords like "nuclear power plant", "meltdown" and "Fukushima".

Data as input for training algorithms

Algorithms interpret search queries and look for the most relevant websites to provide answers to the questions of searchers. These algorithms are developed using a combination of human creativity and machine learning, followed by lab-experiments using panels and subsequent live traffic experiments in the form of A/B testing involving a small fraction of users[17].

Much of the literature on search engines focusses on data as an input for machine learning, i.e. training algorithms.[18] The notion in these papers is that algorithms learn from analysing correlations between search queries and clicking behaviour. The more search queries that can be analysed, the more significant these correlations become, the faster the search algorithm learns, the higher click-

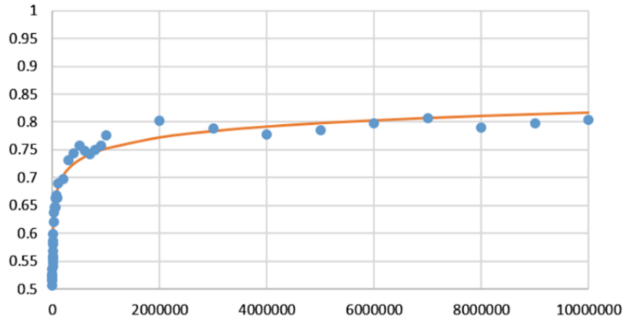
---

11 See https://www.google.com/search/howsearchworks/mission/users/
12 https://searchengineland.com/google-to-deliver-caffeine-after-holidays-29479
13 https://www.intermediair.nl/innovatie/algoritmes/google-laat-grootste-ai-machine-ter-wereld-door-mensen-factchecken
14 See https://www.google.com/search/howsearchworks/mission/users/
15 https://commoncrawl.org/
16 WebCrawler has around 2.6 billion websites indexed, while Google had indexed 130 trillion websites in 2016 (up from 30 trillion in 2013). See http://commoncrawl.org/connect/blog/ and https://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378
17 See https://www.google.com/search/howsearchworks/mission/users/
18 See (He, et al., 2017) (Argenton & Prüfer, 2012) (Prüfer & Schottmüller, 2017) (Anderson, 2020)

through-rates and dwell time, and the less likely will people engage in pogo sticking[19]. Higher click-through-rates and dwell time (and a lower degree of pogo sticking) implies that the user perceives a higher quality of the search engine, which attracts more users, which increases the volume of search queries. This learning effect has also been described as an indirect network effect between past and current users (Argenton & Prüfer, 2012), and as ‘data feedback loop’ (OECD, 2016).

The previous indicates that the learning effect may create a first-mover advantage for search engines with a large market share. Based on this, Microsoft argued in the Microsoft-Yahoo! Merger case (2010) that it lacked the necessary scale. The company claimed that *“increased scale will impact on the user experience, because the trailing volume of search queries holds each party back from improving their search results”*.[20] Various authors (including Google’s chief economist Hal Varian and Amazon’s Pat Bajari) have downplayed this statement by pointing out that the strength of the learning effect diminishes with scale (Anderson, 2020)[21]. Indeed, He et al. (2017) find that that the learning curve of a search algorithm flattens out at low volumes of search queries (see figure below). Moreover, it has been argued that the first-mover advantage expires rather quickly because the shelf-life of data is limited (that is: the value of past search queries to generate relevant search results expires quickly)[22].

Figure 3-1 ***Relation between CTR and historical search queries***



Source (He, et al., 2017)

In relation the Microsoft/Yahoo merger, both the European Commission as well as the U.S. Department of Justice recognised that Google’s competitors indeed had (or could realise) the necessary minimum efficient scale for a majority of (popular) search queries. However, Microsoft claimed that there is a long tail of specific and rare queries (which make up 20%-25% of all queries) with volumes at which the algorithm will still be learning fast. The US Department of Justice (2010) concluded that this fact reinforces scale economies (and hence a first mover advantage) because a query is more likely to be unique when the number of users of a search engine is small[23]. Moreover, it has been argued that rare and unclear queries are interpreted by linking historical examples of related queries and leverage historical examples that differ in minor ways from the

[19] ***Click-through-rates*** refers to how many times has a person with similar search queries clicked on a particular website, ***dwell time*** means the time users spend on a website once they click it, and ***pogo sticking*** is the degree to which users keep going back and forth till they find what they look for.

[20] Case No COMP/M.5727 - MICROSOFT/ YAHOO! SEARCH BUSINESS, para 162

[21] This is not surprising as general statistics teaches us that relevance (R2) increases with the sample size (n) at a diminishing rate; at a certain point, the sample is large enough for getting significant results and an increase in sample size does not lead to higher relevance.

[22] The Internet is constantly updated with new information and content, making the concept of ‘relevance’ a moving target. As such, search algorithms are constantly updated with new combinations of queries and click-through-rates. This implies that the shelf-life of data (and thus any first-mover advantage) is limited.

[23] if the number of users is one, than all search queries are unique.

target query (He, et al., 2017).[24] [25] This technology has the effect of expanding the shelf-life of historical search queries, which reinforces the first-mover advantage. Moreover, in relation to the feedback from search queries to the process of indexing (see above), the fact that data expires quickly enhances the first mover advantage (in terms of number of users) rather than mitigating it[26]. Furthermore, the introduction of technologies such as data mining, machine learning and data fusion propels the number of explanatory variables entering the search algorithm. These technologies allow for solving increasingly more challenging problems, but in order to do that, they require more data (Anderson, 2020)[27].

The previous observations point to data feedback loops and scale effects in the learning process of algorithms. While a vast stream of literature confirms this, little research has been done on the relative importance of this effect (compared to other factors) for the overall quality of search results and the competitive position of a search provider. As explained, machine learning is only one element of developing search algorithms. It also involves the creativity of engineers and a lot of testing. Moreover, there are other quality features, such as loading time[28], that require investments.

3.3.2 Which data are necessary?

To operate as a search engine, one needs a web **index**. To deliver high quality search results, one needs **search query logs** and data on **clicking behaviour** associated with search queries. Correlating these data results in more relevant search results, which in turn attracts more users and thus more data. Moreover, search queries inform web crawlers on when and how to update the index. The resulting learning effects are important for the development of search engine technology, along with other innovative activities. It is not clear if and how fast these feedback loops flatten out, that is, whether there is a minimum efficient scale of data and how difficult it is for others to realise this minimum efficient scale.

In this learning process, access to information on how the **search results** have been presented to the searcher is a "nice-to-have" because it allows for better interpretation of correlations between search queries and clicking behaviour. As such, contextual information on whether the results have been displayed on a desktop computer or a mobile device may help to interpret search queries. Furthermore, there is a causal relation between search results and clicking behaviour, which means that search results may be inferred from clicking behaviour and thereby may create technical opportunities for reversed engineering of search algorithms.

Some search providers service a market niche. For example, Startpage and DuckDuckGo target privacy-sensitive customers. The small scale of these niches prevents these companies from developing their own index or algorithm. The business model of these search providers depends on ‘reselling’ the **search results** of another search engine. Startpage relies on Google results only[29].

---

24 He at al (2017) find that “historical examples of related queries can be linked to seemingly rare queries by applying clusters and graph cutting techniques to the query-document bipartite graph his graph can be used to generate related queries and leverage historical examples that differ in minor ways from the target query.”

25 Google has developed a machine learning technology called *RankBrain* which helps in this process. See https://onlineidealab.com/google-rankbrain-what-is-it-how-can-you-benefit-from-it/

26 Consider the example of the nuclear power plant mentioned above. The crawlers of the search engine with a large global customer base receive earlier and stronger signals to start updating the index with relevant new items about this event (compared to the search engine with only a relatively small and geographically concentrated customer base).

27 From general statistics we that, as the number of explanatory variables (e.g. location, device, sex, or age) goes up, the degrees of freedom go down and the volume of observations (or the number of users of the search engine) must increase if we still want to find relevant and significant correlations.

28 During the interviews, respondents (other than Google) stated that ‘loading time’ is an important quality feature in the perception of end-users. Google happens to outperform others https://www.shoutmeloud.com/google-started-ranking-websites-based-on-load-time-and-speed.html

29 https://www.startpage.com/

DuckDuckGo says it operates as a so-called meta search engine, combining the search results of multiple search sources, though it mostly displays the results from Bing[30].

For the purpose of operating a search engine, some **contextual (personal) information** may be important (such as location, device used, previous search queries and clicking behaviour of the searcher). Other personal information is a nice to have, rather than an essential input (see textbox below).

Role of personal information in search services

Views differ on the value of personal data for search services. Some search engines (such as DuckDuckGo and Startpage) advertise that they do not personalise search results and advertisements and thereby preserve people's privacy and refrain from creating a 'filter bubble'. From interviews it became clear that the strategy of preserving privacy goes at the expense of advertisement revenues, because the click-through-rates of non-personalised advertisements are lower. Google also indicates that (besides contextual information such as a persons' location, past search queries, or the device used) it refrains from further personalising organic search results[31]. It does personalise advertisements, including advertised search results. Notwithstanding that some users and search queries benefit from personalised search results, ranking search results relative to personal profiles is not an essential quality feature. In fact, personalising organic search results may not be regarded a quality feature at all (on the contrary).

Access to personal profiles may be helpful but not essential for search engines that run an advertisement-based revenue model. Without personalisation, advertisements are sold by auctioning 'search queries'. When a searcher enters a query, he or she reveals an instant interest which can be matched with an appropriate advertisement without knowing personal information of the searcher. All in all, it seems that for the purpose of operating a search engine, access to personal information is a nice-to-have, rather than an essential input.

3.3.3 Can access to necessary data facilitate entry or strengthen competition?

Data is not the only entry barrier

One reason for the high concentration of the search market is the presence of entry barriers resulting from the need for high upfront and recurring investments that are largely sunk, including: *"investments in hardware, cost of indexing the web, human capital, cost of developing and updating the algorithm and IP patents"*.[32] In 2010, Microsoft claimed in the Microsoft/Yahoo merger case that entry into the market would require an investment of over USD 2 billion in hardware and human capital and several billions of investments in developing and updating the algorithm[33]. On top of that, a new entrant would have to bear "*significant costs related to the necessity to have a large database"*[34]. Such investments require deep pockets and the ability to realise sufficient scale. The former prevents candidates from entering the market as a new fully integrated search provider and the latter prevents already entered resellers targeting a specific niche (like Startpage or DuckDuckGo) from climbing the ladder of investment[35].

Furthermore, an advertisement-based business model may create another first-mover advantage. This is because such a business model is characterised by indirect network effects which, through investments, become circular. First, a growing number of searchers attracts a growing number of advertisers and growing advertisement revenues. The growing number of advertisers itself does not

30 https://help.duckduckgo.com/results/sources/?redir=1

31 https://searchengineland.com/google-admits-its-using-very-limited-personalization-in-search-results-305469

32 Case No COMP/M.5727 - MICROSOFT/ YAHOO! SEARCH BUSINESS, para 111

33 Ibid.

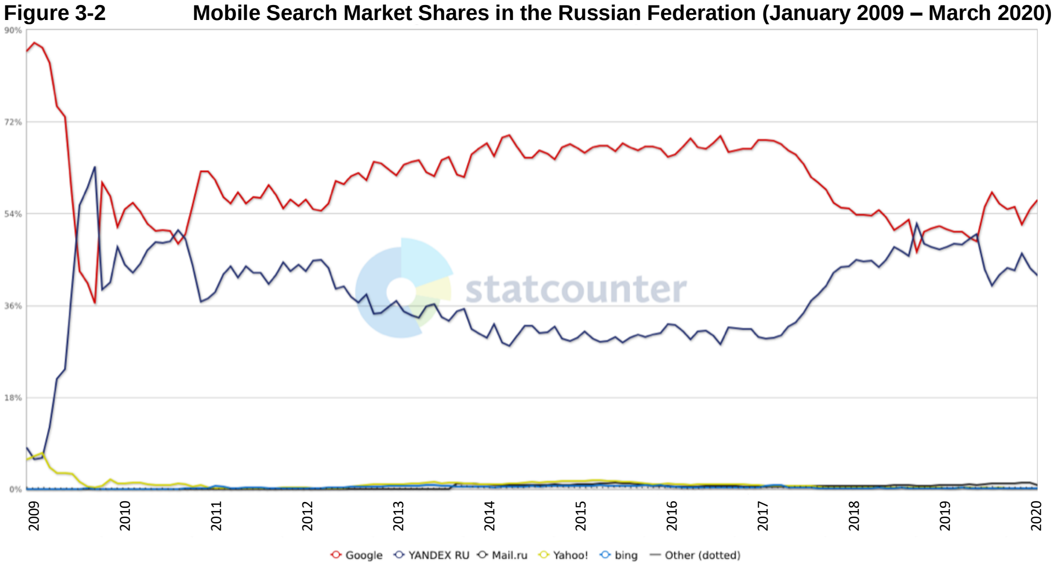
34 Ibid.

35 We have learned during interviews that, even when having access to the data, the cost of developing and regularly updating their own algorithms are too high for these niche players and the scale of the niche they service is too small.

necessarily attract more users. However, when advertisement revenues are invested in upgrades of hardware, indexing, human capital and updating the algorithm, this translates in a higher quality perceived by searchers. This attracts more searchers which in turn attracts more advertisers.

A third reason for high concentration may be related to user stickiness, possibly in combination with strategies of tying and bundling. These arguments were put forth by the European Commission in the Google Android case[36] and by the Russian antitrust authorities in a similar case against Google. These cases were about pre-installing Google Chrome and Google Search on Android devices, which was said to contribute to Google's market share in the search market (at least for mobile search services). In Russia, this tying practice has been resolved for several years since 2017 and it appears to have levelled the playing field in mobile search (see figure below).

Figure 3-2 Mobile Search Market Shares in the Russian Federation (January 2009 – March 2020)



Source: StatCounter Global Statistics

Data sharing may not be sufficient to strengthen competition

If data sharing does not result in entry, the question remains whether data sharing can strengthen competition among incumbents. It has been argued that, as a result of learning effects, "*a small initial data advantage can translate into a significant share of the user base and of the market. In the long run, this self-reinforcing dynamic can lead to market dominance"* (Anderson, 2020). This implies that data sharing would neutralise first-mover advantages and lead to stronger quality competition, i.e. to more investments in indexes and algorithms. Argenton & Prüfer (2012) argue it does and they propose a general and reciprocal obligation for search engines to share data. Reciprocity is necessary, they argue, because an asymmetric obligation will unlevel the playing field for the first mover and tip the market in favour of the second mover[37]. Argenton & Prüfer argue that reciprocity in data sharing will ensure that all search engines compete fiercer by improving their algorithm, rather than by exploiting network effects and first-mover data advantages[38].

---

[36] Case AT. 40099 (Google Android)

[37] Consider the size of the database of the 1st mover as N1 and of the second mover as N2 and consider that N1>N2. Because N1 is larger, the quality of the 1st mover's search results are better and its market share is higher and growing, and the market will eventually tip in favour of the 1st mover. If the 2nd mover is granted access to N1, his database grows to N1+N2 which is by definition larger than N1. If data sharing is not reciprocal, the 2nd mover will see its search results become better than the results of the 1st mover. Over time, the 2nd mover will overtake the market share of the 1st mover and the market would eventually tip in favour of 2nd mover.

[38] "*no search engine can rest on its merits because the only way to sustain its market share and profits is to invest all efficiency gains that come from the exploitation of network externalities in better quality."* (Argenton & Prüfer, 2012).

Before going into the logic behind the reasoning of Argenton and Prüfer, we note that theories supporting the positive effects of data sharing on competition and innovation abstract from the possibility that data sharing may create opportunities for reversed engineering. In that case, the incentives to invest may decline for all parties involved (see also section 3.3.6 below).

The conclusion that data sharing results in positive effects for competition and innovation assumes that data feedback loops and scale effects in the learning process are dominant determinants of quality and competition. However, other factors may also be important determinants of quality and competition. These include investments in innovative activities such as updating the index more efficiently and more frequently[39] [40], developing a separate index specifically for mobile search[41] [42], human-based innovations and testing of search algorithms[43] [44], or speeding up loading time[45]. Moreover, a search engine's leading position may also be explained by strategic behaviour like tying and bundling (as in the above example of mobile search in Russia).

To examine the relative importance of data feedback loops for competition, we discuss two natural experiments in different language areas: English and Russian. Note that search services are language-dependent; this involves the index (because websites provide content in different languages) as well as the algorithm and the associated feedback loops (because search queries are entered in different languages). Further note that the Russian language is fundamentally different from the English language, amongst others, because of the use of a different alphabet[46].

The first experiment concerns the merger between Bing and Yahoo in the United States, which took place in 2010. The merger resulted in an overnight doubling of the query volume for the parties involved. While Bing and Yahoo focus primarily on the English language, one would expect this merger to have some positive impact on the market shares of Bing and Yahoo in (at least) the United States and the United Kingdom. We call this Hypothesis 1. Figure 3-3 below shows that this hypothesis is rejected, since the merger in 2010 had little to no impact on the market shares in the United States[47].

---

39 https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

40 https://www.brafton.co.uk/news/seo-case-study-sites-see-more-pages-indexed-by-google-than-bing-even-post-panda-800527170/

41 https://webmasters.googleblog.com/2016/11/mobile-first-indexing.html

42 http://www.thesempost.com/bing-search-will-not-mobile-first-search-index/

43 See section 3.3.1

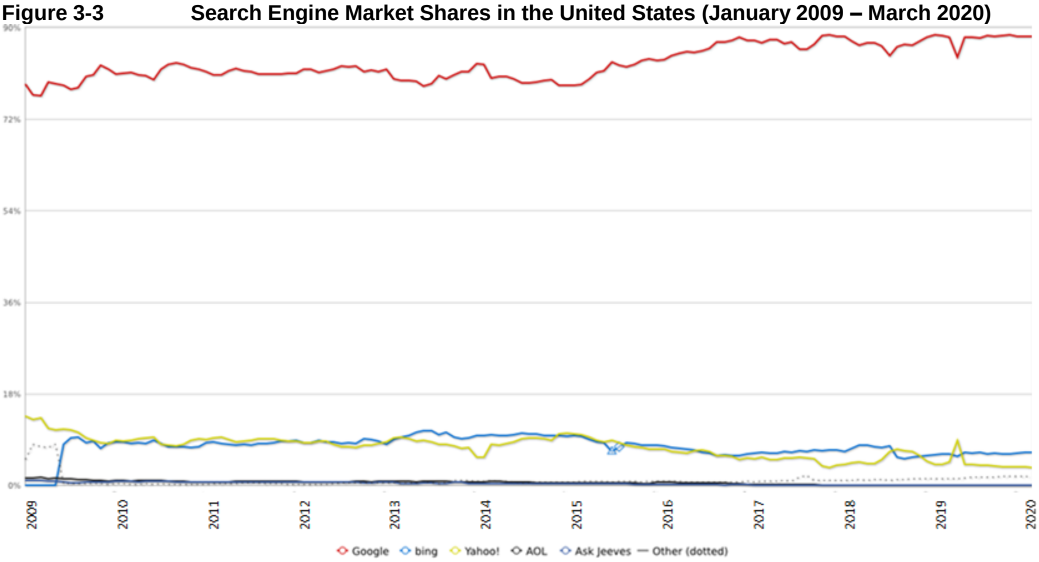
44 https://searchengineland.com/bing-search-quality-rating-guidelines-130592

45 During the interviews, respondents (other than Google) stated that 'loading time' is an important quality feature in the perception of end-users. Google happens to outperform others https://www.shoutmeloud.com/google-started-ranking-websites-based-on-load-time-and-speed.html

46 https://www.nytimes.com/2006/12/18/technology/18google.html

47 In the U.K. the market shares are even more stable over time as in the U.S. (see https://gs.statcounter.com/search-engine-market-share/all/united-kingdom)

Figure 3-3 **Search Engine Market Shares in the United States (January 2009 – March 2020)**



Source: StatCounter Global Statistics

The rejection of hypothesis 1 could mean that the damage (in the form of market tipping) had already been done at the time of the merger and that merging the data sets of Microsoft and Yahoo could not revert this process. It may also mean that the statistics on market shares provide a misleading picture of the volume of data available to Microsoft and Yahoo at the time of the merger. At the time, Microsoft was able to observe search data through Internet Explorer[48], which was the most frequently used browser in desktop environment in the U.S. (with a market share of 70% in 2009)[49]. Accordingly, Microsoft should have been able to gather a significant volume of data on search queries and clicking behaviour (63% to be specific)[50]. The rejection of hypothesis 1, however, could also mean that data feedback loops are not as important as some have argued, and that Google's leading position must be explained by other factors.

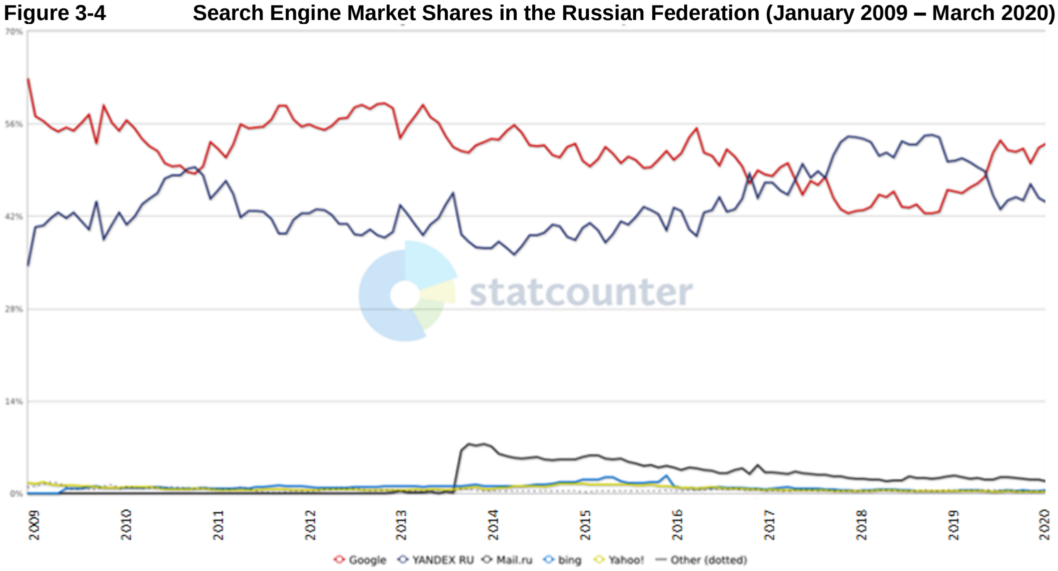
The second experiment is the entry of Google into the Russian market in 2006. Yandex, founded in 1997, already occupied a leading position as Russian's national search engine. If language dependent feedback loops are a dominant driver of a search engine's market share, one would expect Google to fail in gaining any market share in Russia as a result of a second-mover disadvantage. (We call this hypothesis 2a). If feedback loops are not language dependent, however, one would expect Google to crush Yandex because of its global scale advantage. (We call this hypothesis 2b). The figure below shows that both hypotheses are rejected. In other words, Google gains market share but fails to crush Yandex.

48 https://www.businessinsider.com/bing-is-cribbing-from-google-search-results-2011-2?international=true&r=US&IR=T

49 https://gs.statcounter.com/browser-market-share/desktop/united-states-of-america/#monthly-200901-202003

50 If 90% of the market is using Google's search engine, but 70% accesses the Google website via Microsoft's web browsers, one would expect Microsoft to be able to record 63% of all the search queries and clicking behaviour in the American desktop environment.

Figure 3-4 **Search Engine Market Shares in the Russian Federation (January 2009 – March 2020)**



Source: StatCounter Global Statistics

The observed difference between the development of market shares in the U.S. and in Russia is remarkable and cannot be explained by governmental preferential treatment[51]. Indeed, language is important in search engine technology and it did provide Google with serious challenges when entering the Russian market[52] [53]. Google has successfully taken up this challenge but failed to leverage its presumed global scale advantage in data. This implies that data feedback loops are perhaps not a dominant driver of competition, and that other determinants of Google's competitive position are at least equally important. For example, in 2009, Yandex introduced a new machine-learning algorithm MatrixNet which was said to contribute to the restoration of Yandex's market early 2010's[54]. Moreover, a strong indication that anti-competitive tying and bundling is a major determinant of Google's leading position can be found in the earlier mentioned fact that Google's market share in the Russian mobile search market declined rapidly after the Russian competition authorities obliged Google to untie its Chrome browser and search engine from its Android operating system.

All in all, there is strong evidence that data sharing alone is not responsible for the observed unbalanced competition in the search markets in the U.S and Europe. It is thus not obvious that data sharing will lead to a more balanced competition. Experience in Russian indicates that antitrust enforcement is important (and may already be sufficient) for restoring competition in the search market.

3.3.4 Are there bottlenecks for accessing necessary data?

Because of the high and recurring (sunk) investments in developing and maintaining a high-quality index, there are only a few companies that invest in crawling and indexing (e.g. Google, Microsoft, Yandex and Baidu). There are indeed public libraries and open source crawlers (such as Common Crawl). However, during interviews it was explained that their scale and (more importantly) their quality is insufficient for setting up a search service that can effectively compete with the four incumbents. Access to one of the main indexes may be considered a bottleneck (or at least an entry barrier requiring substantial efforts and time to overcome) for supplying search services.

---

51 As could be argued in China.

52 https://www.cnbc.com/2019/01/18/yandex-is-beating-google-in-russia.html
https://www.nytimes.com/2006/12/18/technology/18google.html

53 Although Yandex is on occasion outperforming Google even in the English language (see https://www.youtube.com/watch?v=5HOw01rpG20)

54 https://www.raamoprusland.nl/dossiers/economie/786-how-even-yandex-is-being-brought-under-control

It has been argued that companies other than Google fail to realise the necessary scale in data on search queries and clicking behaviour because of Google's data-related first mover advantage. We noted earlier that the relative importance of these scale effects is not robustly supported by empirical evidence. Although search data can also be collected through alternative sources (such as web browsers and browser plug-ins or extensions), competitors have increasingly been cut off from browser originated data by Google's growth in the browser market[55], possibly supported by anti-competitive conduct[56]. As a result, during the last decade it has become more likely that access to data on search queries and clicking behaviour forms a bottleneck.

Contextual data on location and devices are generally easily accessible. Data on past search behaviour and clicking behaviour at the individual level, however, are not.

3.3.5 Is necessary data shared (voluntarily, on a commercial basis)?

There are no commercial agreements that make any of the four major indexes accessible for search algorithms of third parties. This does not mean that the owners of these indexes are unwilling to provide commercial access. Given the asymmetric distribution of market shares in the retail market for search services, it seems logical for smaller parties (like Bing and Yandex) to be willing to provide such access in return for an appropriate price or in return for data on search queries and clicking behaviour. Such an access service would provide them with additional wholesale revenues or increased access to data. However, it appears there is little to no demand for accessing proprietary indexes because the required investments in the development and maintenance of algorithms are too high (as was confirmed during interviews).

Data on search queries and clicking behaviour are not shared with third parties. During interviews, small niche players expressed no interest in having access to search data and larger players (plural) stated that they are hesitant to share search data on a large scale because of privacy considerations and risks of revealing trade secrets through reverse engineering.

Search results are shared based on commercial agreements. These sharing agreements are made with (small) niche players that complement the main service of the access provider (usually, a service that markets high privacy, e.g. DuckDuckGo using Bing and Startpage using Google). The risk of reverse engineering is perceived as small by the providers of search results because access seekers lack the resources to finance such research. Also, the commercial agreements explicitly prohibit reverse engineering.

3.3.6 What is the impact of mandatory data sharing on innovation incentives?

Section 3.3.3 explains that providing access to search data is unlikely to invite new entry by small-scale search providers. Moreover, it would not spur current resellers of search results to invest in their own algorithms or index because they lack the necessary scale for recouping these investments.

Argenton & Prüfer (2012) show, in a theoretical model, that reciprocal sharing of data between existing search providers proposal could spur innovation, search quality, consumer surplus, and total welfare because, as a result of it, *"no search engine can rest on its merits because the only way to sustain its market share and profits is to invest all efficiency gains that come from the exploitation of network externalities in better quality."* This conclusion assumes that data sharing results in more balanced competition, which is not evident (as argued in section 3.3.3). Moreover, this conclusion is based on the notion that sharing does not negatively affect the incentives to innovate because the data on search queries are assumed to be by-products of previous

---

55 See https://gs.statcounter.com/browser-market-share/desktop/united-states-of-america/#monthly-200901-202003 and https://gs.statcounter.com/browser-market-share/desktop/europe#monthly-200901-202003

56 See Case AT. 40099 (Google Android)

production processes and not the result of dedicated investments. This reasoning may not be correct when access to search data on a large scale allows others to reverse engineer the algorithm or the search results of the current market leader. Copying search results basically boils down to updating one's index based on observed combinations of search queries and clicking behaviour, rather than on examining the true content of a website by crawling it.

The risk of reverse engineering is illustrated by an episode in 2011, when Google conducted an experiment to verify its suspicions of Microsoft reversely engineering Google's search results based on data retrieved via its web browser (Internet Explorer)[57]. The experiment involved a manipulation by Google of its index, relating non-existing words to random websites. Next, Google instructed 20 employees to regularly visit the Google website using Microsoft's Internet explorer, to search for these non-existing words, and to click the results. Google claimed that the 'fake results' started to show up in Bing's search results only after a few weeks. At the time, Microsoft confirmed that it was collecting search and clicking data through its browser and toolbar but denied that it was copying Google results and argued that there are alternative explanations for the outcome of Google's experiment, but did not go into further details[58].

Next to possibly harming innovations in search technology, data sharing obligations may take away incentives to invest in alternative services or products through which search data may be collected, notably web browsers and browser extensions.

3.3.7 What are the risks of data sharing?

Besides risks of reverse engineering and associated negative incentives to innovate (see above), there are possible privacy risks. Data on search queries and clicking behaviour are stored in combination with many personal identifiers (such as IP address, cookies functioning as ID, browser fingerprints, account names, et cetera). Sharing this raw data creates privacy risks. Moreover, it may not be necessary as the data on search queries and clicking behaviour could be anonymised before being shared.

It has been argued that data can always be de-anonymised. An often-cited example is the story of New York Times journalists being able to re-identify 'Searcher No. 4417749' from anonymised AOL search logs[59]. However, the search results in this example were not stripped from identifiers, they were simply replaced by a number allowing to single out an individual. In other words, the data was *pseudonymised* rather than anonymised. Without this identification number, however, and without any other identifiers or contextual information (not even data on location or demographic variables), the data would be fully anonymised and only exist of sets of search queries and subsequent clicking behaviour. In that case, it would not have been possible to single out an individual, nor would be it be possible to trace back the search queries to a specific person.

Data can be anonymised to different degrees[60], but the more it is anonymised, the less useful data tends to become. For example, clicking behaviour may be misinterpreted without any contextual information such as location, device used, and a person's past search behaviour. Adar (2007) provides certain solutions to balance privacy and research needs. Over time, this balance is difficult to maintain, however, as the possibilities for re-identification seem to be increasing over time (Article 29 Working Party, 2014). Rocher et al. (2019) show that re-identification is possible even with a minimum of information about personal characteristics.

57 https://googleblog.blogspot.com/2011/02/microsofts-bing-uses-google-search.html
58 https://www.businessinsider.com/bing-is-cribbing-from-google-search-results-2011-2
59 https://www.nytimes.com/2006/08/09/technology/09aol.html
60 https://searchengineland.com/google-anonymizing-search-records-to-protect-privacy-10736

3.3.8 *What are the options under current legislation to make data sharing mandatory?*

We have established that search data is indispensable for search engine operators and that there may be a bottleneck for accessing such data. Some have argued that mandatory data sharing would increase innovation-based competition. We have concluded above that this does not necessarily lead to more competition, and that it may also lead to less innovations. Having said that, we explore below the options under current legislation to make data sharing mandatory.

The GDPR applies to search engines and search data

Where data on search queries and clicking behaviour can be traced back to an individual, for instance by linking it to an IP address or cookie or by storing it in a profile when users are logged-in to the search service, it qualifies as personal data. This means that the GDPR provisions apply, including the right to data portability.

The personal data included within the scope of application of the right to data portability is personal data 'provided by the data subject'. Search queries are actively inserted by the user and will thus fall within this definition (unless they are stripped from personal identifiers).

A user's clicking behaviour would qualify as 'observed data', which the Article 29 Working Party has argued to be covered by the GDPR's right to data portability. Indeed, the Article 29 Working Party makes explicit reference to search history as a form of observed data falling within the scope of Article 20 GDPR: 'Observed data provided by the data subject by virtue of the use of the service or the device […] may for example include a person's search history' (Article 29 Working Party, 2017).

Effectiveness of the right to data portability under the GDPR is limited

When technically feasible, Article 20(2) GDPR would require that personal data is directly transferred between search engines. However, the GDPR's right to data portability can only be used to port the search history of users one-by-one at their request. This means that the effectiveness of the right to data portability as a tool for mandating data sharing in the sector depends on how actively individuals invoke it.

A second limitation is that it is still unclear to what extent the GDPR's right to data portability can be interpreted as constituting a real-time mechanism with continuous data exchanges between data controllers until the data subject revokes his or her request, or solely provides for one-off portability where personal data is transferred only once at the moment the data subject makes the request. Real-time portability seems more effective to take away bottlenecks in the market than one-off portability.

Note that GDPR cannot compel reciprocity in data sharing, this can only be realised when requested by the individual.

Essential facility doctrine might apply

Another possibility for mandating sharing of data on search queries and clicking behaviour is to rely on the essential facilities doctrine under competition law. A refusal to share data on search queries and clicking behaviour will qualify as abuse of dominance under Article 102 TFEU when the company controlling access to data is dominant and when the following four 'exceptional circumstances' are present: indispensability, exclusion of effective competition, prevention of the introduction of a new product, and absence of objective justification.

- There are clear legal precedents that indicate a dominant position held by Google in the market for online Search.[61]

[61] See Commission's assessment in the *Google Shopping* and *Google Android* decisions

- There are no clear legal precedents on the issue of indispensability of data on search queries and clicking behaviour held by one party. In past merger cases (e.g. Google/DoubleClick) the European Commission argued that Google's data on users' search behaviour with DoubleClick's data on web-browsing behaviour of users did not give the merged entity a competitive advantage that could not be matched by competitors. While this may have been the case at the time (see section 3.3.1), we note that this may no longer be the case today (see section 3.3.4).
- Moreover, there is discussion about how strict the condition of indispensability should be applied to cases of refusals to share data which are "*generated virtually incidentally and without special investment"*. Schweitzer et al. (2018) advocate a less strict application of the condition of indispensability in such case. Argenton & Prüfer (2012) and Prüfer & Schottmüller (2017) state that this condition applies to data on search queries and clicking behaviour. However, as we argue in section 3.3.6, this statement is debatable because people click on search results presented to them which are an output of investments in the index and the algorithm.
- When lowering of existing standards is considered appropriate due to the characteristics of data and the market conditions in online search, competition authorities get the discretion to adapt the application of the competition rules to new circumstances so that there is room to set a new precedent based on the existing legal framework.

Essential facility doctrine may require creative interpretation of the law

When data sharing is desirable, a sharing obligation should be reciprocal according to Argenton & Prüfer (2012). Under competition law, however, only dominant search engines could face an obligation to share information about previous queries[62]. However, it is worth examining whether reciprocity can serve as a form of reasonable price for data access. In that case, competition law could still provide a legal basis for enforcing access. Reciprocity would then at first only apply between the largest player and access seekers, whereby the largest player would still retain a data advantage[63]. However, the other players could reduce this lead by also sharing their data on a reciprocal basis[64]. This would not require an obligation, as they all benefit, but an explicit exception for this cooperation should be included in Article 101 TFEU which prohibits horizontal cooperation between competitors.

Sharing of search data may conflict with Database Directive and/or Trade Secrets Directive

We note that the disclosure of queries and clicks automatically discloses search results on which a user clicked for a particular query. While this is not inferred or derived data, it is data which is produced by a search service provider and results from investments in an index and an algorithm.

---

62 In a recent article about data sharing in general (not specifically related to search engines) Prüfer mitigates his claim on general reciprocity and concludes that only *"the largest two or three providers are obliged to share [provided they have] market shares larger than 20 or 30 per cent"*. Reasons for this are that: i) small platform may otherwise experience disproportional administrative burden, and ii) because it is said that large firms may benefit more from data sharing than small firms because more likely to have access to other sources of information that complement user information from this market (Prüfer, 2019). In our view, these arguments are challengeable. The claim that small platforms would experience a disproportionate administrative burden appears contradictory to Prüfer's own statement that "*sharing of user data is technically and organizationally possible at a large scale and automatically*" (Prüfer, 2019, p. 11). Moreover, if there is such problem, it can be addressed by giving such platforms some time for growth before the obligation become effective. Second, the claim that large firms may benefit more from data sharing than small firms is disputed by Mayer-Schönberg and Ramge (2018). They argue that when *"since smaller firms would have less data to share and machine-learning algorithms produce diminishing returns for each new data point, a company like Amazon would gain far less than its smaller competitors. A data sharing mandate would lift all boats, but to different degrees"* (Mayer-Schönberg & Ramge, 2018, p. 53). Finally, imposing a sharing obligation only on *"the largest two or three"* providers, provided they have "*market shares larger than 20 or 30 per cent*" seems arbitrary: why not the fourth largest player? why not providers with 10 or 40 percent? A suggestion like this inevitably leads to a test of dominance as to avoid arbitrariness. A need to establish dominance would make it more difficult to impose data sharing obligations and possibly results in a repetitive leapfrogging of dominance held the second and first mover (see footnote 37).

63 If A is obliged to share data with B and C on a reciprocal basis, then A would still have a larger data set (of A+B+C) then B and C (of B+A and of C+A, respectively).

64 If B and C decide to voluntarily share their data as well, then all parties have access to the larger data set of A+B+C.

As such, obligations to share search results (or data which may be used to infer search results) may conflict with the protection offered by Database Directive and/or the Trade Secrets Directive.

3.3.9 What are other options for stimulating data sharing?

The European Commission's decision in the Google Android case effectively increased the choice of end-users between mobile browsers. This may allow competitors to increase their market share in the browser market, thereby increasing their ability to collect search data. A similar measure seems to have restored the balance in the Russian search market (see Figure 3-2). However, it is not sure that the same effect will be observed in Europe (this remains to be seen).

An alternative to imposing mandatory data sharing is to support or encourage pooling of anonymised search data by various smaller search engines, browsers, apps, and developers of innovative browser plug-ins and extensions. Such pooling arrangement may require an exemption from Article 101 TFEU on horizontal agreements; possibly there is a need for new block exemptions for data pooling arrangements (Crémer, de Montjoye, & Schweitzer, Competition policy for the digital era, 2019, p. 98). When governed by a proper code of conduct, governments could offer endorsements to data sharing pools and companies that participate in them. Moreover, end-users may be asked to give consent to sharing personalised search data between participants of the data pool. Advocates of general data sharing obligations may argue that as long as Google does not participate in such a pool, the majority of search data remains exclusively at Google's disposal and therefore the measure would not help others overcome their second-mover disadvantage.

Data pooling arrangements might be further stimulated by encouraging data mobility networks such as the Data Transfer Project involving Google, Facebook, Microsoft, Twitter, Apple, and several smaller digital platforms[65]. Data mobility initiatives could be fostered by mediating discussions about the type of data within the scope of the project, establishing rules on reciprocity, and to investigate the possibility of continuous data transfer.

Any measure aimed at stimulating or mandating sharing of search data should be accompanied by measures that, albeit difficult to enforce, discourage reverse engineering of search results.

---

[65] Including Deezer, Mastodon, SmugMug, and Solid.

4 Case: social networks

4.1 Case description

A social network (or social networking service) is an online platform where users connect and interact with other users who share certain interests, professional backgrounds, hobbies, or real-life connections. There are many types of social networks, as well as platforms that include a social networking functionality. Facebook, with consumer functionalities that gravitate toward making connections and sharing "posts", is the prime example of a social network or social media. Other online platforms combine a distinct service with social networking functionalities. An example is Goodreads, a book cataloguing website where users can search for books and reviews, write their own reviews, and create their own reading lists, which they can share with other users. Thus, Goodreads provides (user-generated) content for book lovers with more narrow interactions among readers. Another example is Strava, a platform and app for tracking cycling and running exercises with GPS data. The app records activities which can be shared with followers as well as shared publicly.

The volumes of data collected by social networks can be enormous, as illustrated by the size of the archives of a *New York Times* reporter's data (all numbers refer to a single person's archive size):[66]

- Facebook's archive measured about 650 megabytes (equivalent to 160 hours of music);
- Google's archive for this person was about 8 gigabytes (about 2,000 hours of music) with an extraordinary level of detail;
- special-interest social networks may be different: LinkedIn's archive was less than 0.5 megabyte, containing what one would expect (spreadsheets of LinkedIn contacts and information that the user added to his account).

Social networks do not only gather data, they also share it with business partners. However, in the aftermath of the Cambridge Analytica scandal, platforms like Facebook and Twitter have implemented restrictions to data access via their APIs.

This case pertains to horizontal and vertical data sharing. The initial hypothesis was that (established) social networks are not eager to share their data at all; neither with horizontal competitors (for obvious reasons) nor with downstream companies, especially because they could extend their business value proposition into upstream territory. The latter would intensify horizontal competition. Thus, providing downstream data access can invite entry into the upstream market. Furthermore, according to OECD (2019), the provision of data can require significant investments, reducing incentives to share data. This implies that if data sharing takes place, there must be sufficient possibilities to appropriate the returns on the up-front and follow-on investments, in particular in complementary resources such as data models, algorithms for storage and processing, and secured IT infrastructures. On top of that there are legal issues related to privacy regulation, intellectual property rights and differences between jurisdictions that make parties hesitant to share data.

From the outset we expected that the initial hypothesis would only be partially confirmed. The reason is that one can observe that social network voluntarily share data, although selectively. An example of voluntary data sharing is provided by the social login service ("log in with ..."), which is an easy way for companies to get consumers to register on their websites and apps, and by doing so, collect information about them. Facebook, Google, Twitter and LinkedIn are examples of

[66] "I Downloaded the Information That Facebook Has on Me. Yikes.", *The New York Times* 11 April 2018.

networks providing this login service. It involves data sharing in two directions. The social networks provide data to their partners and receive data from them in return. Interestingly, Facebook used to forbid close connections with (horizontally) competing social networks through its data access conditions. In 2018, however, the company reviewed its policy to block apps built on top of its platform that replicated its core functionality and decided to remove this policy.

A central question is what kind of data sharing agreements currently exist. A related question concerns the desirability of such commercial arrangements: do they benefit consumers? Do they strengthen the market position of social networks? Do they introduce privacy risks? And more generally, do they require regulation to keep them in line with public goals?

4.2 Preliminary analysis

Business model

One can distinguish general-purpose social media, such as Facebook, from social networks dedicated to specific groups, for instance athletes. They have in common that they are typically based on digital platform business models that occupy a central spot in a multi-sided market. Such platforms aim to serve different types of users, for instance consumers and advertisers. More generally, as platforms they arrange, coordinate, and orchestrate interactions between users and across different user groups. A social network's crucial assets include its community and what its members contribute.

The typical business model in social networks aims at generating revenues from data or user attention, often through advertising or viewer targeting. Thus, a social network often is an attention platform, just as search engines and many online news media.[67] Users interested in harvesting attention can be advertisers but also companies and organisation using social media to communicate with customers and stakeholders. Facebook, for instance, offers tools to firms that provide them with information on user engagement with their Facebook pages and their websites. Social networks (e.g. Facebook) may also provide data and insights on social media activities for a broader audience, such as academic researchers.

Social platforms rely on direct and indirect network effects to become successful:

1. Direct network effects drive the size of the customer base: users attract other users because of the increased number of potential interactions among them.
2. Indirect network effects drive profitability: users attract platform users interested in user data, access to user data, or user attention.

Better and more personalized user experiences stimulate these network effects.

Data

A large number of types of data are used to improve the quality of matchings, the relevance and value of user interactions and consumer targeting, and the platform's functionalities. For these purposes, social networks need data about users, such as their social graphs, communication data, and profile data containing information about individual characteristics, preferences, behaviour, and interactions. Hence, they collect various types of data from a broad range of sources.

First, they harvest data "at the surface" where user interactions take place, in particular user-generated content such as status updates and pictures posted by users; users' personal data such

67 Evans (2019) gives the following definition: "Attention platforms supply content to consumers who spend time on their properties" (p. 776).

as name, birth year and address; users' behavioural data such as location history, browsing history and search history, on the social network's website as well as elsewhere; data about connections with other users (the social graph). Most users will understand that this type of data is necessary to support the key user functionalities and may thus be considered provided data. However, users may not always be aware of the data that is collected because the data may also be observed on a very large number of websites other than the social network's own website (e.g. through cookies and invisible pixels that load in the browser to collect information about browsing activities).

Furthermore, social networks may observe data "below the surface" of the basic user interactions, such as: activity logs; inferences from all data types; users device information such as operating system, hardware version, device settings, battery strength, device identifiers; user device locations, including specific geographic locations, such as through GPS, Bluetooth and Wi-Fi signals; connection information such as ISP, browser type, language and time zone, phone number and IP address; non-personally identifying data such as demographic information, preferences and ad interests. Again, such data may also be collected also on websites other than the social network's own website.

At a higher level one can distinguish (i) data about and obtained from users; and (ii) a broad range of data about people who are not users of the social network, also known as shadow profiles. The latter type of data is typically collected on websites other than the social network's own website. Turning off the collection of shadow profile data may only be possible after making a user account, or by installing browser plug-ins that protect one’s privacy.

A more extensive (but possibly non-exhaustive) data typology with relevance for social networks is as follows, noting that several (if not all) categories include data on people without user accounts:
1. primary personal identifiers (name, address, age, phone number, bank account number, etc.);
2. secondary personal identifiers (IP address, data from trackers, cookies, device, technical parameters w.r.t. hardware and software, etc.);
3. data about individual behaviour and interactions on the social network (user-generated content, user activity logs, user settings, etc.);
4. data about individual behaviour and interactions elsewhere (browsing, app usage, online shopping, etc.);
5. social graph within the platform's ecosystem (data about connections, followers and friends on the social network);
6. broad social graph (data obtained from address book, phone calls, outside interactions, etc.);
7. usage context (time, time zone, geographical location, geographical proximity of other users, etc.);
8. network connections (technical parameters w.r.t. Wi-Fi, Bluetooth, ISP, internet routing, etc.).

Social graph information (data types 5 and 6 above) ranges from:
- **narrowly defined social graph data:** at the surface, social graph data pertains to connections, friends, follows likes, etc.;
- **broad/deep without limits social graph data:** below the surface, a social network may collect data about the nature of interactions between connected users (e.g. time, frequency, content, etc.). From such data, the platform may know in detail whether people are friends, colleagues, lovers, sharing a hobby, playing sports matches together, sports teammates, and so on. The platform may also know in detail what type of communication they share, and how their characters match.

Algorithms
The functionalities and technologies that support a platform's activities determine the quality of matches between users and, in turn, the value of their interactions or transactions. A digital platform

learns how to improve interactions from the earlier interactions it facilitated. Accordingly, learning effects create feedback loops. For instance, social networks select and prioritize the posts that users view in their "feeds", suggest new connections to users, and more generally, stimulate user engagement and time spent on the platform. In similar ways they stimulate relevant cross-platform interactions between users and advertisers, or between users and apps within the ecosystem. The behavioural responses by the various types of users generate insights for improving the matching process. Important goals to which underlying algorithms contribute are growth of its user base, the maximization of user engagement through same-side and cross-platform interactions, and revenue generation from these interactions.

Social networks, like many other digital platforms, do not have a "single algorithm", but rather a wide-ranging system of automated intelligence supported by human intelligence[68]. Applications and functionalities range from simple reporting and business intelligence measurements to more advanced machine learning apps, process large quantities of data. In addition, thousands of developers are carrying out ad hoc data analyses on a daily basis. Hence, widely used social media use a tremendous and rapidly growing data volume, which is also due to the increasing number of users and the expansion of functionalities on the platform. Billions of users log in to Facebook on a regular basis to connect and share content with other users; they upload billions of posts on a daily basis[69].

Innovation

Social platforms stimulate innovation in various ways:

1. As discussed above, a social network's innovation is above all driven by learning from user interactions and the resulting feedback loops that drive its growth. Thus, a company has its own developers who are working on innovation on a continuous basis.
2. Many digital platforms also stimulate innovation by third-party developers active within their ecosystems. Think of the large number of third-party apps available on Facebook's platform.
3. Social networks may cooperate, through data sharing agreements, with downstream service providers that use the platform to connect with and serve consumers. An example is the sports app Relive (previously) using data from the social network Strava as an input for its own functionalities. Here, the degree of a social network's openness, also in terms of access to data, influences the scope for downstream innovations developed by third parties.

An implication is that more data is almost always better, even if one abstracts from learning possibilities. In particular:

1. The importance of innovation is a key driver of a social network's data hunger. Each additional piece of information may potentially contribute to the quantity and quality of inferences that can be drawn from data points. Therefore, a company like Facebook has strong incentives to collect as much data as possible.
2. At the same time, social networks have strong incentives to share data as much as possible. This is due to the importance of innovation by third-party developers and downstream service providers. What they develop can add value to the platform's ecosystem, and hence attract more users, advertisers, and so on.

Having pointed out the importance of data for innovation, it is crucial to understand that human creativity is of utmost importance. This is because innovation is not only about finding more and more correlations between data points. It also involves figuring out new and complementary functionalities on and off a social network's platform.

---

68 Thusoo et al. (2010).

69 Feitelson et al. (2013).

The nature of innovations by and on a platform makes social networks different from regular companies, including traditional software manufacturers. The latter type of firm develops a "finite" product, that is, with a delineated scope and a planned completion date. Thus, a regular company typically employs a "linear" development process. Platforms like Facebook tend to be in a "perpetual development mode", if only to keep up with rapid growth of their user bases and user activities[70]. This can be observed in continuously increasing development activities and codebase, an ongoing trend at practically all internet-based companies. There is no grand plan aiming at a final, well-specified social media service. Facebook, for instance, deploys new code on a daily basis day, balancing fast development with good judgment and monitoring. Each new piece of code is reviewed by another engineer, which has as a side effect that knowledge about coding and also the code itself is quickly and continuously shared within the community of engineers.

Twitter illustrates that collecting, processing and analysing high data volumes go hand in hand with an increasing number of people[71]. Its analytics platform (performing the data mining of its big data sets) has experienced high growth over the years, dealing with increasing complexity, an increasing number of users, and more variety of use cases. In 2010, Twitter employed about 100 employees (including an analytics team of four). In 2013, the company already had 2,700 employees (its current size is about five thousand) and was dealing with a daily raw data intake of around one hundred terabytes[72].

A push process of adding new code with a high frequency balances innovation speed with risk control. This is where the organizational culture matters, as illustrated by Facebook[73]. The increasing scale in the team of engineers, the volume of code, and the user base goes hand in hand with growing risks. The higher number of interactions implies that more mistakes and errors can occur. These risks cannot be avoided, but they can be managed. For instance, functionalities that relate to user privacy will be held to a higher standard. A culture of personal responsibility then contributes to quality control: developers have to support the operational use of their code, which gives them a stake in the process of software development. They also get a lot of discretion with regard to the ideas and projects they work on, and the teams they work in. Accordingly, the culture is geared towards freeing up creativity in a decentralised fashion, relying on the efforts and ideas of individuals and teams, and without specifying products in advance.

4.3 Analysis of research questions

4.3.1 Which role does data play in the business model?

The short answer is that almost all data about users add value to social networks. This is because a wide range of data types can be used to figure out how to grow the user base, how to let users interact and engage more, and how to generate revenues from their activities and the follow-on data flows. A wide (virtually unlimited) range of data forms a continuous stream of input the learning, development and innovation processes. Of particular importance for social networks are:

- **user-generated content**, providing the primary value for other users (content to consume) and for users themselves (i.e. “stored value” through the personal posts history);
- **social graph data**, describing connections with other users (and possibly also with non-users, in case data on social relationships can be observed or obtained somewhere);
- **profiling data**, containing information about user characteristics and preferences;

---

70 Feitelson et al (2013).

71 Lin and Ryaboy (2013).

72 There were about 2,700 employees in 2013, 3,900 in 2018, and 4,900 in 2019. https://en.wikipedia.org/wiki/Twitter (retrieved 19 April 2020); https://www.statista.com/statistics/272140/employees-of-twitter/ (retrieved 19 April 2020).

73 Feitelson et al (2013).

- **behavioural data**, collected by observing what users and non-users do on and off the social network's platform.

Network effects, matching and user interactions

The essence of social networks lies in the functionality to form connections with (and follow) other users. These links provide users with access to each other for communicating and sharing content (e.g. posts and pictures) with each other. A user's set of established connections, the "social graph", forms a key input to the functionalities that support user interactions.

Social networks aim at growth of the user base, the maximization of user engagement (see below) and interactions, and revenue generation from engagement and interactions. Network effects form the basis for these goals. Therefore, the benefits of data are driven by:

1. Direct network effects: with a large user base, a platform can collect large amounts of various types of data from users, as well as from persons who do not use the social network. More, better and richer data increase revenues, and may create new profit opportunities.
2. Indirect network effects: the collected data may contribute to revenues through, for instance, advertising and the selling of slices of data. Instead of directly selling data, a social network's business models may sell access to user attention, or the possibility to target users.

Third parties, such as developers of apps on a social network site, may directly or indirectly benefit from the social graph and associated data.[74] For instance, regarding the second point above, Facebook has apparently never sold user data, but instead selectively grants other companies with partial access to its platform as long as this supports its own business model.[75]

User engagement

User engagement is essential for the collection of user-generated content, social graph data and behavioural data. This has two dimensions. First, there is direct user engagement, involving the activities carried out and time spent on the platform. In combination with profile data, these behavioural data may enable the platform to learn how to increase user engagement or to identify and suggest potential (relevant) extensions to one's social graphs. Second, indirect engagement occurs through user involvement with third-party apps and by visiting other websites that connect with the platform in various ways. Again, more information is always better, as it can contribute to better functionalities, better targeting for advertisers, and so on.

In addition to indirect user engagement, platforms may aim at engagement by non-users. Social networks sometimes track people without user accounts and monitor their online behaviour all over the internet. The general purpose is to harvest data anywhere and about anyone. This supports revenue models that are based on delivering targeted advertisements and helping firms to connect with customers.

User targeting

Social networks aim at capturing users' attention, for example, by means of ‘feeds’. For this purpose, it is important to predict the attractiveness of content to particular users[76]. More engaging feeds increase the likelihood that users will keep busy on the platform. Such predictions benefit from detailed information about consumers' preferences and behaviour, which sheds light on the "data hunger" of social networks.[77]

---

[74] Hogan (2018).

[75] "As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants", *The New York Times* 18 December 2018.

[76] Note that this applies more to general-purpose social networks with a large diversified audience than for social networks dedicated to specific groups, In case of the latter, there is less need for personalisation as the user group is more homogeneous.

[77] Calvano and Polo (2020).

Revenues

Facebook's revenues per user are significantly larger than those of rivals like Twitter, Snapchat and Pinterest.[78] The company presumably has the richest user data, allowing it to select ads that users do not find intrusive, resulting in a relatively high portion of ads relative to the amount of other content, and driving the willingness to pay for advertisements.

Data for innovation

Innovation by digital platforms such as social networks is partly (and to a large extent) driven by data. The continuous development process facilitates live experimentation using A/B testing. This involves new functionalities and other innovations to be implemented and deployed immediately, directly followed by an assessment of the impact on user behaviour (initially on small user groups). Such testing quickly generates data about the tested features, which can then be used for fine-tuning and roll-out. At Facebook, for instance, engineers perform tests of new code and run thousands of regression tests to check for mistakes and performance. Thus, developers are constantly working on new code, testing it to generate data and learning from the findings[79].

4.3.2 Which data are indispensable?

A social network, by definition, needs social graph data

Social data containing information on social graphs of users and data on user interactions can be viewed as necessary for any social network. The reason is that these data types are necessary inputs for the core functionality and features for users.

Business models dictate additional data requirements

Beyond social data supporting a social network's core user functionalities, there is no single or clearly identifiable type of data that is essential or necessary. The underlying business model, which may come in many colours and shapes, will dictate which data types are necessary for a viable proposition.

For a subscription-based social network aiming at specific (e.g. professional) user groups, profile data need to be less rich and diverse compared to social networks with an advertising model. For general-purpose social networks that generate revenues from harvested data, the profile data cannot be diverse enough to be of value to the wide variety of third parties such as advertisers, app developers and data resellers. Some types of data will be of immediate value to current third-party users, while other types may depend on future innovations and users. Therefore, some social networks tend to collect virtually all data about users as well as non-users (personal data, behavioural data, device data, inferred data, apps used etc.), whether or not they are using the social network at the time of capturing the data. All such data may ultimately support user growth, innovation and revenue generation.

4.3.3 Can access to indispensable data facilitate entry or strengthen competition?

Impact of data sharing on competition

The literature often mentions data portability as a means to reduce customer lock-in and to increase customer mobility. Even when users multi-home, they may still be "trapped" in a certain platform because they do not want to leave behind their existing connections and the content uploaded by these connections as well as by themselves. An entrant's incentives to innovate will increase if it gets easier to attract customers. Thus, certain types of data sharing, in particular data portability,

---

[78] "Why Facebook generates much more money per user than its rivals", CNBC 1 November 2019. https://www.cnbc.com/2019/11/01/facebook-towers-over-rivals-in-the-critical-metric-of-revenue-per-user.html (retrieved 19 April 2020).

[79] Feitelson et al. (2013).

can make horizontal competition more effective. In markets where consumers multi-home, the impact may be modest though.

Data portability, and data sharing in general, may also lead to the creation of vertical or downstream business models that build upon an incumbent's data. While the impact on competition and innovation is not obvious[80], note that such data sharing strengthens the platform's ecosystem and at the same time leads to new business models (by developers using the platform's data). The former effect may lead to a more dominant position for the platform, which makes it hard to make a general statement. Furthermore, if the data sharing pertains to profiles, the question is what this data is being used for. In case it is used for advertising purposes, data sharing reduces the possibilities for differentiation, which is undesirable for consumers, and also for entrants aiming at niche markets. Alternatively, if the data sharing pertains to social data (describing the social graph and interactions with other users), then it may increase customer mobility and possibly multi-homing. This has a positive impact on differentiation and competition, although in many social network markets, consumers already multi-home.

A social network may circumvent its own unwillingness to facilitate multi-homing with rival platforms through mergers and acquisitions. For instance, Facebook's acquisitions of WhatsApp and Instagram expanded its ecosystem for which it would make sense to allow for user interactions and user data pass-throughs across the three platforms. Such a move could easily lead to further antitrust scrutiny though.

Does data form an entry barrier?

A common example of a potential entry barrier in social media does not relate to data, but to the presence of network effects of an installed user base. Even for such an obvious candidate for a cause of an entry barrier, one must ask if it actually has the supposed impact. For instance, the Dutch social network Hyves, founded in 2004, had built up a strong presence in the Netherlands. This was before people got acquainted with Facebook. Still, by 2013 its user base had decreased substantially and Hyves was discontinued. Network effects notwithstanding, people left Hyves for Facebook, and young users immediately chose Facebook. Note, however, that Facebook's current user base is of a different order, so that the example of Hyves does not imply that a rival with a comparable social network is able to outcompete Facebook. Challengers will probably have to introduce distinct and highly attractive features aiming at new user generations. Moreover, anecdotal evidence indicates that users left Hyves because the company failed to keep up investment levels in server capacity with the growing user base, harming the user experience (Pastoor & Peddemors, 2013).

Regarding the possibility that data forms an entry barrier, note that a new social network that is able to attract users will be able to collect similar data as an incumbent. This can take time though. If consumers face switching costs, incumbents may have an advantage that is hard to overcome[81]. Still, this does not yet imply the presence of an actual entry barrier.

Suppose that an incumbent possesses a data set similar to what newcomers need to have. A possibility for data to engender an entry barrier lies in data-powered economies of scale, resulting in positive feedback loops. For instance, social networks with more users can extract more data, which helps them to improve the quality of user interactions, and in turn attract more users. Thus, the question whether access to data in possession of an incumbent is good for entry and competition then depends on economic characteristics of data. Do data substitutability, complementarity and data returns to scale give rise to a persistent incumbency advantage that puts

80 Engels (2016).
81 Gans (2018).

entrants at a substantial disadvantage for which they cannot find alternative solutions? These are empirical questions that need to be addressed within the context of a specific situation.

In addition, is it mainly because of data characteristics that entry is blocked, or do other aspects play a significant role (possibly in combination with data)? Entry barriers may be caused by a constellation of market characteristics, many of which may be unrelated to data. Examples are network effects, customer switching costs (e.g. due to brand loyalty, administrative burden), and difficulties to get financing in the initial business phases. The extent to which these market characteristics are present, will be an empirical, context-dependent question. Furthermore, there may be legal entry barriers, for instance if privacy protection or intellectual property rights allow an incumbent to exclude entrants from using its data.

Entry barriers may not be persistent. To see this, it helps to envisage different types of competitive threats to an existing social network[82]. First, horizontal competitors — suppliers of somewhat similar attention platforms, possibly based on other types of social networks or even alternative user engagement propositions — may appear on the scene. Second, vertical entrants may threaten an incumbent's position. An example is an independent app developer with (or desiring) a data-sharing agreement with the social network, whose business model evolves into a direct substitute for a part of the social network's activities. For instance, the company may incorporate social features into the app, or may start operating as an attention platform itself.

An observation that goes beyond data is that new social networks can have a hard time to be findable for consumers, making it difficult to grow a significant user base. Downstream firms often face similar challenges. For many apps that use (e.g. through APIs) social network data, having active users is much more important than having access to data. And even then, sometimes a certain aspect may give the impression of an entry barrier, but only for a while. Many new apps and platforms, therefore, employ “growth hacks” during the start-up phase, latching onto third party platforms (whether via APIs or automated processes to obtain data). Thus, creativeness and innovation are key ingredients to overcome hurdles that, at first sight, may look like unsurmountable entry barriers.

Overall, one cannot say in general that data forms an entry barrier. Moreover, innovation in technology, computer intelligence and business models may render assessments of entry barriers obsolete quickly. This is even more so in the light of the fast, ongoing innovation cycles employed by digital platforms. It is important to note though that, for social networks, network effects and other causes of customer lock-in will slow down such disruptions from innovation.

Does mandatory data access stimulate competition?

Because of the observations above, establishing the presence of an entry barrier does not automatically warrant mandatory access to data. Moreover, one should not consider data in isolation. For social networks, having active users may be much more important than having access to data, and new users can be attracted in smart ways (innovative features, growth hacks) and by cooperating with other (software and device) platforms.

Having observed this, access to an existing social network's data can of course alleviate the competition-damping impact of network effects. For instance, data portability may, to some extent, encourage consumers to switch to a similar social network. Interoperability, based on real-time mutual access to data and functionalities, would help to completely overcome customer lock-in due to network effects. Note, however, that interoperability will be hard to specify and implement. It also seems to conflict with the dynamic nature of innovation in business models and functionalities.

82 Cross-sectoral relationships will be discussed elsewhere.

Allowing downstream apps to benefit from a social network's customer base and user data may not introduce direct competition for social networks, but it certainly adds to innovation and market entry in a broader sense. At some point, for instance when developers integrate social networking features into their apps, they may start threatening the position of established social networks.

As discussed in the report, access to user profiles does not automatically increase competition. Sharing such data may, besides privacy issues, put more weight on scale as a precondition to build up a competitive market position. The reason is that when different social networks avail of a common set of profiles, there are less possibilities to differentiate their targeting propositions to advertisers. As a result, platforms will likely differentiate by targeting a theme along with a differentiated user group and offer advertiser an opportunity to associate themselves with that user group or the social networks' theme. The mode of mode of competition may thus change (more specifically, the mode of differentiation) and it may mitigate the need for data by platforms. But it is not clear how this will affect the drive for innovation and ultimately the interests of consumers.

Gans (2018) argues that right to data portability, which typically allows users to take their data to another platform, may not be as effective as one would wish. Portability rules typically address only data that a user has contributed him- or herself. However, switching costs may also stem from data (e.g. content) provided by others. The interactions among users create network effects that also give rise to switching costs. To address the larger switching costs due to network effects, Gans suggests mandatory identity portability as a means to enhance consumer choice: a user who switches to another platform, would then retain his or her permissions to access data provided by other users of the social network. This allows a user to leave the ecosystem without losing the ability to communicate with existing, remaining connections. Note that this goes beyond data portability: it amounts to a form of service interoperability, with has serious drawbacks (as discussed below).

4.3.4 Are there bottlenecks for accessing indispensable data?

A bottleneck or essential facility may be present if there is only a single way of reaching consumers or necessary inputs, possibly including data. Strictly speaking, a given social network's social graph as a whole is typically exclusive and unique, as it cannot exactly be replicated based on data from other sources. To obtain it, one would have to get access through that social network or obtain it through the users themselves, which will be cumbersome. Having said that, a priori it does seem feasible to replicate a (somewhat similar) social graph or build a new one.

If an established social network does not provide access, then alternative data may be available through accessing users' digital address books, web scraping, data markets, or APIs of various other platforms. An example is a (downstream) sports app that latches onto devices (e.g. fitness trackers) and software platforms that provide access to their user data. This makes it convenient (reduced switching cost) for consumers to use the app "on top of" a device's functionality. The benefit is mutual, as a fitness tracker's platform sees its ecosystem value go up, and also gets more brand name recognition (through the app).

This example shows that, as a myriad of data types can have direct or indirect commercial value, there can be room to innovate in smart ways of collecting (possibly different but) more relevant or more timely data. A challenger of an incumbent social network might also find ways to collect data that is similar to the incumbent's data or has similar value. Thus, data does not automatically imply the presence of a bottleneck.

4.3.5 *Is indispensable data being shared (voluntarily, on a commercial basis)?*

Data sharing is very common

Social networks like Facebook and Strava tend to have a very large number of data-sharing agreements with various parties, such as developers of apps that run on the social network's platform, developers of (standalone or off-platform) apps that run on mobile devices, websites, and advertisers. There may also be indirect relationships with third parties that need data, namely through resellers.

In what follows, we discuss in some more detail which types of agreements can be observed in practice. Note, however, that we cannot assess if these agreements pertain to indispensable data. To the contrary, one encounters several examples in which data is very useful for the business model of a complementary app, but strictly speaking not essential. This is illustrated by cases in which the abandonment of an agreement did not inflict fatalities on app developers.

Practices of data sharing

Data-sharing is typically carried out through APIs. For instance, social network may provide "open" (free to use) APIs that give third-party developers access to services or a social network's proprietary data. These developers mix and process this data to be used by their apps. Open APIs are often proprietary software, providing packaged and restricted data without allowing developers to adapt the interface. Developers may not exactly know what an API does, in particular if it is provided as an executable file instead of the source code.[83]

The data shared through APIs is limited and different API tools may provide varying levels of access. Presumably, shared data typically does not contain social graph data, as this is highly important for network effects of social networks. Another reason is that privacy regulations make it hard, if not impossible, to share such data (if only because a user's connections would also have to give consent). On Facebook, for instance, in order to access data beyond public profiles of users, developers must (i) specify their data request (ii) obtain user consent; and (iii) let their apps retrieve the platform's data in line with the user's direction. They have to agree to Facebook's terms of use, which aim to protect Facebook's users and their data, as well as the quality of features and functions that are part of the platform's user experience.

Some examples related to Facebook illustrate data-sharing in practice:[84]

- On mobile devices, Facebook uses various software tools, known as Business Tools, to routinely track users (also if they are logged out) as well as non-users, also outside its platform. App developers share data through a set of development tools known as the Facebook Business SDK (Software Development Kit), allowing them to build apps for specific operating systems. Apps that automatically send data about users and their behaviour on apps to Facebook typically combine it with a unique identifier, the Google advertising ID on Android phones (AAID) or Apple's equivalent on iPhones (IDFA). This ID allows advertisers to link very detailed user data collected from different apps and browsing into a single user profile, providing them with a fine-grained and intimate picture of individual persons. Facebook's Cookies Policy describes how people without Facebook accounts can opt out of cookies to show them ads, but apparently, such data is also being shared, as shown by tests of opt-outs. Facebook lays the responsibility for lawful data practices (collection, usage, sharing) on app developers; the SDK's default implementation automatically sends event data to Facebook. Also, due to Facebook's lack of transparency, it is impossible to know how shared data is being used.[85]
- Partnerships with device makers (e.g. Apple, Amazon, BlackBerry, Microsoft and Samsung) allow them to offer social networking features such as messaging, "like" buttons and address

[83] Qiu (2017).

[84] For an overview see Facebook's Data policy, https://www.facebook.com/full_data_use_policy.

[85] Privacy International (2018).

books. For Facebook, device partners get access to the Facebook app and custom-built apps and integrations belonging to the "Facebook experience". This may allow them to retrieve personal data (in the past also about users' Facebook friends, even if they did not give Facebook permission to share information with third parties).[86]
- Facebook's social login makes life easy for users of third-party apps and websites, as they reduce the hassle with passwords. The mechanism is that Facebook receives data about the users of the app or website (e.g. e-mail addresses) and can contact them with suggestions about users of the app to follow.

It is very hard to provide a complete picture of current data-sharing practices. A series of articles in *The New York Times* gives a fairly detailed account of recent practices, some of which have begun to receive political and legal scrutiny triggered by the Cambridge Analytica scandal. Some more examples of data-sharing during the last decade are:[87]
- Facebook shares (or shared) contact information with hundreds of advertisers, most of which users have never interacted with.
- Advertisers who have obtained personal data elsewhere can upload this into Facebook through a tool called Customer Audiences, which helps them connect this data to Facebook profiles and then serve them targeted ads.
- Some of Facebook's APIs allow users of third-party apps or online services to follow Facebook friends, a feature that helps new apps to speed up user growth. An example was Vine, a video-sharing service. When signing up for the service, new users were given the option of following Facebook friends. Facebook, however, shut down the friends access API, as Vine competed with Facebook' own features. Through Instagram, introduced its own short-form video feature shortly after. Its growth stagnated and Vine closed shop.
- To strengthen Facebook's position as an ecosystem at the heart of online activities, at some point the company tended towards reciprocal data-sharing agreements, giving third-party developers the ability to connect apps to Facebook (allowing users to post app activities on Facebook timelines) in exchange for data that they collect. The idea was to make its platform more attractive for developers. According to Facebook CEO Mark Zuckerberg: social apps "may be good for the world but it’s not good for us unless people also share back to Facebook" (NYT 5 December 2018). Later, the reciprocity rule was weakened by dropping the requirement to send usage data back to Facebook.

The above impression is unlikely to reflect the current practice, as Facebook had to scale back its data practices in response to political and legal scrutiny and investigations.[88] One may expect though that current data deals exist for the sole purpose of supporting its business model, that is, its partners help Facebook to expand and attract new users who spend time on Facebook and contribute to its ad revenues. In case its partners receive user data in return, one may presume that Facebook has put checks and balances in place, at least to make sure that there is no violation of legal requirements.

Motives for data sharing

Social networks benefit from (vertical) data-sharing agreements, even though they tend to form relatively closed ecosystems spanned by their user bases.[89] For instance, Facebook forms data

86 See e.g. "Facebook’s Device Partnerships Explained", *The New York Times* 4 June 2018.

87 "I Downloaded the Information That Facebook Has on Me. Yikes.", *The New York Times* 11 April 2018; "Facebook Gave Device Makers Deep Access to Data on Users and Friends" *The New York Times* 3 June 2018; "Facebook’s Device Partnerships Explained", *The New York Times* 4 June 2018; "Facebook Failed to Police How Its Partners Handled User Data", *The New York Times* 12 November 2018; "Facebook Emails Show Its Real Mission: Making Money and Crushing Competition", *The New York Times* 5 December 2018.

88 "Facebook’s Data Deals Are Under Criminal Investigation", *The New York Times* 13 March 2019.

89 Bruc (2019).

partnerships as a way to integrate Facebook's services into apps, devices and websites by third parties. Indirectly, these partners help Facebook to grow and get new users on board. who spend time on the platform and contribute to ad revenues.[90] Facebook made a myriad of such agreements that are typically subject to confidentiality. In many cases the company shared personal data while ignoring users' privacy settings, and without asking users for consent or informing them about the extent of the data-sharing.

Third-party developers and partners with data agreements can share data with a social network by installing software development kits, pixels, or a social login tool. Such pieces of code can collect user engagement data and share it with the platform. The social login provides an effective illustration of benefits obtained by a social network. While Facebook apparently doesn't receive behavioural data in return from apps and websites, the integrated login functionality allows the company to track usage through metadata. Thus, the social login functionality allows Facebook to track usage from thousands of apps and websites. This provides them with an enormous knowledge, on a real-time basis, about market shares and whatever can be inferred from the metadata.

Third parties (that is, customers) may benefit from (vertical) data-sharing agreements with social networks for various reasons, in particular realizing growth and targeting consumers.[91] Pieces of code implemented through software development kits, pixels, and social login tools collect user engagement data that can be used to help the platform's partners to assess and improve the effectiveness of advertising and targeting, and to obtain insights about their user base. App developers, for instance, can get access to an enormous potential of customers by connecting to an established social network. An option to follow friends on a big social network is a convenient tool for apps to accelerate user growth, in particular if access is free (as is the case for Open APIs). Such interfaces allow developers to combine a social network's proprietary data about users with their own data and integrate it into their apps' functionalities. In addition, data about users and the social graph can give them novel (otherwise unavailable) insights and form valuable inputs for innovation. Advertisers are able to target ads with user data. For instance, Facebook shares a variety of tools and data to increase advertisers' understanding of the effectiveness of ad campaigns. This enables them to measure and track the performance of ads on the platform in real time with metrics for reporting, insights and conversion.

Consumers may benefit from data-sharing because of, for instance, convenience (e.g. social logins), innovative features (e.g. social network functionalities integrated in third-party apps and devices), and more relevant ads.

Bargaining power in data-sharing agreements

While both sides benefit from data-sharing agreements, **bargaining power may be asymmetric**. The big social networks (also in niche markets) tend to assert full discretion over the commercial data-sharing relationships with many third parties, for instance regarding who can access data about users and social graphs, and in which manner. As the provider of the data (API platform), they have all the power to modify or terminate access to the platform. The only bargaining power of an API user stems from size, which occurs when a customer is large enough to enter a custom (long-term) agreements instead of the "as-is" nature of public APIs. Note, however, that closing access for a particular app tends to generate bad publicity for the data provider, as was observed when Strava foreclosed Relive. This tends to be temporary though and hardly impacts end-users.

---

90 "As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants", *The New York Times* 18 December 2018.

91 See e.g. Nicholas and Weinberg (2019) on the usefulness of exported Facebook data.

The closed-source nature of open APIs contributes to the dependence of developers on the APIs' owners and vendors. An API may include a key (a text string that identifies a developer as the source of an API request), which provides possibilities to monitor users and to control (and possibly suspend) their API usage.[92]

There are various ways in which APIs can be used to exert power over third parties, as is clear from Facebook's Platform Policy:[93] (i) Monitoring API usage: according to Facebook's policy, it reserves the right to create apps similar to apps that integrate with its Graph API. (ii) Limiting or cutting off API access: according to Facebook's policy, it reserves the right to block Graph API access. (iii) Change the available data or the structure of the data: Facebook changed (e.g. through depreciation) the Graph API functionalities. The Graph API allowed third-party apps to post updates on users' profiles, which were visible to friends. These developers also obtained data about Facebook users, including social graph data and locations.

Some examples in which such practices have been applied by various social networks are the following:

- During (roughly) the last decade, Facebook and other social network have limited third parties' access to data. For instance, in 2007 a third party on Facebook had access to individual profile data, friends lists, and selected data about friends (conditional on a user's permission). In 2015, a typical third party could only access a user's friends also using its app. Facebook implemented these changes through the "end points" where third parties can download user data and the rules for accessing them (hence through the APIs).[94]
- Twitter uses APIs that require authentication of data requests and impose additional terms on developers with high-volume usage. It also provides Twitter the possibility to control third-party apps that pose a competitive threat.[95] In 2015, DataSift (a data reseller) announced that Twitter was ending their relationship. Customers who wanted continued access to Twitter's "firehose" would have to set up a connect with Gnip, a data provider (social media API aggregation company) acquired by Twitter in 2014.[96] Another example was the termination by Twitter of a sharing agreement with PeopleBrowsr, apparently because it viewed PeopleBrowsr's services as incompatible with its adapted business model.[97] The overall picture is that Twitter initially offered broad access to its data and encouraged spinoff innovations, and restricted access later on, parallel to acquiring Tweetdeck (a dashboard app for managing Twitter accounts) and Gnip.[98]
- Shortly before its launch in 2013, Twitter acquired Vine, a short-form video hosting service. New users of Vine could choose to follow their Facebook friends, a feature that was implemented through an API of Facebook. However, Facebook quickly shut down this type of access. Not much later, Instagram introduced ts own short-form video functionality, and in 2016 Vine quit the market.[99]
- Running and cycling apps Strava and Relive used to complement each other. Strava is a social network for athletes that tracks speed, altitude and routes. The Dutch app Relive accessed this data to build a moving map. When the latter company introduced the feature to follow friends within its app, Strava terminated the data-sharing agreement[100].

92 Qiu (2017).
93 Nicholas and Weinberg (2019), Appendix A.
94 Hogan (2018).
95 Qiu (2017).
96 "Twitter Ending Firehose Access for DataSift, Other Data Resellers", *ProgrammableWeb* 13 April 2015, https://www.programmableweb.com/news/twitter-ending-firehose-access-datasift-other-data-resellers/2015/04/13.
97 Puschmann and Burgess (2014).
98 Puschmann and Ausserhofer (2014).
99 "Facebook Emails Show Its Real Mission: Making Money and Crushing Competition", *The New York Times* 5 December 2018.
100 "Strava op ramkoers met Nederlandse sportapp Relive", *Het Financieele Dagblad* 1 August 2019.

In the recent past, Facebook used to explicitly forbid close connections with (horizontally) competing social networks through its data access conditions[101]. The conditions said that a partner must not use Facebook to export user data into a competing social network without our permission, and that apps on Facebook may not integrate, link to, promote, distribute, or redirect to any app on competing social networks. The current version of these conditions no longer contains this restriction[102]. In 2018, Facebook reviewed its policy to block apps built on top of its platform that replicated its core functionality and decided to remove this policy[103].

For would-be competitors with a large user base, social networks may (wisely) choose not to foreclose them with access to data. Doing so would disappoint a large number of consumers. This may explain why Facebook maintained its social log-in agreement with Spotify. After signing in on a Facebook account in Spotify's app, a user could send and receive messages (through Facebook) without leaving the app. This worked by using an API that provided Spotify with access to personal messages[104].

An important point related to the observations above is that platform owners will compare the substitution and complementary effects between their own apps and third-party apps. Independent developers, in particular smaller ones, tend to be at the receiving end of commercial agreements. Moreover, the impression is that existing (voluntary) data-sharing agreements tend to be vertical in nature. They are driven by mutual interests. The social network benefits from the complementary value that third parties bring to their ecosystem, or from more direct revenues from data. Independent developers benefit from the growth opportunities provided by an incumbent's user base and data, or from more direct opportunities to use data as a way to grab user attention. When business partners are active in the same arena as the platform (e.g. social interactions), they start to form a horizontal competitive threat which typically results in a termination of the agreement initiated by the social network.

4.3.6 What is the impact of mandatory data sharing on innovation incentives?

A general concern is that a mandatory data-sharing regime may lead to underinvestment in the collection and production of data, and that it undermines data-driven business models[105].

Horizontal data sharing

Policy recommendations in the literature gravitate towards mandatory data portability as a means to reduce customer lock-in and increase customer mobility. A drawback of an obligation to provide portability is the reduced return on investment, which may affect incentives to innovate. For potential entrants, the efforts (including organisational adaptations and implementing legal safeguards) required to make data portable might be prohibitive in the first phase of their existence — say the first two years of the start-up phase. This can be overcome by granting newcomers a short exemption period. Note that one has to specify to which types of data obligations apply. As was already discussed, to stimulate consumer mobility it may make sense to focus on social data and user-generated content; although the effect may be limited – see reference above to Gans (2018).

---

101 "Facebook blocks competitors from its social data", Macworld 25 January 2013. https://www.macworld.com/article/2026233/facebook-blocks-competitors-from-its-social-data.html (retrieved 22 April 2020).

102 https://developers.facebook.com/policy/ (retrieved 22 April 2020.

103 "Facebook ends platform policy banning apps that copy its features", Techcrunch 5 December 2018 (retrieved 24 April 2020).

104 https://about.fb.com/news/2018/12/facebooks-partners/ (retrieved 22 April 2020).

105 Borsenberger et al. (2018).

It follows that there would have to be a substantial problem with horizontal competition to impose horizontal data sharing (assuming this intervention would be able to solve it). The presence of network effects signals that such problems could well be present, but the burden of proof is a hurdle, as demonstrated by market developments such as the fast emergence of social networks like Snapchat and TikTok, and the demise of the (nationally) established social network Hyves. Entrants would benefit from demand-side scale economies while there may not be a solid rationale in terms of helping them to overcome network barriers.

Mandatory sharing of profile data is somewhat similar to number portability in telecommunications markets. Extending this parallel, would service interoperability be able to solve the problem of network effects, similar to the impact of network interconnection and interoperability requirements in telecoms? These obligations had the obvious purpose of helping entrants to overcome entry barrier caused by network effects. In digital markets, such as for social networks, the logic is different. These services are not inseparable from the underlying infrastructure, as used to be the case in telecommunications. The costs of switching and multi-homing are therefore much lower. Data portability in social networks may still contribute to the advancement of competition by facilitating multi-homing. However, multi-homing already is common so the need to mandate such measure is much less pressing. Therefore, the argument of helping entrants to overcome entry barriers is not immediately compelling for online services like social media. Note that this line of reasoning depends on how one views the relevant product market: what are relevant substitutes? For instance, does Facebook serve its own relevant market or does is compete with Reddit, Snapchat and so on?

A drawback of mandatory data sharing is that such measures could interfere with innovation in business models of social networks, and of services that go beyond social media. They could also lock a market into a specific standard for data sharing for years to come. (Van Gorp N. , et al., 2016) discuss this in the context of short messaging services (SMS). This telecoms service has, for good reasons, benefited from mandatory interconnection and interoperability. At the same time, however, the (legal and technical) requirements to conform to the legacy standard of SMS restricted the possibilities for future adaptations based on new technologies and business models, and for evolution into new territory. Thus, there is a risk of putting a brake on (unanticipated) innovations.

Vertical data-sharing

Vertical data-sharing concerns agreements between a social network and a third party, typically complementing each other. As long as there is no direct or anticipated competitive threat, parties will (or can be expected to) be willing to make voluntary agreements, because there is a mutual benefit[106]. One may ask though to what extent data will actually be shared, as there may be information asymmetries and uncertainty about the value. Also, for small third parties, an established social network calls the shots.[107] Hence the question is whether additional requirements, most likely to the benefit of third parties, are desirable and what their impact on innovation is.

Note that the social network will judge data-sharing requests on third parties' contribution to its ecosystem. In principle, the platform, in particular if it is not vertically integrated with firms similar to the third parties in question, will not want to harm them. Their success contributes to the social network's success. From a broader perspective though, one may want to evaluate data-sharing in the light of the merits for the whole market (or the economy). For instance, it may lead to spinoff innovations that are neutral to, or possibly harm, the platform's ecosystem, but generate value in

---

106 Engels (2016).

107 Nicholas and Weinberg (2019) provide examples illustrating how Facebook's policy may put third parties at the mercy of the social network.

other settings that outweigh the impact on the social network. To come back to the question above: one has to compare the possible reduction in the platform's profitability (and innovation incentives) with the potential increase in third parties' initiatives to develop new business models, technologies, content and apps.

What type of reasoning would be required to conclude that mandatory data-sharing is desirable? An assessment of the divergence of the social network's private incentives and the broader benefits from innovation is the starting point. Moreover, is there an asymmetric bargaining situation such that the social network may neglect the benefits from data-sharing that go beyond the value to its ecosystem? If so, the question is: what is the desirable scope for data access, in various dimensions such as the type of data, the timing (intermittent, delayed, real-time, live), the level of identification and tracking? As these questions require a very detailed analysis of business models and market characteristics, a more practical approach is more effective. We refer to the case on vertical exclusion for a further discussion.

Overall

Maintaining an API platform to provide data access (more broadly related to implementation, maintenance, and security) requires resources. It also exposes a firm to new risks, as data users may abuse the data which may reflect back on the platform. Therefore, for newcomers, it may not be a good idea to impose restrictions on them that hinder them in getting traction. A short exemption period (say two years) may solve this problem.

4.3.7 What are the risks of data sharing?

By sharing data with third parties, there is a risk that privacy and security protections get diluted. As a result, personal data may get compromised, resulting in harm to persons. They may include users of social networks, consumers of apps and people who are not users or consumers. The same may be true for the incentives to for compliance or perform audits on data partners (which are basically cost components, not revenue drivers). The more data is collected and shared, and the larger the number of device makers, app developers and ad companies who get access to data and perhaps store data on their own servers, the larger the privacy and security risks.

Contractual requirements only go so far to reduce this risk, as illustrated by the Cambridge Analytica scandal: it can be difficult to control what happens to user data that has left a platform. Similarly, it may be hard to prevent that user data leaves user devices, for instance when a device synchronizes or backs up data to cloud services.

4.3.8 What are the options under current legislation to make data sharing mandatory?

Options under GDPR

One of the purposes of the GDPR is to stimulate the free movement of personal data. It strengthens data subjects' control over personal data and impose duties on data controllers to restrict the collection and use of personal data. The GDPR contains a right to data portability that facilitates sharing and reuse of personal data at the request of the data subject. For instance, the regulation provides data subjects with the right to have personal data transmitted from one controller to another (if technically feasible), which constitutes a right to portability of one's personal data. Thus, current legislation allows for a restricted form of data sharing: data portability falls short of interoperability of social networks. More elaborate forms of data sharing, such as real-time portability might also be included under the GDPR's obligations (subject to technical feasibility), but the actual scope (and bite) of the obligations is not yet fully clear.

Options related to abuses of dominance in competition law (including the essential facilities doctrine)

Foreclosing access to data could, in certain situations, correspond to an abuse of a dominant position. This possibility is discussed in more detail in the case on vertical exclusion for an elaboration on this route to mandating data access.

The essential facilities doctrine (EFD) is particularly relevant as it pertains to situations where a dominant company does not want to share data. Of course, one of the questions then is whether this data is "essential". The direct consequence of applying EFD would be to impose obligations to share the essential data. However, the options to apply the essential facilities doctrine seem to be limited, due to the difficulty (or impossibility) to satisfy the necessary conditions establishing the presence of a bottleneck. Exceptions may exist, but the burden of proof will in any case be very high, while the range of situations that correspond to a bottleneck situation is likely to be very narrow.

Remedies resulting from the EFD are not ex ante, since they require an ex post assessment of an abuse. Still, once determined, obligations call for monitoring. Also, once implemented in a certain market, such remedies may incite players in other markets to share data, if only to avoid fines and liability.

Note that in case of other forms of abuse (e.g. discrimination), data sharing obligations could also be used as a remedy, supposing it would solve the problem.

Compatibility with the Database Directive and Trade Secrets Directive

Data may be subject to intellectual property rights including copyright, sui generis (that is, unique) database protection and trade secret protection. These regimes might get in the way of data sharing. In response, invasions into intellectual property rights could be made proportional by balancing interests into data sharing measures.

Copyrights may protect original content provided by users, such as posts and pictures on social networks. While users may hold copyrights over the content they upload, platforms often require users to give the provider a license to use this content, thus transferring users' intellectual property rights.

Sui generis database protection becomes relevant when a social network invests in obtaining, verifying or presenting pre-existing content (e.g. provided by users). However, it only applies to the database (the structuring of the data), not to the data itself.

Trade secret protection may also apply, conditional on data being secret, that secrecy renders it commercially valuable, while efforts have been made to keep it secret. If a social network processes user data to make targeted advertising possible, it may be the case that secrecy makes such advertising more profitable. For example, knowing what type of data feeds the algorithm may (in theory) reveal information about how the algorithm operates.

Thus, there are various layers of protection that may apply, and they are not necessarily compatible with each other. For instance, if user A takes a picture of user B, then A has a copyright while the picture contains personal data about B. When posted on Facebook, the picture becomes part of the database protection applying to the platform's data structure. The various layers may, at some point, come into conflict with each other.

An example illustrating that social networks indeed invoke trade secrets and other intellectual property rights to protect the data they hold about users is the dispute between Max Schrems (a data protection activist) and Facebook regarding his right to access his social network data. Facebook did not provide access to all personal data in response to the Schrems' requests for access in 2011. As justification, the social network provider referred to one of the sections of the Irish Data Protection Acts (as applicable to Facebook because of its Irish headquarters) that 'carves out an exception to subject access requests where the disclosures in response would adversely affect trade secrets or intellectual property'[108]. This also shows that data protection rights, such as the right to access personal data but also the right to data portability, may clash with the intellectual property rights held by data controllers. To solve this conflict, the two rights will need to be balanced against each other.

4.3.9 *What are other options for stimulating data sharing?*

Since the risks associated with data sharing (and mandating it) depend on case-specific characteristics as well as the nature and scope of data-sharing, it makes sense to try to identify no-regret policies. At first sight, a narrow set of obligations might apply to (i) data about individual behaviour and social interactions, and (ii) the social graph. Not so much with the purpose of realising full services interoperability, but to inject some convenience with regard to cross-network interactions. It is an open question if that would be worth the effort, as a priori there is not an immediate case supporting that data forms a substantial entry barrier, or that data access eliminates an entry barrier.

Data pooling arrangements such as the Data Transfer Project (DTP) of the big tech firms as well as some smaller companies could stimulate data mobility. The scope of such projects determines the scope of data sharing that it facilitates: does it merely aim at access to and portability of personal data subject to a common standard, or does it have full-service interoperability as a goal? OECD (2019) mentions as its use cases: data portability between services, trying out new services, and backing up data. DTP seems to gravitate around porting data.

To achieve a higher degree of interoperability (assuming it would be desirable), one probably needs a more elaborate common data format. It would have to enables "syntactic" compatibility (such that the target system can decode received data) as well as "semantic" compatibility (such that the target system also understands data)[109].

One may consider imposing restrictions on API access by regulating which types of restrictions companies may impose on this type of access. This could reduce situations in which social networks "arbitrarily" block downstream firms from data access.

---

[108] E. Protalinski, 'Facebook: Releasing your personal data reveals our trade secrets', ZD Net, 12 October 2011, available at https://www.zdnet.com/article/facebook-releasing-your-personal-data-reveals-our-trade-secrets/ .

[109] OECD (2019).

5 Case: banking

5.1 Case description

This case explores the impact on the banking sector of the asymmetric data sharing obligation for banks (PSD2). The case analyses the provision of financial services by Big Tech firms[110], its drivers, the resulting benefits, and its implications for competition and innovation. In doing so, it draws on examples from specific firms. The hypothesis is as follows: due to an absence of reciprocity in data sharing there is an unlevel playing field between traditional financial service providers on the one hand and the big tech companies on the other.

We test this hypothesis in a legal and practical (business economics) dimension. Therefore, the case consists of:

- a legal analysis of the degree to which tech companies are allowed to combine banking data with other (non-financial personal) data, and of the degree to which banks are allowed to combine these data types; and
- a factual analysis of the types of data with which financial banking data is combined by big tech companies and whether banks (can) have access to the same non-financial personal data used by tech companies.

If differences exist, the purpose of the case is to illustrate, with recent product innovations, that the competitiveness of banks is indeed hampered by the asymmetric treatment and not for other reasons (for example, by a rigidity of banks to adjust their business model). If the hypothesis is confirmed, a central question is whether reciprocity in data sharing obligations could restore the playing field, how one should scope and enforce such reciprocity requirements.

This case pertains to different types of data sharing between providers in different markets and of data obtained through different services. This does not necessarily refer to cross-sectoral data sharing. The latter term can be confusing. The fact that PSD2 provides outsiders with access to banking data does not amount to cross-sectoral data sharing, because these outsiders use the data to enter the banking sector (which contributes to horizontal competition). Cross-sectoral data sharing involves sharing of data with sectors other than those where the data is being collected. If banks would get access to data from other sectors, for instance data from tech firms, there would be cross-sectoral data sharing.

5.2 Analysis of research questions

5.2.1 Which role does data play in the business model, and what are the benefits that data provide?

Financial firms generate revenues from banking transactions and other financial services related to consumer banking, commercial banking, investment banking, and asset & wealth management. In contrast, some of the Big Tech firms generate revenues primarily from collecting and monetizing data, or from selling hardware, software and related IT services. What drives then the motivation of

110 One can think of Big Tech as the large US tech companies Google, Amazon, Facebook, Apple and Microsoft, but there is no need to restrict to these examples. Note also that these companies are interested in financial services to different degrees and for different reasons.

Big Tech firms to expand into financial services? The firms themselves argue that this is mainly to improve customer satisfaction and offer convenience to their customers.[111]

According to commentators in the Banking Industry[112], Big Tech firms have expanded into financial services for three main reasons. First, offering financial services allows them to diversify their business and opens new revenue opportunities. Second, the provision of financial services enables Big Tech firms to provide higher levels of customer experience to their end-users by integrating financial services into their existing platforms, thereby also increasing revenues from their core businesses. Third, offering payment or account information services allows Big Tech firms to collect additional data on the spending habits and financial positions of their customers.

Looking at it from a strategic angle, some Big Tech companies offer financial services as part of their super-app or single-ecosystem strategies.[113] As part of such a strategy, they are trying to make their app a central part of their customers' life and moving beyond search, social media, eCommerce, or ride-sharing activities. With this strategy, they aim to reduce frictions and deliver hyper-personalized, superb experiences that consumers find helpful.[114]

Banks and financial service providers also understand the importance of data. Financial Service Providers collect customer-provided data and observe transactional data about their clients. The banks collect customer-provided data through applications for checking accounts, credit cards, loans, and mortgages. This data may include salary payments, annual income information, employer details, or other data elements. In addition, banks also observe transactional data while capturing the customers' activities. Accessing other data (hereafter referred to as "non-traditional data") is not regulated; they may be observed by the banks or obtained, either bought or freely accessed, from third parties.

The banking industry is among the most data-driven industries, and firms are leveraging big data and already employ advanced analytics to grow their business. According to McKinsey, big data can generate more value to financial institutions in three ways. First, it can boost traditional Profit & Loss levels by accelerating growth, enhancing productivity, and improving risk control. Second, it can help them to find new business models and new sources of growth. Third, it can improve the customer experience by gathering real-time data and using analytics to understand the customer and build the proper customer journey.[115]

When we contrast these motives with those of the Big Techs, we must realize that the motives are very similar. Data is used by both the Big Tech firms and the Banks to (i) improve the P&L of their legacy core business, (ii) build a better customer experience, and (iii) find new businesses and business models.

The Big Tech's move into the domain of financial services does not in itself result in level playing field issues. The potential issue is their first mover's advantage in building "Super Apps".[116] The key advantage of the Big Tech's is their legacy market position and data access advantage to build better customer engagement and the capability of offering superior services using their super-app

[111] For example, Amazon claims that speed, security, purchase guarantees, trust, and convenience are the top benefits of Amazon Pay service. See https://pay.amazon.com/blog/top-reasons-why-consumers-like-amazon-pay

[112] Financial Stability Board, BigTech in finance. See https://www.fsb.org/wp-content/uploads/P091219-1.pdf

[113] Super apps can serve as a single portal to a wide range of virtual products and services. See https://home.kpmg/xx/en/home/insights/2019/06/super-app-or-super-disruption.html

[114] The most sophisticated apps like WeChat and AliPay in China bundle together online messaging, social media, eCommerce platforms (like eBay), and ride-sharing services. One super app, one sign-in, one user experience for virtually any product or service a customer may want to use.

[115] See https://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value

[116] It would be interesting to run an experiment and check if people would rather make a payment via their (super) social media account, or be willing to chat and post pictures via their (super) banking app.

platforms. Even if traditional financial service providers would be allowed to broaden the scope of their services and gain access to Big Tech data, it is highly unlikely that they can build super apps that can effectively compete with those of Big Techs. We shall also add to this that user experience (UX) is critical to market success in the platform economy, and hyper-personalization plays an important role in the UX design. Larger, more extensive data sets can be used to drive better insights and better insights can lead to better UX design. This itself, in addition to their first mover's position in building super-apps, may create another advantage for the Big Techs. Even if banks get a chance to access non-traditional data of selected customers, they would not have access to non-traditional data of customers on their Serviceable Available Market (SAM), therefore, they could likely not match the UX that Big Tech's are capable to offering to their end customers.

5.2.2 Which data and types of data are indispensable?

Regarding cross-sectoral data sharing from Big Tech firms to financial service providers, it is hard to define what data and data types would be indispensable for the business model.

It is straightforward to define what data is necessary for the provision of given financial services. Indispensable data in this context means data that is needed for payment related services or for account information services offered by e.g. Big Tech firms. According to PSD2, banks must provide *'all information on the initiation of the payment transaction and all information accessible to the account servicing payment service provider regarding the execution of the payment transaction[117]'*. Note that Big Tech firms can request all essential information from the financial service provider, but only with the consent of the end-user. In the case of payment services, the needed data is minimal. Here, one can think of data related to the identification of the customer and data associated with the particular payment transaction (e.g., the transaction amount, time stamp, location identifier, and confirmation if sufficient funds are available to complete the transaction). In the case of account information services, the data exchange is more extensive, but the exact data content is more ambiguous. Banks will need to provide all information related to the account information service. This information will include data associated with the identification of the customer, account history, information about other accounts, account balances, and any other data that is collected by the bank and is available to the end-users when they would log into their respective accounts. For payment transactions and account information services, this type of data access is enabled by PSD2. Derived and implied information, which is not available to end-users but might be used by the bank for customer analyses or other assessments, are not covered and currently there is no obligation for Banks to share them.

The nature of indispensable data is different for banks. The primary reason why they want to get access to non-traditional data is not to develop services outside of the financial services market but to enhance their core activities: enhance productivity, improve risk control, and enhance the customer experience. One can argue that obtaining non-traditional data is "indispensable" for Banks to provide traditional banking services, so they can get the efficiency and quality required in the digital age. In this context, non-traditional data is important to improve their efficiency and to withstand the increased competitive pressures from (Big)Techs, but it is not "indispensable" for the provision of traditional banking services.[118]

[117] PSD2, Art 66, 4.b.

[118] Banks do not develop services outside of the financial services market. In certain jurisdictions, Banks are not allowed to go into non-banking activities. In other jurisdictions, they could go into certain non-banking activities, however doing it would be very costly for them due to banking regulations (financial stability, prudential banking) and compliance requirements, such as Anti-Money Laundering (AML), Know-Your-Customer (KYC), and on-boarding processes.

By way of illustration, let us consider the example of lending. Traditionally, players in the financial industry assess consumers' creditworthiness based on their financial history.[119] An increasing number of lenders now also rely on non-traditional data to assess credit risk. For example, Lenddo assigns credit scores based on information in users' social networking profiles (with their permission). This information includes education and employment history, how many followers they have, who they are friends with, and information about those friends (Rusli, 2013). Rusli's research also implies that an individual's credit assessment is likely to be affected by the creditworthiness of his or her most relevant social media contacts. Bertrand and Kamenica (2018) show that owning an Apple iOS device is one of the best predictors for being in the top quartile of the income distribution. Gambacorta et al. (2019) found that financial models based on machine learning and non-traditional data are better able to predict losses and defaults than traditional models in the presence of a negative shock to the aggregate credit supply.

These findings suggest that credit scoring models based on machine learning and non-traditional financial data used by the Big Tech firms are better able to predict borrowers' losses and defaults than traditional empirical models alone. Although Ruiz et al. (2019) argues that a digital footprint[120] complements rather than substitutes for traditional credit bureau information, it is logical to reach the conclusion that analysing both traditional and non-traditional data will predict defaults better than analysing only one of them.

Payment services and Account Information Services allow Big Tech firms to collect information about the payment history of individuals. This data access is far-reaching because recent research suggests that cash-flow details provide a more detailed and timelier picture of how applicants manage their finances than traditional credit reports. When this cash-flow information is combined with other data collected about end-users (the digital footprint), Big Techs might be better positioned to develop risk models than relying solely on financial data or only on the digital footprint. Having access to spending habits and financial situation may also improve the accuracy of their customer profiling and allow them to increase revenues in their core businesses. If the combination of financial information with digital footprints yields superior information on predicting defaults, then Big Techs firms may challenge the traditional financial institutions based on an information advantage, rather than on merits (alone).

5.2.3 Can access to indispensable data facilitate entry or strengthen competition?

Access to banking data under PSD2 certainly helps to facilitate market entry. Big Tech firms have applied for PSD2 licenses and they can offer financial services. This may intensify competition in the financial sector. However, some argue that when Big Tech firms' access banking data, this may ultimately weaken competition in financial services, given their ability to tip markets. However, even when Big Techs were to conquer the market for financial services, we note that the nature of competition between Big Tech companies is inherently more dynamic then the competition between banks. As such, it remains to be seen whether the net effect of entry by Big Tech companies is good or bad for consumers.

According to some sources, the rationale of PSD2 regulation is viewed as pro-competitive: the business case of new entrants who want to provide online payment-related services largely depends on the availability of customers' account data. Because the latter is at the sole disposal of incumbents, and due to limited competition upstream, banks do not have sufficient incentives to share consumer data with downstream companies. The obligation outlined in Art. 36 of PDS2 aims

119 The FICO score, for example, uses data from five categories, namely payment history (35%), amounts owed (30%), length of credit history (15%), new credit (10%) and credit mix (10%). See https://www.myfico.com/credit-education/credit-scores/whats-not-in-your-credit-score

120 Ruiz et al. (2019) used the following ten variables to define digital footprint: digital device type, device operating system, host name of the email provider, channel, transaction time, do not track settings, name in the email, number in the email, lower case letters in address, and email validity.

at preventing banks from refusing *"to let qualified third parties access customer's account data despite the latter's consent"*, and therefore to foster competition (Porto & Ghidini, 2020).

In contrast, running commercial platforms and providing payment services allows Big Tech firms to accumulate a broader set of information on the merchants' turnover and scale of business as well as to make better judgment on their financial soundness (broader and better compared to traditional financial service providers). They can also accumulate broader information on an individual's financial position. These make it possible for the Big Tech firms to step into the lending business providing superior lending to small and medium-sized firms as well as to end-users. According to other research, within a few years, Big Tech companies will be able to succeed in overtaking the origination and distribution of loans to consumers and SMEs. Mano and Padilla (2018) argue that this may ultimately harm competition, reduce consumer welfare, and bring an increased risk of financial instability in the medium or long term. Mano and Padilla thus argue that the current oligopoly of a few banks is replaced by an even worse performing oligopoly of a few global Big Tech companies. However, one may argue that the global Big Tech companies are not competing in a traditional oligopolistic market, but in highly dynamic moligopolistic market. The conclusion that consumers will be worse off will then be debatable.

In theory, however, the competitive process in the banking sector may benefit from reciprocity in data sharing regardless of the scenario being played out. The question is whether this is also practical and what are the costs in terms of privacy and incentives to innovate?

5.2.4 Are there bottlenecks for accessing data?

Given existing PSD2 regulations (European Commission, 2015), there are no bottlenecks for the Big Tech firms to access data related to customers' accounts at banks. Therefore, at present, there are no bottlenecks for the Big Techs for accessing essential data for the provision of financial services (under PSD2). Under this regulation, Big Techs are not required to make non-traditional data available to banks that can complement the banks' financial services. Under the GDPR, however, platform users may instruct platforms to share data with their banks. Furthermore, banks and financial institutions can also obtain non-traditional data elsewhere, and they do. Banks have long realized that non-traditional data, when combined with traditional data and analytics, can improve the acquisition and retention of clients, and it allows for cross-selling and upselling.[121] As a result, financial industry players collect non-traditional data via the Internet and mobile networks, obtain data from IoT devices, swipe cards, ATMs, card readers, and sensors. They can observe non-traditional data directly and also get non-traditional data from Big Tech firms and other third-party data suppliers.

Banks and financial institutions cannot, however, access "all" Big Tech data. There are a few examples when Big Tech platforms decided not to grant financial institutions access to their data for fear of negative side-effects on their own core business. For example, Facebook blocked one of the UK's biggest insurers, Admiral Insurance, from using its user data. This insurance company wanted to use social data for risk profiling and to set insurance premiums accordingly.[122] Facebook blocked data access claiming that they protect the privacy of its users, and with such profiling, Admiral

[121] For example, McKinsey (2019) cites an example of a bank using credit-card transactional data "*to develop offers that gave customers incentives to make regular purchases from one of the bank's merchants*". See https://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value

[122] Admiral Insurance wanted to identify personality traits through examining posts and likes of Facebook users. The insurance company would identify users *"who write in short, concise sentences, use lists, and arrange to meet friends at a set time and place, rather than just "tonight", would be identified as conscientious. In contrast, those who frequently use exclamation marks and phrases such as "always" or "never" rather than "maybe" could be overconfident." See https://www.theguardian.com/money/2016/nov/02/facebook-admiral-car-insurance-privacy-data*

Insurance would violate Facebook's platform policy.[123] A privacy issue is that Facebook users may find it challenging to opt out because the financial disadvantage in doing so becomes so significant that they have no other option but to grant the insurance company access to their data. Another problem is the adverse selection being leveraged into the platform. High-risk clients would be less inclined to join (or actively use) the social network to escape the risk premium that the insurer would otherwise apply to them.

Big Tech firms have better, more exhaustive access to non-traditional data (e.g., social media relationship, behavioural profiles) than banks, due to their own market activity. Therefore, using the previous example of lending, Big Tech firms are likely able to better predict borrowers' losses and defaults than traditional financial institutions. These data become increasingly important for banks to withstand the growing competitive pressure from Big Techs in the provision of financial services. Not having access to the Big Tech data might create a competitive disadvantage to banks regarding certain financial services. Possibly, banks are forced to focus on providing some core functions in the financial system and leave specific services to others. In such scenario, banks would face similar challenges as telecom operators who have also been forced by Big Techs to focus their activities on the provision a key infrastructural service. Such development implies a certain degree of commoditisation of the services that banks offer.

5.2.5 *Is data currently being shared (voluntarily, on a commercial basis)?*

Banks can collect a lot of non-traditional data directly or buy them from third parties; they do not need have a data-sharing agreement with Big Tech firms.

Not all commercial agreements are in the public domain. Still, at the time of writing, we are not aware that personalized data on search (queries and clicking behaviour) or detailed data on social network relationships would be directly accessible to financial institutions, even with customer consent.[124] Big Tech firms, however, did create data-sharing agreements with third parties (developers), to support apps that run on their platforms. Banking and financial services apps have access to the same Big Tech firms' data, under the same terms and conditions as other app developers.[125] There are a few examples, however, where Big Tech data is mined and used to gain insights. Kabbage, for instance, provides small businesses with an automated line of credit. Kabbage automates its credit decisions based on several data factors, including business volume, time in business, transaction volume, social media activity, owner's credit score.[126]

Essential data from Banks to Big Tech firms are shared under PSD2.[127] At the time of writing, we are not aware that banking data outside of PSD2 would be available to Big Tech firms (and existing banking regulations would not even allow for such data exchange). For limited sharing of banking data on a commercial basis, we can mention certain partnership agreements between Big Tech firms and banks where the primary nature of the deal is not related to the data, but a limited data exchange must take place to fulfil the agreement. One such example could be Amazon's

123 Facebook's Platform Policy (3.15) states that developers cannot "use data obtained from Facebook to make decisions about eligibility, including whether to approve or reject an application or how much interest to charge on a loan." See https://developers.facebook.com/policy/

124 Article 6(1)(b) of the GDPR would allow the combination of personal data held by the BigTech firm with the traditional financial information of the bank, but we are not aware of commercial data exchange of this type.

125 For terms and conditions to access data, for example, you can visit: https://developers.google.com/terms/api-services-user-data-policy (Google), https://developer.linkedin.com/docs/rest-api (LinkedIn), or https://developers.facebook.com/docs/graph-api/using-graph-api (Facebook).

126 https://www.valuepenguin.com/small-business/kabbage-business-loan-review

127 Or, it is shared under similar regulations, such as Open Banking rules in the United Kingdom.

partnership with Bank of America for its lending program,[128] or Google's partnership with Citigroup and Stanford Federal Credit Union to offer checking accounts.[129]

5.2.6 What is the impact of mandatory data sharing on innovation incentives?

Regarding data sharing by banks under PSD2, there will be no impact, because the data involved is typically observed/provided and can be regarded a by-product of providing financial services. A possible obligation for Big Tech firms to share data may not necessarily affect their incentives to invest in producing these data, provided it only concerns data that was provided to or observed by them, and not the data which was inferred or derived from these data.

The entry of technology-focused firms into financial services offers innovation and social benefits. The benefits include possible reductions in the cost of financial services for consumers, both at the retail and institutional level, through improved efficiencies. Customers will be able to access financial products that are cheaper, more convenient, tailored, and accessible.

PSD2 created incentives for the Big Techs to develop new, innovative services because their services are typically delivered through a platform already used extensively by the customer. Big Tech firms may also improve financial inclusion and facilitate access to markets that were previously untapped, a particularly important benefit for the unbanked population and a large portion of SMEs.[130]

Banks argue, however, that PSD2 introduced asymmetric regulation and that the mandatory and asymmetric data sharing between financial service providers and Big Tech companies could cut off part of the banks' current revenue stream, decrease their profit margin, and ultimately lower the banks' incentives for innovation. Reciprocity in data sharing would ensure that Big Techs and banks would compete on a level playing field.

Big Techs' position is that such an asymmetry does not exist, Banks today can obtain all essential information to offer their financial services. They also argue that a mandatory data sharing regime could possibly lead to underinvestment in data monetization and undermine data-driven business models.

5.2.7 What are the risks of data sharing?

The financial services sector handles sensitive information about individuals and enterprises. With the digital transformation in the finance sector, more and more customer-provided data is available in digital formats, and more and more information is observed during transactions. Combining big data sets makes it easier to generate insights but also makes the data more susceptible to security breaches. Also, the combination of large data sets, even if the data is de-anonymized, has the potential to privacy vulnerabilities.[131] Because of this, security and privacy vulnerabilities are threats to the rise of big data solutions in the finance sector.

Consumers may benefit from data sharing, but data ubiquity, and consequently, data security, are proving to be a significant challenge for banks and Big Tech firms alike. The more data is collected

---

[128] In this deal, Bank of America is providing capital for Amazon Lending and the primary reason for the deal agreement was not to exchange banking data. See https://www.americanbanker.com/slideshow/how-amazon-is-shaking-up-financial-services

[129] See https://www.reuters.com/article/us-google-finance/google-pay-to-offer-checking-accounts-through-citi-stanford-federal-idUSKBN1XN1IQ

[130] Big Tech firms' such offerings are on markets which are typically outside of the European Union.

[131] Data sets that are anonymized can often be de-anynomized. De-anonymization is the practice of matching anonymous data with a publicly available data set(s) to discover the person to which the data belong (Lubarsky, 2017).

and shared, and the larger the number of participants in the data sharing process, the more significant the security and privacy risks.

Of course, the risks of data sharing can be managed. Banks already must comply with strict regulations for safeguarding people's privacy and the integrity of the retail payment infrastructure, and they have a good track record in doing so. Big Tech firms also have gained ample knowledge and experience to handle data security, privacy, and compliance requirements by design.[132]

5.2.8 What are the options under current legislation (general and sector-specific) to make data sharing mandatory?

The question for this specific case is not the data sharing, but if (and how) reciprocity might be reached. PSD2 already allows for data sharing from the finance firms to Big Techs.

- According to Article 36 of the PSD2 Directive, Payment Initiation Services Providers (PISPs) and Account Information Services Providers (AISP) have a right to freely access their customers' accounts data, under the same customer's consent to provide payment services or account information services.
- According to Article 31 of the Regulatory Technical Standards (RTS), banks shall comply with each of the following requirements:
  - they shall provide account information service providers with the same information from designated payment accounts and associated payment transactions made available to the payment service user when directly requesting access to the account information;
  - they shall provide payment initiation service providers with the same information on the initiation and execution of the payment transaction provided or made available to the payment service user when the transaction is initiated directly by the latter;
  - they shall provide payment service providers with a confirmation whether the amount necessary for the execution of a payment transaction is available on the payment account of the payer.

Big Techs collect personal data about individuals, and this personal data could be ported to Banks and Financial Service providers using GDPR provisions and with the users' consent.

So, the argument that the PSD2's data sharing obligations are asymmetric, and that this asymmetry harms the competitive position of banks vis-à-vis Big Techs, is not entirely valid. This does not mean, however, there is no asymmetry at all, but this does not relate to data sharing. Currently, new business activities of the banks are subject to the same regulations as their traditional banking activities. Therefore, as they argue, banks do not compete with new entrants on the same rules and same terms, A possible solution lies in the notion of "activity-based regulation" that poses fewer restrictions on banks that want to innovate. Currently, new business activities of the banks are under the same stringent regulatory requirements as the traditional business activities. Therefore, as they argue, banks do not compete with new entrants on the same rules and same terms.[133, 134]

Banks are in favour of cross-sectoral data-sharing based on a mix of GDPR and PSD2. A problem with GDPR is that there is no imposed format for the data exchange, nor are there specific requirements and conditions for real-time data transfers. The existing GDPR regulation could be amended (maybe in the form of a domain-specific Regulatory Technical Standard) that addresses the challenges of continuous mechanisms for such (near real-time) data exchange. A non-domain

---

[132] For more details, for example, how Google addresses the requirements of the finance industry, See https://cloud.google.com/security/compliance/financial-services

[133] ING argues for the principle of "same services, same risks, same rules". See https://www.ing.com/About-us/Regulatory-international-affairs/Viewpoints/Financial-innovation-and-licensing.htm

[134] It is also worth noting that the activities of Big Tech firms in financial services may be subject to regulation in certain jurisdictions. Similarly, there are other jurisdictions where the legal framework prohibits deposit-taking banks, or their corporate affiliates, from engaging in commercial activities.

specific amendment to the GDPR could also be introduced with that applies to all sectors (banking and non-banking), with real-time access through APIs, and standardized.

5.2.9 Policy options

There are two fundamentally different models for how reciprocal access might be addressed. The first model is to introduce data sharing requirements only within the domain of financial services and only between intra-industry participants. This model would assume Big Tech firms would be required to separate their activities in the financial markets from their other businesses. In this model, the Big Tech firms would be allowed to enter the financial services market via a structurally separated entity (which would not be allowed to share the obtained data with other entities), and they would need to provide data access to the incumbent banks and other financial services providers in the same terms. With such structural separation, Big Tech firms themselves would not offer financial services and they could stay neutral to all players in the financial industry.

The second model would be opening data sharing requirements to non-industry participants. This model would address cross-sector data exchange. The raw data held by companies in all industries would be accessible by any firm on similar terms and conditions, upon the request of an end-customer. Banks would have to make traditional data accessible to non-industry players. In contrast, other companies would have to do the same with their raw data. This would enable the end-users to manage full access to their personal data to any firm and take advantage of advanced product offerings from companies across many industries.[135]

The first model is easier to implement. The second model is more difficult as it requires coordination amongst several regulatory bodies across many sectors in many countries. One would also have to explore the technical feasibility (and possibly the delineation) with regard to the sharing of raw data. For both models to work, it would be desired to come-up with regulation at the EU level not to create fragmented regulatory instruments and to avoid regulatory shopping.[136]

From a purely technical point of view, reciprocity in data exchange can be achieved in either model and in various ways. For example, in return for getting access to the financial data of a particular customer, Big Techs could be required to make all raw customer data available to the Banks, in real-time. Alternatively, while providing financial services, Big Techs could be prohibited to combine financial data with personal data that they uniquely have about the particular customer. Technical feasibility, however, does not mean that it is worthwhile to implement reciprocity in practice. The main reason for this is that, currently, there are no guarantees that the cost of implementing reciprocity in data exchange would outweigh its perceived benefits considering the banking sector alone. Providing ultra-fast access to real-time big data has its implementation costs, and it is not simple to define reciprocity rules that do not leave much room for interpretation by lawyers in various jurisdictions, and simple to verify compliance. A level playing field in the domain of financial services is primarily not driven by the lack of reciprocal data access but by the lack of same rules for the same services. Big Techs do not fall under the same regulatory and supervisory requirements as banks, therefore, implementing reciprocal data access will, in itself, not create a level playing field and may introduce possible negative side effects related to privacy and data security risks due to potentially ambiguous data exchange obligations. While there may be good reasons for regulatory differentiation between banks and Big Techs, it is beyond the scope of this study to evaluate these reasons.

---

135 Institute of International Finance: “Reciprocity in Customer Data Sharing Frameworks”. See https://www.iif.com/portals/0/Files/private/32370132_reciprocity_in_customer_data_sharing_frameworks_20170730.pdf

136 The supervision of banking, payments, and e-money institutions are rather fragmented in the European Union. Banking regulation resides at the EU level, while payments and e-money institutions are nationally licensed and supervised. This led to a somewhat fragmented regulatory framework and resulted in regulatory shopping within the finance sector.

Rather than drafting new rules on reciprocity, the same effect can to large extent be realised on the basis of PSD2 and GDPR. However, in practice users (still) tend to be hesitant to instruct a bank and a platform to exchange data, which may be a sign of lack of trust. To overcome this barrier, the creation of a trusted (digital) environment for data sharing may help. ‘Trusted’ in this context means, among other things, that there are transparent and non-discriminatory rules for data exchange. In this trusted environment, the data access is transparently monitored, and rule violations need to be prevented. The rules shall allow parties to benefit from data sharing without experiencing negative consequences. This requires control by those involved about who has access to data that can be traced back to the data subject (person or company). The platform must also be based on a robust and reliable infrastructure that guarantees continuity and security. Simply imposing a law that obligates data sharing does not create such a trusted environment, nor does it create the technical standards for data exchange. This is something that must come from the market or driven by the policy makers.

6 Case: administrative business software

6.1 Case description

This case zooms in on possibilities for and potential effects of data sharing between companies in the market for administrative business software. The market for business software developers is characterized by a comparatively small group of software developers, such as Oracle, Microsoft, SAP and AWS, that is capable of developing and hosting complex, large-scale software infrastructures for large corporations and public authorities at national and decentralized levels of government. The complexity and scale of these systems originates from the need for integration of various databases, applications and cloud services, often supplemented with specialized solutions from smaller-scale developers that can run on the infrastructure of the leading developers. Alternatively, a supplier may provide both the data infrastructure and the key administrative functionalities, in an integrated software system (still, in such cases, the application may have to be able to work together with software from other suppliers). Important parameters are the interoperability of autonomous systems, as well as the possibility of interaction and data sharing between them, called data portability.

The role of data in the case of business software

As in the other cases, data represents a core function in the market for business software. However, its role is fundamentally different in this case: data is a key input for the users of the administrative software. Relevant types of data cover all administered information from both internal and external processes and operations. Software providers control access to that data by managing the infrastructure in which organisations structure and analyse the data retrieved from their portfolio of data sets and applications. Efficient use by the business users is facilitated through uninterrupted data portability and interoperability between solutions of various software suppliers. New solutions built upon the legacy software of a software infrastructure provider increase efficiency of the comprehensive network of applications and databases, facilitating economic benefit, such as synergies, more efficient internal processes, better informed decision-making, and increased innovation.

Business software users' operations can become dependent on their software infrastructure, which can, in turn, create a certain dependency on the provider of the software running on top. Dependency can frustrate the optimal functioning of an organisation's software infrastructure, especially when it limits the flexibility to integrate software applications of other developers to the infrastructure through data portability and interoperability. Predominantly, this occurs in a horizontal setting between providers of similar software, to the extent where users cannot (easily) cut the core supplier out or change to a competitor ('vendor lock-in').

The development, implementation and maintenance of data portability between solutions is a valuable part of software suppliers' business model, while the costs and risks involved with switching software suppliers enables existing suppliers to raise prices. As a result, suppliers can create significant barriers for switching between software providers and for alternative software providers to enter the market.

Therefore, this case explores the relationship between (limitations in) data portability and lock-in effects in the market for enterprise software. Since there may be a range of causes that make it hard for customers to switch to another provider, some of them not related to data, a key question is to what extent voluntary or mandatory horizontal data sharing (through data portability) can alleviate customer lock-in. The case study also aims to reveal what the potential impact would be on competition and innovation in the market for business software. It is not clear to what extent data portability (alone) could lower the barriers for switching and entry, nor is its impact on the need for acquiring expert knowledge for implementation and maintenance of new business software. This depends largely on its ability to circumvent legal licensing issues and technical legacy software issues, as well as what the involved costs would be.

6.2 Preliminary analysis

6.2.1 The market for business software

Demand side

The deployment of information technology (IT) products and services by governments and businesses have shown steady growth over the last decades. Figures on investment and maintenance costs in Table 6.1 underline this development. Especially the size and growth rate of investment costs for software stand out, compared to hardware and electronic networks in 2005, 2011 and 2017. Costs for intermediary utilization services, such as maintenance, show a similar growing trend in the reference period. The size and increase of IT investment led to studies on its drivers and on possibilities for cost containment and efficiencies.

Table 6.1: Overview of IT investment and maintenance costs in 2005, 2011 and 2017.

	Category	2005	2011	2017
IT investment costs by businesses and governments	Hardware	€ 3.7 bln	€ 3.8 bln	€ 4.1 bln
	Software	€ 10.8 bln	€ 15.1 bln	€ 21.7 bln
	Electronic networks	€ 1.9 bln	€ 2.0 bln	€ 2.8 bln
	Total	€ 16.4 bln	€ 20.9 bln	€ 28.7 bln
IT intermediary utilization* by businesses and governments	IT products & services	€ 28.7 bln	€ 33.1 bln	€ 51.2 bln

Source: CBS reports 'ICT, kennis en economie 2016' and 'ICT, kennis en economie 2019'.

For a part of the market, the demand for IT solutions is characterised by needs for flexibility and customisation. This drives the reliance of governments and businesses on external expertise.[137] The use of external expertise is relatively costly, and results from ongoing shortages of skilled in-house software experts. The shortage of skilled labour in software used to be predominantly an issue for public institutions[138], who could not compete with labour conditions offered by private parties. As a result, governments lacked in-house knowledge and experience, and had to outsource the development and maintenance of administrative software to specialized, private companies. The described deficiency is not restricted to government anymore; today, private businesses report similar problems in attracting sufficient software professionals.[139]

Supply side

With the rising demand for business software solutions, the supply side of the market has matured over the last decades. Globally, there are conservatively tens of thousands of business software vendors of all sorts of solutions, while for each solution there are many differentiated alternatives which compete to offer that combination of features and functionality that customers will want to

137 Ollongren, K. (2019). Toelichting op toename externe inhuur (ICT) personeel Rijk. Letter of government. Available at: https://www.rijksoverheid.nl/documenten/kamerstukken/2019/07/02/kamerbrief-over-toename-externe-inhuur-ict-personeel-rijk

138 Algemene Rekenkamer. (2019). Rijksoverheid heeft informatiebeveiliging en IT beheer nog niet op orde. Available at: https://www.rekenkamer.nl/actueel/nieuws/2019/05/15/rijksoverheid-heeft-informatiebeveiliging-en-it-beheer-nog-niet-op-orde

139 Trouw. (2018). Krapte ICT-arbeidsmarkt leidt tot problemen voor werkgevers. Available at: https://www.trouw.nl/nieuws/krapte-ict-arbeidsmarkt-leidt-tot-problemen-voor-werkgevers~b328e333/

choose. Moreover, with the advent of the public cloud, increasingly more organisations are - constructing their own enterprise software solutions.

At a first glance, the market for business software consists of a diverse and sizeable group of software providers. It is, however, important to distinguish providers at a more granular level. A crucial development is the shift of the business software market to a more cloud-based environment. Traditional and offline IT systems are supplemented with, or in some cases replaced by, cloud business services. Generally, the most accurate distinction is made through a division in Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). These classifications each offer a progressive layer of abstraction on top of traditional IT environments, where all products and services were managed by the end user.

Source: IBM. (2020). IaaS. Available at: https://www.ibm.com/cloud/learn/iaas#toc-iaas-vs-pa-w12c-z4b

First, IaaS delivers the most fundamental type of cloud services, including virtualization, servers, networks, and storage to business users, while the users still have a range of other platforms and applications on the service provider's infrastructure. The cloud services are provided online, on-demand and on a pay-as-you-go basis, thereby enabling scaling resources up and down. For customers, this reduces the need for up-front capital expenditures or investment in unnecessary resources. For their service, IaaS providers have to host large physical data centres, which contain physical machines required to power the various layers of abstraction developed on top of them.

Because building and hosting a (global) network of servers and data centres requires huge upfront capital investment, the IaaS market is characterised by only a few competitors – with a top-tier of three to four leading competitors. For instance, organisations that are predominantly interested in systems that enable them to transfer massive data packages, like a streaming service such as Netflix, are bound to a few competitors that provide services that meet the requirements.

Supplementary to storage and other computing resources offered to IaaS users, PaaS offers a cloud-based environment, where customers can develop, run and manage prebuilt tools as tailored applications. Providers create the underlying infrastructure and offer solutions for managing, tailoring and integrating of applications, through abstracting away management of the operating system, middleware, and runtime. Purchasing organisations can either choose to do this management themselves or involve a third-party provider.

On the platform layer, each of the IaaS competitors offers their services, but there are lots of smaller competitors that build or integrate their own solutions on top of the infrastructures of the IaaS providers. Because of the more comprehensive and integrated nature of PaaS and the absence of a need to invest in physical data centres, there are significantly fewer barriers for smaller software vendors to compete in the market. Customers can demand tailored solutions that fit their needs in terms of features and functionality, as a result of the large variation and scalability of solutions. When compared to IaaS solutions, competition among competitors can be relatively limited, while the large extent of differentiation and customisation creates more frequent data portability and interoperability issues when users want to switch.

Finally, business users can even run single, specialized applications as a cloud service. This is called SaaS. Business users are granted access to vendor's cloud-based software, where they can use remotely hosted applications on their local devices, through the web or an application programming interface (API). Software can be either for a particular amount of time or for the lifetime. Mainly, SaaS application do not require any downloads as they are used directly through the web browser. The supply of software solutions is diverse: business users can make use of applications for enterprise resource planning, messaging or office software.

As with PaaS, SaaS providers build and integrate solutions to run on the systems of IaaS-providers. Given the large variation in possible SaaS solutions, their scalability in size and complexity, as well as their operability via web browsers, the barriers for providers to compete on this market are relatively limited. Also, given the market's fragmentation, customers are able to demand tailored solutions that fit their needs in terms of features and functionality.

All in all, as with each type and form of business software, the level of competition in the market for cloud-based business software is influenced by the scale and sophistication of software services required.

Causes of vendor lock-in

The scale and sophistication required for business software projects of governments and large corporations limits the number of software providers that is able of designing and hosting the core infrastructure, especially for IaaS and PaaS. In general, procuring actors will seek an integral solution for their administrative systems. In such a solution, various internal sub-systems can communicate and interact with each other and share data (*data portability and interoperability*). For example, municipalities may easily operate over a hundred software applications for their public services, many of which need to be interconnected to varying extents.[140] Furthermore, providers take care of ongoing management and development of integration of internal technical solutions. As a result, business users have by and large one primary point of external contact for IT related matters. The need for comprehensive, sizeable infrastructure favours large solution providers over smaller counterparts.

To illustrate this concentration: on a total of 89 ongoing Dutch government IT projects with a value of over € 5 million in 2018, Capgemini was active in 20 of these projects, IBM in 14 and ATOS in 11.[141] From a competition perspective, this is a relatively small group of leading, large software firms that is serving a considerable share of the market for large-scale software projects. Overall, their scale, resources and technical expertise create barriers to enter the market for newcomers and smaller software firms.

---

140 NRC. (2015). Gegijzeld door de softwareboer. Available at: http://vvoj.org/wp-content/uploads/Productie-De-falende-ICT-bij-de-overheid-06.pdf

141 Computable. (2019). Ict-kosten Rijk fors gestegen in 2018. Available at: https://www.computable.nl/artikel/nieuws/overheid/6666680/250449/ict-kosten-rijk-fors-gestegen-in-2018.html

In addition, purchasers tend to stick to the systems created by these software suppliers, as maintaining existing systems through continuous development and maintenance is economically preferable over investing in comprehensive system upgrades. As a result, the core of software infrastructures is often built on licensed solutions which become so-called *legacy* systems over time – systems which run on outdated technology and need continuous updates and upgrades to comply with today's technical and functional software requirements.[142]

Furthermore, these systems offer limited technical compatibility with software solutions from other and sometimes even the same supplier. To function properly within the internal processes of an organisation, these legacy systems either need to be replaced by new systems or require continuous maintenance and development. Although the cost of fixing the legacy systems recur regularly, they are generally lower compared to the cost of implementing a new system, making them an appealing business case to purchasers – especially because the pace of software innovation quickly transforms a newly implemented system into a legacy system as well.[143]

The key problem with legacy software is its limited compatibility with new software from other suppliers, which is amplified by the use of proprietary, licensed solutions by software suppliers.[144] Data accessibility and exchange is core to the optimization of organisation's processes, but is not naturally supported between solutions. This reinforces the edge of incumbent big suppliers over smaller or new competitors.

The preferential position of big software suppliers causes organizations to become locked-in with their dominant software supplier.[145] This 'vendor lock-in' entails that organizations cannot (easily) integrate software solutions into the system of the primary provider or switch the entire system altogether, without risking the continuation of their administrative system's functioning.[146] Especially at the horizontal level, where users' fear of disruptive and costly software switches to competitors strengthens the position of incumbent providers in keeping their clients and charging them considerable service fees for continuation of their services. As an illustration, an estimated 72% of companies' IT budgets are dedicated to keep existing software systems running. [147] Each new application has to be made compatible to the existing network of applications for optimal functioning. The European Commission reported that 42% of the organizations monitored in a study on vendor lock-in, experienced an ICT lock-in.[148]

Some business software customers deliberately select smaller providers for their software, or providers that ensure operability with other software providers; however, when these smaller providers gain momentum, leading providers might acquire them to integrate their knowledge and skills, and avert the rise of a significant competitor. As a result, customers may end up a large software providers, even if they try to prevent this. This reinforces vendor lock-in.

---

142 The Global Treasurer. (2019). The problem of legacy tech. Available at: https://www.theglobaltreasurer.com/2019/10/25/the-problem-of-legacy-tech/

143 Hernaes. (2019). Why is legacy tech a problem and how do we fix it? Available at: https://hernaes.com/2019/03/04/why-is-legacy-tech-problem-and-how-do-we-fix-it/

144 Proprietary, or closed-source software holds the source code safe and encrypted. Meaning, the user can't copy, modify, or delete parts of the code without some type of consequence. It can go from voiding the warranty to even legal repercussions. Source: https://www.veriday.com/blog/open-source-vs-closed-source/

145 NRC. (2007). Belastingdienst werd afhankelijk van IBM. Available at: https://www.nrc.nl/nieuws/2007/08/11/belastingdienst-werd-afhankelijk-van-ibm-11373262-a158178

146 Webwereld. (2010). Vendor lock-in grootste probleem cloudhosting. Available at: https://webwereld.nl/nieuws/business/vendor-lock-grootste-probleem-cloudhosting-3748618/

147 The Global Treasurer. (2019). The problem of legacy tech. Available at: https://www.theglobaltreasurer.com/2019/10/25/the-problem-of-legacy-tech/

148 European Commission. (2016). Study on best practices for ICT procurement based on standards in order to promote efficiency and reduce lock-in. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=14434

Consequences of vendor lock-in

Although, a trustworthy software supplier that understands a client's existing systems and future software needs is in essence a perfect business partner, the situation of a vendor lock-in creates perverse incentives for software providers. Reports of how software providers make use of their powerful position at the expense of their clients are numerous and date back to 2011.

With a recent publication, Dutch chief information officers (CIOs) renewed political attention for the matter, leading to Parliamentary questions about the dependency on software providers.[149]. United in the CIO Platform Nederland, over a hundred large software buyers expressed concerns about the stranglehold of software giants. According to the CIO Platform and news reports, these firms abuse the asymmetric distribution of technical expertise and the limited mobility to change from one supplier to another in various ways to optimize their revenues. For instance, by overcharging clients for their services, by delivering incomplete or outdated services that need additional add-ons, by stopping the development of (security) updates of software, or through systematic budget overruns. [150] [151] [152] [153] [154]

Another frequently applied method is through conducting onsite inspections at the client's premises. These inspections aim to reveal unauthorized use of software licenses and are agreed upon in contracts; however, if organisations have used too many licenses, which often happens unintentionally, software providers issue claims ranging up to billions of euros by charging list prices instead of the discounted user terms agreed upon in initial contracts. [155] [156] [157] Moreover, software providers surprise clients with unannounced and one-sided mark-ups in license agreements (increasing prices by a multiple of two or three).[158] [159] In some cases, it was reported that instead of warning clients in advance that the terms and conditions of licenses had changed, providers decided retrospectively that clients were 'non-compliant' to new rules, and issued claims adding up to more than three billion euros.[160] [161]

Nonetheless, the CIO-platform provides just a sub-sample of Dutch multinationals, and not all organisations report similar issues. It could be argued that the switching costs are simply considerable and inevitable given the nature of software products. However, these switching costs are not so much locking clients in at the back-end of the service agreement with a provider, but are the result of considerations of purchasers at the frontend of the service agreement, when

---

149 Ollongren, K. (2019). Kamerbrief over berichtgeving Financieele Dagblad IT leverancier Oracle claimde miljarden bij Rijksoverheid. Letter of government. Available at: https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/28/kamerbrief-over-berichtgeving-fd-it-leverancier-oracle-claimde-miljarden-bij-rijksoverheid

150 Webwereld. (2014). XP-plakken kost Rijk minstens 2,5 miljoen. Available at: https://webwereld.nl/nieuws/business/xp-plakken-kost-rijk-minstens-25-miljoen-3768401/

151 NRC. (2015). Gegijzeld door de softwareboer. Available at:http://vvoj.org/wp-content/uploads/Productie-De-falende-ICT-bij-de-overheid-06.pdf

152 Computable. (2011). Rijk verspeelt doelbewust 5 miljard aan ICT. Available at: https://www.computable.nl/artikel/nieuws/overheid/4230545/1277202/rijk-verspeelt-doelbewust-5-miljard-aan-ict.html

153 Computable. (2014). ICT-stelsel bij de overheid is rot. Available at: https://www.computable.nl/artikel/achtergrond/overheid/5085768/1444691/ict-stelsel-bij-de-overheid-is-rot.html

154 Computerworld. (2014). Belastingdienst verzuipt in moeras van IBM en Oracle. Available at: https://computerworld.nl/development/84068-belastingdienst-verzuipt-in-moeras-van-ibm-en-oracle

155 NU.nl. (2013). Rijk wil snijden in softwarekosten Oracle, SAP en Microsoft. Available at: https://www.nu.nl/internet/3378521/rijk-wil-snijden-in-softwarekosten-oracle-sap-en-microsoft.html

156 Therp. Overheid leert van Oracle incident. Available at: https://therp.nl/blog/our-news-1/post/overheid-leert-van-oracle-incident-38

157 Het Financieele Dagblad. (2019). IT-leverancier Oracle claimde miljarden bij Rijksoverheid. Available at: https://fd.nl/economie-politiek/1320235/it-leverancier-oracle-claimde-miljarden-bij-rijksoverheid

158 Channelweb. (2016). Tariefstijging Vicrea kan oplopen tot 1150 procent. Available at: https://www.channelweb.nl/artikel/nieuws/overheid/5865604/5226433/tariefstijging-vicrea-kan-oplopen-tot-1150-procent.html

159 NRC. (2015). Gegijzeld door de softwareboer. Available at:http://vvoj.org/wp-content/uploads/Productie-De-falende-ICT-bij-de-overheid-06.pdf

160 Het Financieele Dagblad. (2019). The stranglehold of software giants. Available at: https://fd.nl/ondernemen/1319877/the-stranglehold-of-the-software-giants

161 Het Financieele Dagblad. (2019). IT-leverancier Oracle claimde miljarden bij Rijksoverheid. Available at: https://fd.nl/economie-politiek/1320235/it-leverancier-oracle-claimde-miljarden-bij-rijksoverheid

negotiating a deal and choosing one provider over another. Notwithstanding whether these considerations were either deliberate and thoughtful, or insufficiently informed initially, they could lead to considerable transaction costs for future software-related choices.

The accelerating pace of technological development and (its corresponding) customer demand for business software causes a continuous need for new, innovative connections between systems and new functionalities. However, uncertainty and fear for high costs lead both customers and developers to ty down contracts for software use and development to the last detail. This lack of flexibility causes tensions between both parties in the remainder of the contract, as the time frame of the contract does not correspond with the dynamics of the market and technological development.

Although both supplier and purchaser play a part in concluding risk averse contracts, the supplier can direct negotiations in the preferred direction through their superior bargaining power. As a result of the considerable transition costs when changing software supplier, suppliers are in a position to narrow down possibilities in future contracts once the first contract is concluded, and demand considerable price premiums and upgrades for additional software solutions.

The considerable costs for transition and lack of flexibility in existing software contracts limit purchasers. The degree of limitation is disputed, however. An expert indicated that transitions from one large software supplier to the other happen at most once every ten to twenty years. Uncertainty and limitations would thus only be experienced incidentally. The other way around, it could also be argued that costs for transition and limited portability and interoperability in case of transition create a disincentive to switch software providers in the first place.

Either way, experts agree that switching software suppliers is a disruptive and costly process in terms of time and costs, which influences the frequency of transitions and the degree to which purchasers perceive it as a barrier – and software providers are aware of this anxiety and exploit it by raising prices for its clients.

This makes upfront negotiating of the terms of the initial service contract crucial. Investing time and effort in negotiating contracts with software providers that are mutually beneficial, and creating an in-house, skilled IT (procurement) capacity that reflects the size and growth of an organisation, as well as the complexity of its (software-related) operations is vital. [162] [163]

6.3 Analysis of research questions

6.3.1 Which role does data play in the business model and what are the benefits that data provides?

Operational data from clients represents a core function in the market for business software. It is the key input for the business software users, and as such is at the core of the business model of software suppliers. Software providers control access to that data by managing the infrastructure in which organisations structure and analyse data.

For business software providers, as opposed to other cases about data use in companies, it is therefore not so much about the extensive creation, collection and analysis of valuable or intimate personal data, or about which company has more or better access than others. Privacy is also a less relevant issue, because this type of software providers typically maintain the infrastructure in

162 Het Financieele Dagblad. (2019). The stranglehold of software giants. Available at: https://fd.nl/ondernemen/1319877/the-stranglehold-of-the-software-giants

163 Het Financieele Dagblad. (2019). IT-leverancier Oracle claimde miljarden bij Rijksoverheid. Available at: https://fd.nl/economie-politiek/1320235/it-leverancier-oracle-claimde-miljarden-bij-rijksoverheid

which the actual data is stored and have no access to the underlying data. If they do have access to data, it is because the data is meant to be archived, aggregated, analysed, combined, and so on, for administrative, operational, or other business purposes.

In this case, impact is rather created through the organisation's data infrastructure in a portfolio of applications, data sets and sources. Uninterrupted data portability and interoperability between licensed, legacy solutions of software suppliers facilitates efficient use by business users. New solutions built upon that legacy increase efficiency of the comprehensive network of applications and databases, facilitating economic benefit, such as synergies, more efficient internal processes, better informed decision-making, and increased innovation.

Benefits can be expected from a societal perspective as well. First, a well-functioning data infrastructure offers a more secure mode of data exchange between applications, as compared to, for instance, sharing data via e-mail. Moreover, governments can use the integration of applications and their datasets for more evidence-based monitoring and policymaking, better informed policy and services to citizens.

However, unifying the categorization of personal or business data throughout the economy, to facilitate its use and transfer through a standardized and exchangeable machine-readable format is a complex matter. Early examples of attempts to standardize all generic standards and data formats for software systems in one sector, such as in the oil industry[164], proved to be a slow and painful process within one sector – let alone introducing such standards for the entire European economy.

The concentrated market for business software frustrates free data flow between solutions, as market players have no incentive to contribute to the proposed standardization. In sectors where vertical and horizontal data sharing is (potentially) beneficial to both involved parties, such as in e-commerce or social media, market players have been putting considerably more effort into data consolidation and have been more successful in creating infrastructures that are interoperable and allow for portability of data.

6.3.2 Which data and types of data are indispensable?

Instead of focusing on direct data sharing between software providers, the focus is on the infrastructure for data portability and the impact of mandatory or voluntary initiatives to develop such infrastructures on the freedom of choice for business software users and its impact on competition between business software developers.

Relevant types of data cover operational data from internal and external processes and operations administered by software users, the organisation's social data stored in CRM and SRM systems, and transactional data. Just how versatile operational data can be, is shown by looking at the example of a medium-sized Dutch municipality. Such an organisation easily offers tens of public services – such as garbage collection, services related to permits and personal documentation, social security, education and urban planning – which are all administered internally to keep track of relevant characteristics, such as the progress, frequency and quality of service delivery. Each task, in turn, requires the collection of different (types of) specific data. Furthermore, for optimal internal processes and service delivery to clients, various datasets within each of these services need to be connected through applications that are interoperable and able of exchanging data. Moreover, in some cases, communication between applications for disparate services is required, creating the need for interaction between administrative software applications. All in all, municipalities easily have over a hundred different software applications for the public services they provide.

[164] ISO. (2014). ISO 10303-209:2014. Available at: https://www.iso.org/standard/59780.html

6.3.3 *Can data portability facilitate entry or strengthen competition?*

As remarked above, in this case we do not address access to indispensable data, but the possibilities for data portability between administrative systems (and their infrastructures in particular).

Reasons for and impact of limited software switching

Discontinuity of software or obstacles in the transition to alternative software pose a serious threat for customer's business processes, hindering free movement between software providers once the infrastructure is implemented. After a first contract period, generally agreed for a period of three years, software users have to renegotiate the contracts terms; however, no customer would like to switch software provider again, because of the significant transition costs, implementation issues and corresponding internal resistance to change. This strongly limits free customer movement and competition in the market for business software.

Three factors influence the costs for transition between software systems. First, the costs in money and time to renegotiate the contract. Second, the integration of new software in the customer's internal processes. And finally, uncertainty about the ability to reuse all data in the new software system. Data has to be transferred in formats that operate in both the old and new system – if these systems do not offer data portability, data will first have to be transformed.

Once a client is locked-in, 'aftermarkets' start playing a significant role. Initially, costs for acquiring software solutions are relatively low, but once a customer is in, the supplier works with considerably higher fees for continuation of services. Incumbent software providers are aware of the transition costs faced by users. They often can raise prices in their extension offers just below the price for transition to a competitor. As a result, customers tend to stick to their primary supplier as it is more economically advantageous than implementing a new software system.

Moreover, interview partners claimed that software providers are primarily interested in selling more solutions to their customers, rather than in optimizing the use of solutions that are already in the customer's software portfolio. This is reflected in the incentives created for sales personnel that is rewarded based on targets for sales of additional software solutions and not for optimization of software that is already in use. For instance, providers promote cloud services to their clients as an option to increase flexibility and portability, but only offer it as an add-on to their existing software system. As such, generally, customers can only scale-up their system with new solutions, while providers are reluctant in offering possibilities to scale-down costs through repelling redundant solutions. This is especially the case when a provider initially offered 'on-premise' software and starts offering cloud services.

Overall, business clients are tied to their primary software supplier, due to limited data portability and limited interoperability between administrative systems of various providers. Especially at the horizontal level, where users' fear of disruptive and costly software switches to competitors strengthens the position of incumbent providers in keeping their clients and charging them considerable service fees for continuation of their services. As such, choosing a software supplier creates path dependency within an organization: with each additional solutions or service from a given supplier, a client becomes more locked-in with this supplier. The more unilaterally software is provided by one supplier, the more dependent a buyer becomes on their software.[165] In turn, it becomes increasingly costly to replace the existing administrative software with competitors'

---

165 Het Financieele Dagblad. (2019). The stranglehold of software giants. Available at: https://fd.nl/ondernemen/1319877/the-stranglehold-of-the-software-giants

software, while it is comparatively less expensive to get future software applications from the same supplier.[166]

Effect of data portability on competition

With the emergence of cloud services in business software in recent years, more and more developers moved away from IT systems based on offline, purely closed-source or proprietary software services towards solutions that allow for more transparent and open software systems. These modern services operate in the cloud, based on open standards or on open source.

Leading software developers have offered more transparency in the ground rules and codes of their software and partially published them online. This, in turn, creates platforms on top of which smaller SaaS developers of niche or small-scale software solutions are capable of developing software solutions that can run on the platforms and infrastructures of PaaS and IaaS market leaders. This enhances the business case of both leaders and newcomers: smaller developers can offer their solutions to all users of a market leader's infrastructure, while market leaders can offer a more differentiated portfolio of solutions that can be hosted on their infrastructure. This sharing incentive works predominantly in a vertical setting between large infrastructure or platform providers and smaller, specialized software solution providers.

From a horizontal perspective, however, there is less incentive for solution developers to coordinate and tailor solutions to enable easy switching. For instance, technical issues complicate a switch from one primary cloud storage provider to another resulting in risks of discontinuity of software systems or data loss. In addition to the technical barrier, facilitating a successful transition of solutions (in terms of portability and interoperability) is a costly process which further propels switching costs. An additional advantage of considerable switching costs to developers is that it boosts their ability to demand higher service fees for continuation of their services once they are in. All in all, there is no stimulus for suppliers to lower the barriers for a switch from their solution to horizontal competitors' solutions. Initiatives to lower lock-in should, therefore, mainly focus on horizontal competition.

Data portability could (in theory) lower some of the technical barriers for horizontal competition between providers of similar software solutions on the market, through establishing the fundament for interchangeability of their applications. As such, it mitigates the risk of a vendor lock-in and it facilitates entry into the market.[167]

We note that data portability is not the only restricting factor in competition. For example, in the market for IaaS, competition is limited to a handful of operators in the market, because these are the only ones financially capable of providing the solutions. Their financial position enables them to build and run a global network of large-scale data centres, which is impossible for smaller competing firms due to the complexity and the high investments costs involved.

6.3.4 Are there bottlenecks for porting data?

The need for sharing or integrating with other data sources or organisations developed quickly only in the last decades. However, most initial information systems were initially not designed with the idea that its contents or functionalities should be replaceable, sharable or integrative with other

[166] Verdonck, Klooster & Associates, Berenschot and RAND Corporation. (2016). Mogelijkheden om de afhankelijkheid van ict-leveranciers te verminderen. Available at: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2016/02/15/mogelijkheden-om-de-afhankelijkheid-van-ict-leveranciers-te-verminderen.pdf

[167] Computable. (2020). ACM onderzoekt ict-systemen in de zorg. Available at: https://www.computable.nl/artikel/nieuws/zorg/6876203/1276929/acm-onderzoekt-ict-systemen-in-de-zorg.html

systems. Often, the systems were offline and proprietary, functioning solely internal purposes, such as measuring inventory and production or storing personnel information.

Consequently, traditional proprietary software can frustrate the optimal functioning of an organisation. A manifesto drawn up by employers' associations gives insight in the consequences of this for Dutch healthcare organisations.[168] In the manifesto, issues arising from IT systems that cannot exchange data are said to have a detrimental impact on both efficiency and privacy. For instance, without the possibility of secure and scalable data exchange the use of technology is inefficient. Data exchange is currently difficult for health providers between the infrastructures of various national and regional institutions as open standards are not implemented at all or in different ways. Also, because data exchange between IT systems is impossible, it happens through outdated methods that are unsecured and diffuse.

This example shows the drawback of not being able to port and share data across different IT systems within a sector. Across sectors, the barriers for porting and exchanging data may even be larger because each organisation has different internal systems and needs, their operational issues are different as well. While a municipality has to store and integrate various data sources for a variety of diverging public services, a globally operating multinational faces issues with systems throughout the world that should measure similar activities in a similar manner and format, overcoming latency issues and legal issues regarding software licences. Alternatively, for a medical organisation the time-critical element of a system might be most important, while for a chemical firm the accuracy of the information in the system is most important.

Technical, legal and financial bottlenecks

In section 6.3.3, three influential factors for transition were mentioned: complexity of integration of new software time and costs, uncertainty about the ability to reuse all data in the new software system, and finally the costs involved in terms of money and time. The first two relate to the technical feasibility of facilitating data portability between software from different software providers. Beside the actual technical feasibility, it is also relevant to asses a user's legal power to enforce disposal of his own data in agreements with software providers: for instance, could they legally reject an organisation's desire to withdraw his data from their software? The third, then, relates to the financial barriers to enforce data portability, which in this case predominantly comes down to the switching costs involved.

Technical feasibility

As described in detail in section 6.3.3, developers' willingness to bridge technical barriers between similar systems ('substitutes') is limited, as it does not benefit their business case. Currently, if there are technical limitations to making legacy core software interoperable, business users may be forced to migrate to entire new core system and will have to incur substantial switching costs. In this case, data portability is essential for migration processes. As a result, one might argue that mandatory data sharing requires a broader policy solution that focusses on portability between, and interoperability of, legacy core software (including access to data bases) which facilitate access for third party developers. Still, the issue remains whether it is technically feasible to create portability between each and every software solution.

Not so long ago, in the 90s and 00s, facilitating portability and interoperability between disparate systems and solutions of various providers still was technically challenging and required intensive commitment from skilled experts to create the necessary APIs. Nowadays, it is no longer an engineering challenge. In universities, students can easily do it through using Representational

---

[168] MKB Nederland & VNO-NCW. (2019). Samen vooruit: gegevensuitwisseling een ambitie voor in de zorg. Available at: https://www.mkb.nl/sites/default/files/nln19085_manifest_samen_vooruit_05.pdf

state transfer (REST) interfaces[169]. Technological advancement has made the creation of interoperable interfaces significantly less complex. As a result, stakeholders from both customer and vendor perspective have no doubt that creating interoperability is technically feasible for software providers: data portability can be done at different levels and in a unified way.

Although from IaaS to PaaS to SaaS, the portability of data becomes more challenging with each layer, there are ways of unifying this in terms of transparency on data portability. For IaaS services, data portability can create data lakes, such as envisioned in the Open Data Initiative, which can be facilitated relatively easily between the hand full of providers. Although there are just a few large-scale cloud infrastructure providers, competition among them appears fierce. For PaaS and SaaS, the extent of differentiation and customisation at the level of creates more frequent data portability and interoperability issues when users want to switch. Switching could be facilitated if the sector adopts standards that facilitate portability and interoperability between software providers.

Stakeholders agreed that almost everything is technologically possible in terms of portability and interoperability, as long as there is a well-defined customer demand. In many instances, sophisticated business users nowadays demand the ability to access and transfer their data. Solutions that cannot meet that bar will not be competitive in the crowded software markets.

Legal feasibility

Business user representatives feel locked-in to their arrangements with business providers, due to technical dependency on the provider and service contracts with rigid legal restrictions. However, according to providers, theorized restrictions from a legal perspective have no significant impact on portability and interoperability. Obviously both sides have a different perspective on this and there may be need for more legal clarity.

Financial feasibility

Providers argue that the underlying reasons for the perceived technical and legal lock-in could be traced back to the financial barriers for switching software (providers). In any level of software switch, there is going to be switching costs involved. Vendors argue that although this might assume an issue, this not problematic in itself: clients want to avoid lock-in with one provider, but switching costs are simply inevitable. However, switching costs are not so much locking clients in at the back-end of a service agreement, but are the result of deliberate commercial considerations of purchasers at the frontend of the service agreement, when negotiating a deal and choosing one provider over another. This implies an information asymmetry during the tendering process which may be taken away by a general legal framework.

The implementation of mandatory data sharing obligations, preferably at least at EU level, that introduces and enforces uniform standards to facilitate data portability might be a necessary step to create more technical flexibility and less uncertainties for business users.

6.3.5 Is essential data currently being shared (voluntarily, on a commercial basis)?

In essence, there is not yet a comprehensive, generally applied framework in place at national or European level that facilitates data sharing between companies in the market for administrative software – either voluntary or mandatory. Nonetheless, emerging initiatives stimulate increased vertical coordination and exchange between various market players.

An important recent development is the increased application of cloud infrastructures in business software in recent years. Leading software infrastructure and platform providers make use of more open and transparent systems, which created a platform infrastructure upon which smaller

169 These interfaces function like industry standards by facilitating uniform and predefined sets of stateless operations, not bound by a specific language or runtime, and format data in machine readable JSON formats.

developers of niche or small-scale software solutions are able to develop their own software solutions. This has enhanced the diversification of the market and the business case of both leaders and newcomers on the market. However, this sharing predominantly takes place in a vertical setting between large infrastructure or platform providers and smaller, specialized software solution providers.

From a horizontal perspective, there is less incentive for solution developers to coordinate and tailor solutions to enable easy switching, as this harms their own business model. In recent years, private companies as well as national and European public institutions have launched initiatives to facilitate horizontal portability and interoperability in an attempt to mitigate lock-in and switching issues. So far, however, these have failed to materialize a breakthrough in the creation of sector-wide horizontal data portability.

Voluntary initiatives

One such attempt by private companies is the Open Data Initiative, orchestrated by three big IT suppliers – Adobe, Microsoft and SAP. It aims to create a common data model, which enhances interoperability and data exchange between their disparate applications and platforms[170]. Under this model, business application data is pooled, and clients are able to read and write data from all three companies' business applications and build intelligent applications on them. In addition, the model understands relationships between data sources and harmonizes siloed data to create new value.[171]

Although the three software giants compete, there are rationales behind their partnership. First, they share a strong customer base. Second, if they decide not to partner up, other firms will become active in creating interoperability and data portability between data sources. As such, the initiative comes as a response to a broader trend of software integration, even among market rivals. This offers customers more supplier flexibility and urges software companies to meet clients' need for a tailored, integrated network of applications.[172] According to a representative of one of the three organisations, this initiative mirrors the need of the market, as well as their broader view on how the sector should look and work together.

Some experts are sceptic about the added value of these private sector initiatives, as such actors generally are driven by self-interest. In other cases where larger software suppliers offered no or low code software on which organisations could build their own software, the complexity and costliness of switching sustain the incentive to exploit customers.

National alternative

At the national level, several attempts have been initiated. In 2014, a parliamentary scrutiny[173] of government IT projects was launched, which led to the establishment of Supplier Lifecycle Management (SLM), as a concentrated effort to deal with the power of big IT providers. SLM functions as a centralized knowledge and management facility for its most important software suppliers, such as Microsoft, SAP, Oracle, IBM, KPN and Ordina.[174] It brings together technical and

---

170 More specifically, they mention Adobe Experience Cloud and Adobe Experience Platform, Microsoft Dynamics 365, SAP C/4HANA and S/4HANA in their press release. Source: https://news.microsoft.com/2018/09/24/adobe-microsoft-and-sap-announce-the-open-data-initiative-to-empower-a-new-generation-of-customer-experiences/

171 Microsoft. (2018). Adobe, Microsoft and SAP announce the Open Data Initiative to empower a new generation of customer experiences. Available at: https://news.microsoft.com/2018/09/24/adobe-microsoft-and-sap-announce-the-open-data-initiative-to-empower-a-new-generation-of-customer-experiences/

172 ISG. Data Sharing Is the New Normal for Enterprise Software. Available at: https://isg-one.com/research/articles/data-sharing-is-the-new-normal-for-enterprise-software

173 Tweede Kamer. (2014). Parlementair onderzoek naar ICT-projecten bij de overheid. Available at: https://www.tweedekamer.nl/sites/default/files/field_uploads/33326-5-Eindrapport_tcm181-239826.pdf

174 Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). Strategische I-agenda Rijksdienst 2019-2021. Available at: https://www.rijksoverheid.nl/documenten/rapporten/2019/01/01/rapport-strategische-i-agenda-rijksdienst-2019-2021

user experience, as well as procurement expertise with these suppliers. SLM focuses on enhancing IT's contribution to the government's strategic goals, and aims to optimize government-wide efficiencies and economies of scale in IT contracts and licences. The core of these activities concentrate on facing software suppliers as one integral public sector purchaser, that internally re-uses and shares contracts, licenses and application platforms.[175] [176]

The idea behind SLM is that better coordinated demand management of the government reduces the risk that organisations are played off against one another, or that redundant software licenses are acquired. Since its introduction, the SLM teams have been successful in achieving their objective. The concentration of central government's software demand in one organisation has increased their bargaining power vis-à-vis software suppliers. SLM teams created a coherent and strong counterpart, that forced big suppliers, such as Microsoft, to make concessions to their terms, functionalities and licences. Moreover, the SLM teams have forced Microsoft to make these concessions enforceable for all other Dutch software customers of Microsoft. Although this new procurement strategy needed considerable initial investment (about € 5mln), the saved investment costs are in the tens of millions of euros.

Still, in essence, the reach of SLM teams only covers ministries and other central government institutions, due to restrictions on central procurement for other (semi-)public institutions, such as municipalities or hospitals. Enhanced possibilities for aggregation of demand in IT procurement could increase the bargaining power of customers in more sectors.

Notwithstanding the successful approach of the SLM teams for central government institutions, many business software users still perceive to be locked-in. Recently, the Dutch national government presented its 'Strategic I-agenda Rijksdienst 2019-2021'[177], where it expressed its further ambitions regarding solutions for IT procurement issues. In essence, their goal is to transform their IT infrastructure into an information-driven system, where functionalities and data are at the core, rather than the suppliers of these functionalities. Implementation of government-wide interoperability and data portability of autonomous systems are ongoing and crucial in this regard, with the goal of enabling re-use and exchange of data between government branches.

In practice, the strategy comes down to a situation where government makes use of the market when possible, but it relies on own software development when necessary or favourable. An elaborate sourcing strategy will guide the use of market expertise. This strategy promotes a careful consideration of the use of products and services from the public cloud, and possibilities for own development and management of functionalities and applications. Also, the strategy should enable smaller innovative players to compete in procurement and reduce vendor lock-in. Specific attention is raised for cloud hosting by private operators, because this entails the storage of government data at privately run datacentres. In this case, it is a crucial consideration to limit the dependency on one supplier through consequent use of open standards.

The use of open standards and open source software

Open standards

The use of open standards is advocated as an instrument to mitigate the risk of vendor lock-in. As opposed to supplier-bound standards, open standards are not software-specific and therefore not

---

175 Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). Jaarrapportage Bedrijfsvoering Rijk. Available at: https://www.rijksoverheid.nl/documenten/rapporten/2019/05/01/jaarrapportage-bedrijfsvoering-rijk-2018

176 Therp. Overheid leert van Oracle incident. Available at: https://therp.nl/blog/our-news-1/post/overheid-leert-van-oracle-incident-38

177 Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). Strategische I-agenda Rijksdienst 2019-2021. Available at: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2019/01/01/rapport-strategische-i-agenda-rijksdienst-2019-2021/rapport-strategische-i-agenda-rijksdienst-2019-2021.pdf

bound to licenses. Rather, open standards focus on the required functionalities of software solutions and on enabling data portability and interoperability between software with similar open standards. Open standards are not only used in open-source software, but also in proprietary software, and their application by suppliers can as such bridge the existing vendor lock-in.[178] The manifesto for healthcare providers employs open standards in the (semi-) public sector, as does the Forum Standaardisatie, an executive agency for the national government, in a guidance document for IT procurement.[179]

The use of open standards is finding its way quite successfully even in proprietary business software, but open standards are unlikely to live up to the expectation of removing portability and interoperability issues indefinitely. Software providers argue that it is utopian to believe that good enough standards would resolve all problems, in particular because of the strong demand for customised solutions. Although nowadays it would be technically feasible to design APIs and implement protocols that could facilitate a uniform software standard, neither developers nor users would be served by a "one-size-fits-all" approach – simply because customisation and uniqueness cannot go together with and standardisations cannot go together.

Open source

Open-source software, interfaces and ecosystems are said to be vital for interoperability of internal systems, while closed systems hinder integration of data from various data-driven systems.[180]

Open-source software is a collaboration and organisation model for the production of software resources in all layers of the technology stack.[181] Due to the absence of costs for licensing, open-source software seems beneficial in terms of potential cost savings. Costless disposition of the source code stimulates the possibility that multiple suppliers develop and maintain the software, solving the dependency on a single supplier.

Open source is more cost efficient for users and is said to solve dependency on a small group of suppliers to fix legacy and license problems. Nonetheless, in practice the mitigating effect of open-source software on vendor lock-in is disputed.[182] [183] In essence, the difference between open-source and proprietary code lies in their licensing model, and the quality of software and ability to port data are in principle not impacted by this difference. In effect, open-source only offers a comprehensive software solution when combined with proprietary software, and as such the increased use of open-source software is unlikely to significantly reduce the need for licensed (proprietary) software. Eventually, open-source solutions have to evolve into or supplemented with commercially packages to become a viable option for business users, because open-source is not providing users the necessary support. Consequently, users will need a software provider's expertise for tested and sophisticated services, and the implementation and maintenance thereof.

178 PIANOo. (2016). Open standaarden ICT en bestekteksten. Available at: https://www.pianoo.nl/nl/markten/ict/inkopen-van-ict/open-standaarden-ict-bestekteksten

179 Rombouts, R. & B. Knubben. (2016). Vragen om open standaarden bij inkoop. Forum Standaardisatie. Available at: https://www.pianoo.nl/sites/default/files/documents/documents/openstandandaardenbijinkoopict-productenen-diensten-december2016.pdf

180 Computerworld. (2019). Wie is er nog bang voor vendor lock-in? Available at: https://computerworld.nl/open-source/110385-wie-is-er-nog-bang-voor-vendor-lock-in

181 Definition retrieved from report of a study commissioned by national government and executed by Verdonck, Klooster & Associates, Berenschot and Rand Corporation (2016), Mogelijkheden om de afhankelijkheid van ICT-leveranciers te verminderen. Available at: https://www.rijksoverheid.nl/documenten/rapporten/2017/10/11/onderzoek-open-source-software

182 Idem.

183 De Pous, V. (2020). Rijksoverheid ontbeert digitaal sourcingsbeleid. IT Executive. Available at: https://itexecutive.nl/wp-content/uploads/2020/01/Rijksoverheid-ontbeert-digitaal-sourcingsbeleid.pdf

EU-level initiatives on standardisation

Initiatives to tackle vendor lock-ins have also been undertaken at the EU-level. For example, DG CNECT (European Commission) presented best practices for ICT procurement based on standards in order to promote efficiency and reduce lock-in to organisations suffering from a lock-in.[184]

The most promising initiative at EU level is conducted by the SWIPO (switching and porting) Codes of Conduct Working Group.[185] This working group was established as one of two Digital Single Market (DSM) cloud stakeholder working groups in 2017. Their objective was to conduct self-regulatory work in the areas of cloud security and porting data/switching cloud service providers, ultimately leading to self-regulatory Codes of Conduct on data portability across different cloud infrastructures and across different cloud-based applications that reduce the risk of vendor lock-in by cloud service providers.

SWIPO's Codes could boost the European market for cloud services, by setting pre-defined regulation and quality requirements for software providers to compete on the market. In more traditional types of software, regulation could only be introduced after the fact, potentially leading to complex issues of non-compliance of fully integrated business software systems. The market for cloud services, on the other hand, is still establishing and could implement such requirements more easily. This could make the European market for cloud services more fluid and enables smaller companies and new market entrants to compete.

The group – co-chaired by representatives from the cloud service industry and from business users of cloud services, with a balanced mix of legal, technical and economical expertise and professional experience in the field of cloud computing – published its outcomes in November 2019.[186] These included Codes of Conduct for both cloud infrastructures (IaaS) and cloud-based applications (SaaS), and were presented as a milestone for the European data economy, reducing vendor lock-in and facilitating competition between cloud service providers. For instance, the Codes define best practices and information requirements for sufficiently detailed, clear and transparent information before a cloud service agreement enters into force.

However, the Codes' added value were subject of debate for involved stakeholders. Important stakeholders like Cigref (representing 150 large French companies and public administrations) responded that they could not recognise the legitimacy of the documents, as the documents failed to represent equally the interests of all stakeholders involved. The failure is "the result of a systemic asymmetry of skills, resources and objectives between those of some of the world's leading cloud service providers, on the one hand, who defend the core of their business and their ability to lock their customers, and on the other hand those users whose lobbying in this area is not the business".[187] This argumentation was supported by statements from stakeholders involved in this study.

Developer representatives had significantly more interest in a satisfactory outcome than their counterparts in the working group, as it touched upon the core of their activities and business model. Developers had a unified objective and maximized lobby power in terms of money and time, effectively capturing the policy process. Although customer representatives from ministries and

---

184 PwC. (2016). Study on best practices for ICT procurement based on standards in order to promote efficiency and reduce lock-in. European Commission: DG CNECT. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=14434

185 European Commission (2020). DSM cloud stakeholder working groups on cloud switching and cloud security certification. Available at: https://ec.europa.eu/digital-single-market/en/dsm-cloud-stakeholder-working-groups-cloud-switching-and-cloud-security-certification

186 SWIPO. (2019). Multi-stakeholder group presents Codes of Conduct to enable competition and data portability for cloud service customers across Europe. Available at: https://swipo.eu/media/SWIPO_press_release.pdf

187 Cigref. (2019). SWIPO: Failure to regulate the European cloud market. Available at: https://www.cigref.fr/swipo-failure-regulate-european-cloud-market

large corporations contributed, their interests are primarily concerned with their core business and were as such less direct, as was their corresponding capacity in terms of time and money. Finally, customers' individual interests were rather divers.

The Commission's SWIPO initiative could potentially facilitate data portability by laying down its technical ground rules. The self-regulatory Codes of Conduct for data portability across different cloud infrastructures and across different cloud-based applications are defined with consent of software suppliers. The Codes potentially make the European market for cloud services more fluid and enables smaller companies and new market entrants to compete there as well. However, so far, the added value has been disputed by involved stakeholders, as the Codes were highly influenced by the perspective of the world's leading cloud service providers, while input of other user stakeholders was limited.

6.3.6 What is the impact of mandatory data sharing on innovation incentives?

Given its technical and legal feasibility, the impact of mandatory data sharing on innovation incentives is twofold.

Potential benefits for innovation

On the one hand, obligatory data portability between software applications lowers technical barriers to smaller providers for offering their software applications on the market, as it creates the fundament for compatibility of their applications to the systems of larger providers. As such, it facilitates lower entry barriers for market entrants and smaller providers and contributes to a more diversified market supply.

Still, experts argue that given the size and complexity of software systems for large organisations, purchasers are reluctant to select small software developers for their projects. Notwithstanding small competitors' ability to implement large-scale IT projects, the diversification of supply in combination with the possibility to integrate solutions from various providers in one system, offers purchasers more possibilities to shop between providers. The increased ease and reduced transition costs of switching between suppliers generate some leverage for purchasers with current and competing suppliers, both during the duration of the contract and in their negotiations for a new contract or contract renewal.

As a result, dominant providers will have somewhat less ground to apply their usual their pricing strategies in negotiations with software purchasers – possibly leading to some cost efficiencies for purchasers. Moreover, they suppliers may experience more discipline to provide quality, tailored to the needs of their customers, rather than pushing for upselling of additional software solutions. Note that we envisage such effects to occur only to a certain extent, due to other (persistent) causes of customer lock-in, as discussed above.

Nonetheless, the tendency for big software firms to acquire software companies before they can develop into a serious market competitors will not be prevented by data sharing. For venture capitalists who invest in innovative and small software start-ups, being taken over by a leading software firms can be an objective.

Increased integration of currently autonomous applications and datasets offer several novel data connections that could lead to innovations for end-users through re-use and integration. Also, increased integration of systems reduces internal duration of processes and a reduction of the margin of error. Clients that have had to pay for data sharing through data portability—or who decided not to invest in further development of their applications or functionalities due to considerable scale-up costs—will receive benefits of data portability.

Such innovations can lead to economic benefits, through more efficient internal processes, better informed decision-making, and more synergies, as well as in more secure infrastructure for data exchange between applications, and evidence-based policymaking.

Potential risks for innovation

On the one hand, there is a risk of underinvestment in data portability under a mandatory data sharing framework. Experiences from the SWIPO Working Groups learned that leading software suppliers apply great lobby power, by investing significantly time and money to alter relevant policy debates in their preferred direction. They do so because implementing large-scale frameworks that enable data portability between software providers harms their business models. If policymakers would somehow succeed in implementing a comprehensive framework for data portability, leading software developers that have relied on profits received for proprietary software development and creating data portability for legacy systems, will be forced to open up their services and find a new revenue model. These activities constitute an important element in the work of software suppliers, and enforcing (mandatory) data portability between providers would cut off part of their revenue stream.

Some experts anticipate that suppliers will protect their business models at great cost, and in case data sharing policies come into existence, they might try to find ways to make them less effective. At present it is hard to foresee to what extent that may happen though.

More importantly, the development and implementation of the required standards would complicate innovation in terms of meeting demand for customised solutions. Mandated data portability will create cost efficiencies for business users, but they might come at the expense of increased complexity and a more limited scope for innovation for software developers. The services provided by software developers are unique, and business users often make deliberate considerations based on preference for the unique features of one software provider over another. The implementation of standardisation in the sector would harm the tailored and differentiated supply in the sector, especially when implemented as a 'one-size-fits-all' across differentiated business software solutions.

Some providers fear that they continuously enhance their products but get confronted with the need to adapt their products when industry standards support a particular feature which their product cannot meet as easily as a rival's offering.

6.3.7 What are the risks of data sharing?

In policy initiatives to facilitate data portability, such as the Open Data Initiative or SWIPO, software developers have been actively present and supportive. Of course, their position consists of finding a balance between coordination versus room for innovation and diversification.

Experts argue that from a perspective on technical feasibility, overcoming data portability and interoperability issues in software systems is already possible. A risk of mandated portability approaches could then be that the implementation of government mandated 'one-size-fits-all' or 'lowest common denominator approaches' across numerous differentiated business software solutions might create some cost efficiencies for business users, at the expense of increased complexity and limited innovation.

6.3.8 What are the options under current legislation (general and sector-specific) to make data sharing mandatory?

As opposed to the other case studies, the case of data sharing in administrative software does not specifically concern the use and transfer of personal data. Consequently, the right to data portability as laid down in GDPR does not apply here.

Instead, SWIPO's self-regulatory codes of conduct for portability between cloud providers as introduced through the Regulation on non-personal data may be relevant. Article 6(1) requires the European Commission to encourage and facilitate the development of self-regulatory codes of conduct at EU level, which include 'best practices for facilitating the switching of service providers and the porting of data in a structured, commonly used and machine-readable format including open standard formats. This type of intervention is an example of facilitating data sharing through non-legally binding measures. So far, the self-regulatory codes of conduct have been insufficiently effective in creating and maintaining a competitive data economy, which the Regulation aimed to achieve.

Finally, data portability might be imposed as a competition law remedy under the essential facilities doctrine but the same limitations as discussed in the other case studies apply, including the dominance of the firm involved and the indispensability of the data to be ported. Wider requirements of portability and interoperability are better mandated outside of the competition framework, for instance by building upon the self-regulatory codes of conduct that are now facilitated by the Commission.

6.3.9 Policy options

As discussed in section 6.3.8, options under current legislations, such as competition law, are limited and have a self-regulatory and non-mandatory nature.

Fixes at the national level

Progress made by the SLM teams in negotiations with big software suppliers on service agreements could incentivize policymakers to upgrade the SLM's mandate and offer leniency in the application of procurement rules for organisations that are active in services of general economic interest. Strengthening their bargaining power could be a promising possibility to create a more equal market structure.

Currently, the SLM teams are sometimes in a position where their powers technically do not meet the level of discretion needed to negotiate the terms of a deal – for instance, in terms of aggregation of demand. In addition, they could benefit from a more concentrated software procurement – for instance, by procuring as a representative of all Dutch municipalities or hospitals. For example, because the health care sector is a patchwork of health care providers, ranging from completely private to public, the sector is not allowed to concentrate their procurement via the Dutch Hospitals Association (NVZ) under the existing competition and procurement regulation. As a result, their IT costs are an estimated 50 to 70 per cent higher than those of ministries.

To expand the width of their mandate, the responsible minister could grant the SLM teams exemptions to existing procurement rules, like exemptions made for centralized procurement of crucial medicines. Under current regulation, some flexibility in aggregation of demand is possible when there is a disbalanced playing field between suppliers and purchasers in a market; however, there is reluctance, for instance among municipalities, in applying this flexibility as it could lead to legal procedures with software providers.[188]

The need for supranational policy options

Moreover, stakeholders emphasized that for data sharing policy options to be effective, they had to be discussed, created and implemented at a European or even global, rather than a national level. In that regard, ministries and representative associations like the CIO Platform maintain in close

---

[188] Computable. (2019). VNG zet aanbesteding GT Software stop. Available at: https://www.computable.nl/artikel/nieuws/wie-gunt-wat/6844872/3152533/vng-zet-aanbesteding-gt-software-stop.html

contact with their European counterparts – predominantly with the Commission, EDPS and their Belgian, French, German and other (North-)Western European partners. Although SWIPO has potential at the EU level, it is not yet in the stage of aligning political and technical realities.

One expert indicated that an important step could be the creation or appointment of a central authority or gatekeeper that supervises market participation in the market for cloud services. For existing software products and services in more mature markets, it might be complex to set and enforce quality standards, but this authority could design, implement and enforce central standards for crucial elements of cloud services, such as information security, data portability and interoperability. A relevant development in this regard is the establishment at EU level of the European Union Agency for Cybersecurity (ENISA), which pursues the implementation of central standards for information security.

From a software providers perspective, if self-regulatory sectoral agreements, such as SWIPO, would be supervised by a third-party authority, it remains to be seen whether it would benefit or harm their implementation case. They would benefit most from a global taxonomy, for instance through adopting ISO standards in this field at EU-level, instead of defining own standards. Own EU-level standards could become problematic in the framework of conflicting developments in other parts of the world (e.g. certain discussions currently going on in SWIPO have already been resolved in the context of ISO).

Nonetheless, although implementation and supervision of collective standards could be beneficial for business users, it should be noted that it follows from this case that it might be a necessary condition to remove issues regarding business users' perceived lock-ins, but is probably not a sufficient condition. Lock-ins are the result of high switching costs in software systems, which could be reduced in terms of uncertainty of data access and risk of discontinuity of services by diminishing technical and legal barriers, but would still be significant of investment required.

7 Case: vertical exclusion

7.1 Case description

This case is somewhat exceptional, since it does not zero in on a specific market or sector. Instead, it explores the blocking of vertical data sharing between vertically integrated platforms and third-party services. Firms and platforms may start out with a complementary vertical relationship fostered through access to data held by the platform. By adding new features, a service provider using a platform's data may move closer to the platform's business model, which introduces a horizontal relationship, or at least a horizontal, competitive threat. In such cases, the platform may decide to block data access. In the past, we have seen examples of services being suddenly blocked from data access. For example, Facebook blocked Vine from data access, and recently, Strava blocked Relive's data access.

This case elaborates on the example of Strava and Relive. Based on publicly available information, this example allows us to clearly illustrate the answers to several questions below. Note that some of the questions explored in the previous cases are left out of the case analysis, as they do not apply.

Strava allows athletes to track and analyse running and cycling activities, using data collected via phones, watches, and third-party tracking devices. It developed into a social network allowing athletes to connect and interact with each other. Users benefit from third-party apps that complement Strava's functionalities. These third-party apps rely on user data uploaded to Strava and shared with the third party. An example is Relive, which generates workout movies and visualises tracking data in a 3D map. At certain point in time, Relive was denied access to Strava's API for violating certain terms of agreement. This development illustrates that in general, platforms are unlikely to restrain data access for services that complement their own services, but this may be different for competing services[189].

7.2 Analysis of research questions

7.2.1 Which role does data play in the business model and what are the benefits that data provides?

In general, platforms typically facilitate vertical data sharing with downstream parties because it increases the value of their ecosystems. Third parties benefit from data-sharing agreements because it enhances their service or accelerates their growth. Thus, one may expect that vertical partnerships emerge automatically. However, as downstream apps move closer to a platform's business core functionalities, they may transform in a horizontal competitive threat and (depending on the terms of the agreement) could be cut off from data access.

It is instructive to illustrate data usage in business models. Strava started out as a tracking app which allows athletes to track and analyse running or cycling activities. Tracking data is collected via smart devices (phones or watches) or via third-party tracking devices (such as Garmin or Wahoo). Strava soon developed into a social network app allowing athletes to connect and interact with other athletes about sports activities and engage in competitions, training and challenges with them. Strava is now also open for sports clubs, brands and (sports)retailers to engage in marketing activities. Its users can also make use of many third-party apps that complement the functionalities

[189] "Strava op ramkoers met Nederlandse sportapp Relive", *Het Financieele Dagblad* 1 August 2019.

offered by Strava, including analytics, interaction, training, and visualisers. These third-party apps rely on user data uploaded to Strava and shared through Strava's API.

Strava collects the following types of data[190]:
- basic account information (name address, age, gender, contact details, etc.);
- profile, activities and user information (provided or observed, including videos, foto's, kudo's, smileys etc.);
- location data, including speed, altitude, and routes (via Strava app, third party app, third party device);
- health data (via Strava app, third party app, third party device);
- shared content (provided by users);
- social graph data (imported from devices or other accounts such as Facebook or Gmail)[191];
- technical data and logfiles.

Relive used to be such a third-party app. Its core functionality is to generate short movies of workouts. It visualised Strava's tracking data in a 3D map and added photos and videos that the athlete(s) shot during the exercise[192]. At the time, Relive retrieved the following data from Strava:
- location data (for describing the routes that athletes have taken);
- social graph (for the social feed and networking functions).

Strava was just one of Relive's providers of tracking data. Its users can also import data from several other platforms with tracking functionalities[193], as well directly from a smartphone's GPS. In particular, Relive (itself as well as through various sources) gathered the following data types:
- basic account information (name address, age, gender, contact details, etc.);
- location data (via devices with GPS functionality)
- photos and videos uploaded directly (and timestamped by the device)[194];
- map data and imagery providers[195] .

Thus, one can observe that Strava is a social network that facilitates member interactions. The main growth driver of the platform consists of direct network effects. Next, Strava provides a platform for app developers, driving indirect network effects. Data about user profiles is essential as well as data on users' social graph. Strava as well as other parties develop a range of functionalities. These third-party developers need access to data collected or aggregated by Strava. Relive used to make use of tracking data aggregated by Strava as well as by Relive itself.

Business models tend to adapt and change. At some point in time, Relive developed and introduced social interaction functionalities for which it needed access to Strava's social graph data; or more specifically, it needed access to an individual's social graph on Strava and aimed to target users' friends through Strava. For Relive this was a means to leverage direct network effects from Strava into its own app. Below, we discuss the subsequent developments.

7.2.2 Which data and types of data are indispensable?

In our illustration, we focus on data which is indispensable for Relive. Initially, its business model only required observed/provided location data and user-generated content (specifically photos and

190 https://www.strava.com/legal/privacy#aggregate_information

191 https://support.strava.com/hc/nl/articles/360016371512

192 It is possible to merge photo and video material of persons being within 50 meters distance of each other during the majority of the exercise. https://www.relive.cc/support.

193 Garmin Connect, Wahoo, Polar Flow, Suunto, Apple Health, Endomondo, adidas Running by Runtastic, Under Armour (inclusief: MapMyRun, MapMyWalk, MapMyRide, MapMyHike, etc.) of Ride met GPS

194 Timestamps are used to place photos and videos on the route

195 https://www.relive.cc/map-attribution

videos) to offer its core functionality. After the evolution towards a social networking feature, additional data (in particular social data) became a necessary, additional ingredient.

7.2.3 *Can access to indispensable data facilitate entry or strengthen competition?*

Although this case is about vertical relationships, it is useful to also consider horizontal issues. Strava and Relive started out with a complementary vertical relationship. By adding new features, a Relive moved closer to the platform's business model. This development gave rise to a horizontal competitive threat for Strava. By blocking data access, Strava may have been trying to defend its market position against this horizontal threat.

As we will point out in more detail below, Strava's API agreement excludes partners offering apps similar to or competing with Strava. Forbidding this type of clause would support horizontal competition and stimulate platforms as well as downstream app developers to innovate more. We note that it is not uncommon for platforms to include such terms in API agreements. In the past, Facebook also did not provide data access to apps that replicated its core functionality. However, the company dropped this policy in 2018.

Regarding horizontal data sharing, as explained in the cases about online search and social networks, a lack of access to an incumbent's data may hamper horizontal competition between same-service-providers. Data sharing may mitigate first-mover advantages related to learning effects, and data sharing may stimulate multi-homing which mitigates the competitive advantage stemming from network effects. The question remains, however, whether data access is essential for newcomers in the market, or if they can find different routes to access data or build an innovative business model to effectively compete with incumbents with a data advantage.

Coming back to vertical data sharing, in traditional markets it is not logical for the owner of an essential facility to block a downstream service that does not compete with its own downstream services. A refusal to deal must then be motivated by efficiency or quality reasons. In digital markets, however, the threat for a platform of being intermediated or disintermediated by another player is much larger, and even comes from players currently offering complementary services in the downstream market. A platform may want to defend itself against competitive threats like these (or at least not want to facilitate them) by terminating or limiting the partnership with the downstream player or by limiting its growth otherwise. Such defensive actions come at the expense of the richness of the platform's ecosystem and may drive users away to rival platforms unless the platform is dominant[196]. Thus, there can be an anti-competitive motivation behind a refusal to provide access to data. However, note that a refusal can be legitimate, for example when data sharing damages the privacy of data subjects (and would undermine confidence in the platform), or when it would lead to free riding.

Based on this type of reasoning, Strava's termination of the agreement with Relive could perhaps qualify as vertical exclusion under competition law. A pre-condition (to be proven by the competition authority) would then be that Strava is dominant. A possible reason could be that it controls access to an "essential facility" (i.e. it has exclusive access to unique data that is indispensable to become active in the market) or because of other unsurmountable entry barriers. As we will discuss below, it tends to be quite hard for a competition authority to prove that an incumbent's data is an essential facility. However, we are not aware of any current or past cases in which (access to) data has been labelled as essential. The following sub-section argues that in the example of Strava, there does not appear to be an essential facility. In general, though, competition authorities can still assess the role of data in competition problems.

196 It may also be the case that the excluded service was not appealing enough to end-users, but in that case there would not be any reason for the platform to feel threatened by it, and thus, it would not have a reason to exclude the service.

7.2.4 Are there bottlenecks for accessing indispensable data?

There might be bottlenecks for accessing indispensable data, in particular if the data in question is unique and there is only one way of obtaining it; for instance, when the data cannot be obtained without legal privilege, or because data collection requires large (sunk) investments. Examples of the former case include health data and smart meter data (see textbox below). We are not aware of examples of the latter case. Thus, only under exceptional circumstances can (access to) data be regarded as an essential facility. What is more likely to happen is that a combination of data, algorithms, consumer switching costs, network effects, and so on, provides a competitive advantage and possibly market power (which is not the same as saying that such a combination gives rise to an unsurmountable entry barrier).

Legal bottlenecks for accessing indispensable data

An example of essential data is provided by smart meters for energy consumption[197]. In some countries (e.g. the Netherlands), distribution system operators have a legal monopoly to connect households to the energy grid and install smart metering equipment in households' premises. Thus, they have direct access to a range of a household's energy data. EU regulation, however, provides in limited access to energy consumption data. This gives outsiders the possibility to offer households a smart thermostat that receives data from the smart meter. Based on that, such a company can provide energy savings services.

Another sector with strict requirements about data is health care. These restrictions also provide data holders with privileged or restricted access to data. Health care organizations and vendors supplying software to them can, thanks to the involvement with patients, collect, generate and control unique data. They have to power to decide if and under which conditions they share data with others[198]. Complementary legislations, such as provided by the GDPR, may empower consumers to get access to such data, or to give their consent so that other parties can use their data.

Let us come back to the example of Strava. Shortly after Relive had introduced a social networking functionality, Strava denied Relive access to its API. Here it is relevant to observe that Strava's API agreement says that partners *"... may not create an application that merely replicates or competes with Strava"*[199]*.* It also states that *"... Strava may currently or in the future develop products and services that may be similar to or compete with your Developer Applications"*[200]. Hence, Strava may initially provide a certain developer with access to its data, then copy the partner's functionality, after which it reserves the right to block access to the data[201].

Regarding the question we are addressing here, it is important to note that Strava's blocking of API access did not impose a fatal injury on Relive. Relive relied partly on Strava for access to location data, but Relive's customers can also connect the app directly with a wide range of tracking devices, such as those manufactured by Suunto, Polar, Apple, Garmin and Adidas[202]. The blocking of data access may also not be fatal for Relive's social networking functionalities, as Relive can probably rely on other sources, such as Facebook, for accessing users' social data. More generally, social graph data can be accessed through sources like email directories, phone address books, and social logins offered by various platforms (besides Facebook also including Google and

197 See e.g. ACM (2019), discussing the underlying Clean Energy Package that contains proposals to eliminate bottlenecks in access to energy data. The draft directive requires, among other things, transparent and non-discriminatory access to data.

198 Savage et al. (2019).

199 https://www.strava.com/legal/api, second bullet under introductory paragraph (accessed 10 June 2020).

200 https://www.strava.com/legal/api, §11.B (accessed 10 June 2020).

201 This happened to "Segment Challenge"; see https://webbindustries.com/hackernews/story/20421625.

202 See https://www.relive.cc/ (accessed 10 June 2020).

LinkedIn). The latter type of options allows for communicating with friends (i.e. share streams and updates).

Nevertheless, another app developer that was cut off from data access by Strava, Segment Challenge, was fully dependent on Strava's API[203] and was shut down soon after Strava cut off its data supply.[204] Having said that, even with such a dramatic outcome, the data in question need not be essential or non-replicable, as illustrated by Relive's strategy to rely on a wide range of data sources.

More generally, it may be hard to prove that (access to) an incumbent's data is essential. Competition authorities can, however, include data in assessments of competition problems. For example, suppose that Strava blocking Relive (a relatively popular app used by about 5%-10% of Strava users, as can be observed from company websites) did not drive users away from Strava. This would serve as an indication of a dominant position for Strava. The next question is then if this behaviour is anti-competitive in the sense that it allows the dominant platform to avoid competition on the merits. Strava's API agreement, which restricts data access to developers offering similar or competing services, suggests that this is the objective of the behaviour.

7.2.5 What is the impact of mandatory data sharing on innovation incentives?

Preventing exclusion by imposing a general obligation on platforms to vertically share data may be excessive. First, platforms typically have an incentive to share data with vertical players to enrich the services offered on the platform. Second, a general obligation to vertically share data may reduce the incentives of downstream players to innovate and find new ways around the upstream platform or diversify their data sources.

An obligation to share may be desirable if a platform does not share (exclusive, unique) data in order to keep out (potential) competitors (cf. the data sharing obligations for banks, electricity companies and health providers). In that case, the primary objective of the sharing obligation is to stimulate horizontal competition that would otherwise not arise. In this type of situation, one must however be aware of negative impacts on competition that may occur in response to such obligations. In particular, note that data sharing requires certain standards and data formats. This pushes the type of innovation that obligations engender in the direction of ideas based on the data types in question and on the forms in which they can be accessed. By implication, competitors experience less drive to aim at more radical innovations that could replace the incumbent, or that are based on other types of (elsewhere available) data. The result could be less diversity in functionalities, due to the limits of the selected data protocols and standards, and also to interference with business models that go beyond innovations that require data access[205]. The relevance of this type of argument will apply to the market in question, including, for instance, the rate of technological change and innovation in business models.

When platforms already vertically share data with complementary downstream players (like Strava), there may be reasons to impose non-discrimination obligations. The primary objective of such non-discrimination obligations is to prevent anti-competitive behaviour towards downstream services that threaten to (dis)intermediate the platform or threaten to evolve into a horizontal competitor. In the case of Strava and Relive, for example, a non-discrimination obligation would have prevented Strava from blocking Relive's data access.

---

203 https://news.ycombinator.com/item?id=20421625

204 https://twitter.com/segmntchallenge

205 For this argument applied to instant messaging, see De Bijl and Van Gorp (2020).

A relevant question is whether data sharing obligations or non-discrimination obligations reduce the incentives to invest in data. The answer depends on the type of platform and the nature of the data in question.

For observed and provided social, contextual, or profile data, it is unlikely that (non-discriminatory) data sharing obligations undermine the incentives to collect such data because they are typically gathered while providing the service (financial information is gathered while offering banking services, data on electricity consumption is gathered while supplying electricity, health data is gathered while treating people, and social data is observed while providing social media services)[206]. These data are often shared voluntarily with companies offering complementary downstream services, but even when downstream partners evolve into horizontal competitors, sharing the data most likely does not affect the incentives to collect them. In our illustration, an obligation to share location and social data with a potential competitor like Relive would not harm Strava's incentives to collect these data. These data are a by-product of Strava's core functionality.

In cases involving derived or inferred data, the impact on the incentives to produce the data may be different, but it depends on how the data is employed by both parties. For example, inferred or derived social and profile data may contain deeper insights into individual preferences. This makes them highly valuable for targeting consumers, which is particularly relevant for platforms that monetise consumer attention through selling advertising space. Sharing derived data undermines its exclusivity and may invite free riding by others. The risk of freeriding is less with vertical partners whose value materialises within a platform's ecosystem, so that the platform can capture some of that value through its pricing. The risk of freeriding is more pertinent when inferred or derived data must be shared with horizontal competitors (including vertical partners that evolve into horizontal competitors). That may reduce the potential for revenue generation, so that mandatory sharing could harm the willingness to invest. It follows that discriminatory access to inferred and derived data may be based on an objective reason that it harms the upstream platform's incentives to invest in the production of the data.

In the example of Strava and Relive, an obligation for Strava to share observed and provided location and social data, would not have affected its incentives to collect these data. If the data concerned derived or inferred data, there might have been objective reasons for Strava to restrict data access to partners offering apps similar to or competing with Strava.

7.2.6 *What are the options under current legislation (general and sector-specific) to make data sharing mandatory?*

At present, there are little options to make data sharing mandatory under competition law. The burden of proof to establish dominance and/or apply the essential facilities doctrine is very high while it takes a long time to build a case and reach a verdict. By that time, the damage may already be done and may be irreversible (see also Van Gorp and De Bijl, 2019).

It helps competition authorities when a vertically integrated platform applies for approval of a merger or acquisition. In such cases, competition authorities have an escape, because remedies can be imposed as conditions for approval. For instance, a remedy may consist of an obligation to the provide non-discriminatory access to data. ACM has already pioneered data sharing as a remedy in the Phase II review of such a case. In the Dutch market for educational materials publishing, the competition authority conditionally cleared the acquisition of Iddink Group by

---

[206] For behavioural data, the answer depends on the nature of the business model and the risk of reverse engineering (see for instance the case on online search).

Sanoma Learning[207]. Iddink Group owns the learning management app Magister, which is used by secondary schools. Sanoma Learning owns publisher of educational materials Malmberg. The remedy included conditions to facilitate access for Malmberg's competitors to Magister, under equal conditions and in the same way as Malmberg can access it. Furthermore, Sanoma had to ensure that commercially sensitive information of competing publishers cannot find its way to the company through Iddink.

7.2.7 *Policy options*

Van Gorp and De Bijl (2019) analyse vertical discrimination by digital platforms. In general, discrimination by vertically integrated digital platforms should be frowned upon, but there may be objective reasons for such behaviour. They discuss the main options for prohibiting discriminatory behaviour, based either on a new ex ante rules or on ex post interventions within competition law, the latter one complemented with specific guidelines on inferring dominance from observed behaviour. We briefly summarise their findings.

In case of ex ante options, discrimination is not allowed unless a platform can convince the competition authority that it as objective reasons for it. An objective reason could be that discrimination of data access is necessary for maximising downstream innovation while preventing freeriding and reverse engineering. One can apply such rules to all vertically integrated platforms, or only to platforms with a (dominant) gatekeeper position. The first option seems more effective in mitigating the risk of irreversible harm as it would considerably speed up the decision-making by competition authorities. In both cases, however, there may be harm when objective reasons to discriminate turn out to be valid. For example, when a platform must share its behavioural data until it has successfully proven the risk of reverse engineering, the damage may already be done.

Under ex post intervention, discrimination is allowed unless someone files a complaint and the platform fails to provide objective reasons for its behaviour. Since only dominant firms can harm competition through discrimination, this approach only relates to dominant platforms. While establishing that a firm is dominant can be complicated, here dominance can be inferred from the occurrence of discrimination by a vertically integrated platform that does not have objective reasons for it. Only when an efficiency defence is valid, it becomes necessary to define the relevant market and assess dominance, namely to weigh the efficiency gains against possible harm to competition. This ex post approach speeds up competition cases, mitigates the risk of irreversible harm, and reinstalls the deterrent effect of competition law on exclusionary abuse. Contrary to ex ante rules, the ex post approach does not slow down or harm innovations, as platforms can initiate efficiency enhancing discriminatory practices without having to wait for permission.

[207] "ACM conditionally clears acquisition of Iddink Group by Sanoma Learning", ACM press release 29 August 2019. https://www.acm.nl/en/publications/acm-conditionally-clears-acquisition-iddink-group-sanoma-learning (retrieved 28 April 2020).

8 References

ACM (2019), "Visie datagovernance energie", ACM/INT/384185, 21 March.

Adar, E. (2007). User 4XXXXX9: Anonymizing Query Logs. *presented at the World Wide Web Conference.* Retrieved from http://www.cond.org/anonlogs.pdf

Agrawal, A., Gans, J., & Goldfarb, A. (2019). Economic policy for artificial intelligence. *Innovation Policy and the Economy, 19*(1), 139-159.

Anderson, J. (2020, February 11). *The dynamics of data accumulation*. Retrieved from Bruegel Blog: https://bruegel.org/2020/02/the-dynamics-of-data-accumulation/

Argenton, C., & Prüfer, J. (2012). Search Engine Competition with Network Externalities. *Journal of Competition Law & Economics*, 73-105.

Bertrand, M., & Kamenica, E. (2018). *Coming Apart? Cultural Distances in the United States over Time.* Cambridge, MA: NBER Working Paper No. 24771.

CBS. (2016). *ICT, Kennis en economie 2016.* The Hague: Centraal Bureau voor de Statistiek.

CBS. (2019). *ICT, kennis en economie 2019.* The Hague: Centraal Bureau voor de Statistiek.

Choi, J. P., & Jeon, D. S. (2016). *A Leverage Theory of Tying in Two-Sided Markets.* Working Paper.

Choudary, S. P. (2015). *Platform Scale: How an emerging business model helps startups build large empires with minimum investment .* San Francisco: Platform Thinking Labs.

Cloud Security Agency. (2014, September). *Big Data Taxonomy.* Retrieved from cloudsecurityalliance.org: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf

Crémer, J., de Montjoye, Y., & H., S. (2019). *Competition policy for the digital era.* European Commission, a study commissioned by DG Competition.

Crémer, J., de Montjoye, Y., & Schweitzer, H. (2019). *Competition policy for the digital era.* European Commission, a study commissioned by DG Competition.

de Bas, P., Batura, O., Yagafarova, A., van Gorp, N., ..., & Pichler, E. (2017). *Business-to-Business relations in the online platform environment.* Rotterdam: Ecorys/WIFO commissioned by the European Commission DG Connect.

De Bijl, P., & Van Gorp, N. (2016-2019). Gastcolleges in Digital Business Models and Strategy. *Modular Executive MBA in Business & IT: Digital Strategy and Transformation.* Breukelen: Nyenrode Univserity.

De Bijl, P., & Van Gorp, N. (2018, December). Publieke belangen en de regulering van digitale platformen. *Economische Statistische Berichten (ESB), 103*(4768S), 63-38.

De Bijl, P., & Werner, G. (2019). Platformbedrijven als marktmeesters. *TPEdigitaal*(jaargang 13(2)), 33-49.

De Standaard. (2017, 05 11). *Zo kwam Snap aan 2 miljard dollar verlies*. Retrieved 12 31, 2019, from standaard.be: https://www.standaard.be/cnt/dmf20170511_02876608

Deahl, R. (2019, August 8). *How to Write a Compelling Profile of a Person*. Retrieved from thebalancecareers.com: https://www.thebalancecareers.com/writing-a-compelling-profile-of-a-person-2316038

Digital Transport & Logistics Forum. (2018). *Data sharing in supply and logistics as commodity.* Retrieved april 6, 2020, from ec.europa.eu: https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=14275

EC-Council. (2016). *Computer Forensics: Investigating File and Operating Systems, Wireless Networks, and Storage* (2 ed.). Cengage Learning.

Ecorys. (2008). *A study on EU spending.* Rotterdam: European Commission.

Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform Envelopment. *Strategic Management Journal , 32*(12), 1270-1285.

European Commission. (2015, November 25). Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 2007/64/EC. Brussels.

European Commission. (2016). *Study on best practices for ICT procurement based on standards in order to promote efficiency and reduce lock-in.* Brussels.

European Commission. (2020, februari 19). *A European strategy for data.* COM(2020) 66 final. Retrieved from ec.europa.eu: https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

Evans, P., & Gawer, A. (2016). The Rise of the Platform Enterprise: A Global Survey. *The Emerging Platform Economy Series, Vol. 1*. Retrieved from http://thecge.net/wp-content/uploads/2016/01/PDF-WEB-Platform-Survey_01_12.pdf

Ezrachi, A., & Stucke, M. (2017). Artificial Intelligence & Collusion: When Computers Inhibit Competition. *University of Illinois Law Review*, 1775-1810.

FCC. (2010). *2010 open internet order.* FCC 10-201.

Federal Trade Commission. (2016). *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues.* Washington: FTC Report. Retrieved from https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report

Frischmann, B. M. (2012). *Infrastructure: The social value of shared resources.* Oxford: Oxford University.

Furman, J., Coyle, D., Fletcher, A., McAuley, D., & Marsden, P. (2019). *Unlocking digital competition.* Report of the Digital Competition Expert Panel.

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm.* Basel: BIS Working Papers 834.

Gawer, A. (2016). *Competition Policy and Regulatory Reforms for Big Data: Propositions to Harness the Power of Big Data while Curbing Platforms' Abuse of Dominance.* Organisation for Economic Co-operation and Development.

He, D., Kannan, A., Liu, T. Y., McAfee, R. P., Qin, T., & Rao, J. M. (2017). Scale Effects in Web Search. *International Conference on Web and Internet Economics* (pp. 294-310). Springer, Cham.

Hogendorn, C., & Frischmann, B. (2020). Infrastructure and general purpose technologies: a technology flow framework. *European Journal of Law and Economics*. Retrieved from https://doi.org/10.1007/s10657-020-09642-w

IBM. (2020, March 12). *What is Derived Data*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/SSQVNT_3.8.0/ilog.odms.ide.odm.enterprise.help/ODME/Userman/topics/Creating/ADM_AdvDerivedData.html

ISO/IEC. (2017, September). *Information technology — Cloud computing — Cloud services and devices: Data flow, data categories and data use.* Retrieved from iso.org: https://standards.iso.org/ittf/PubliclyAvailableStandards/c066674_ISO_IEC_19944_2017.zip

Kathuria, V. (2019). Greed for data and exclusionary conduct in data-driven markets. *Computer law & security review, 35*(1), 89-102.

Kessel, J. M., & Mozur, P. (2016, augustus 9). *How China is changing the internet.* Retrieved from New Yort Times: https://www.nytimes.com/video/technology/100000004574648/china-internet-wechat.html

Lubarsky, B. (2017). Re-Identification of "Anonymized" Data. *Georgetown Law Technology Review*, 202-213.

Mano, M. d., & Padilla, J. (2018). Big Tech Banking. *Journal of Competition Law & Economics, 14*(4), 494-526.

Mayer-Schönberg, V., & Ramge, T. (2018). A Big Choice for Big Tech, Share Data or Suffer the Consequences. *Foreign Affairs, 97*(5), 48-54.

Meng, T., Jing, X., Yan, Z., & Pedrycz, W. (2020). A Survey on Machine Learning for Data Fusion. *Information Fusion, 2020*(57), 115-129.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). *Jaarrapportage Bedrijfsvoering Rijk.* The Hague: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2019). *Strategische I-agenda Rijksdienst 2019-2021.* The Hague: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Nerdwriter1. (2018, March 28). *How Dark Patterns Trick You Online*. Retrieved from YouTube.come: https://youtu.be/kxkrdLI6e6M

Nooren, P., Van Gorp, N., Van Eijk, N., & Fathaigh, R. Ó. (2018). Should we regulate digital platforms? A new framework for Evaluating policy options. *Policy & Internet, 10*(3), 264-301.

Nooren, P., Van Gorp, N., Van Eijk, N., & Fathaigh, R. Ó. (2018). Should we regulate digital platforms? A new framework for Evaluating policy options. *Policy & Internet, 10*(3), 264-301.

OECD. (2014). *Summary of the OECD privacy expert roundtable, Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking".* Working Party on Security and Privacy in the Digital Economy. DSTI/ICCP/REG(2014)3: OECD.

OECD. (2016). *Big data: Bringing competition policy to the digital era.* Background note by the Secretariat , Competition Committee Directorate for Financial and Enterpires Affairs. Paris: of the. Retrieved from https://one.oecd.org/document/DAF/COMP(2016)14/en/pdf

Pariser, E. (2011, May 2). *Beware online "filter bubbles".* Retrieved from YouTube: https://youtu.be/B8ofWFx525s

Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform revolution: how networked markets are transforming the economyand how to make them work for you.* New York City: WW Norton & Company.

Pastoor, R., & Peddemors, K. (2013, juni 28). *De ondergang van Hyves: Een reconstructie van het grote TMG-debacle*. Retrieved from The Post Online: https://biz.tpo.nl/2013/06/28/de-ondergang-van-hyves/

Petit, N. (2018). American tech giants are fiercely competitive monopolies. *ESB, 103*(4768S).

Porto, F. D., & Ghidini, G. (2020). I Access Your Data, You Access Mine. *International Review of Intellectual Property and Competition Law, 51*, 307-329.

Prüfer, J. (2019). *Competition Policy and Data Sharing on Data-driven Markets; Steps Towards Legal Implementation.* Bonn: A Friedrich-Ebert-Stiftung.

Prüfer, J., & Schottmüller, C. (2017). Competing with Big Data. *TILEC Discussion Paper No. 2017-006; CentER Discussion Paper 2017-007*.

PwC. (2016). *Study on best practices for ICT procurement based on standards in order to promote efficiency and reduce lock-in.* Brussels: European Commission: DG CNECT.

Radicand Economics. (2018). *A review of ACM's game theory analysis.* Woerden: VodafoneZiggo.

Retail 90. (2017, September 14). *Winning the Amazon Buy Box: What You Need to Know...* Retrieved from YouTube.com: https://www.youtube.com/watch?v=9jOSLkCsAik

Rocher, L., Hendrickx, J., & De Montjoye, Y. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications, 10*(3069). Retrieved from https://doi.org/10.1038/s41467-019-10933-3

Rogers, D. L. (2016). *The digital transformation playbook: Rethink your business for the digital age.* New York: Columbia University Press.

Rombouts, R., & Knubben, B. (2016). *Vragen om open standaarden bij inkoop.* The Hague: Forum Standaardisatie.

Ruiz, S., Gomes, P., Rodrigues, L., & Gama, J. (2019). Credit scoring for microfinance using behavioral data in emerging markets. *Intelligent Data Analysis, 23*(6), 1355-1378.

Rusli, E. M. (2013, April 1). Bad Credit? Start Tweeting. *The Wall Street Journal*.

Schweitzer, H., Haucap, J., Kerber, W., & Welker, R. (2018). *Modernising the law on abuse of market power.* German Federal Ministry for Economic Affairs and Energy. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3250742

Scott, M. (2020, march 3). *Coronavirus crisis shows Big Tech for what it is — a 21st century public utility.* Retrieved from Politico: https://www.politico.eu/pro/coronavirus-big-tech-utility-google-facebook/

Senellart, P. (2016, September 15). What do Internet companies know about you. *Data & Competition Policy.* Brussel: CERRE Private Seminar . Retrieved from https://pierre.senellart.com/talks/cerre-20160915.pdf

Stiglitz, J. E., & Rosengard, J. K. (2015). *Economics of the public sector* (Fourth international student edition ed.). New York: WW Norton & Company.

The Guardian. (2016, november 2). *Facebook forces Admiral to pull plan to price car insurance based on posts.* Retrieved april 6, 2020, from theguardian.com: https://www.theguardian.com/money/2016/nov/02/facebook-admiral-car-insurance-privacy-data

The Verge. (2016, november 2). *Facebook blocks insurer exploiting user data to find 'conscientious' drivers.* Retrieved april 6, 2020, from theverge.com: https://www.theverge.com/2016/11/2/13496316/facebook-blocks-car-insurer-from-using-user-data-to-set-insurance-rate

TNO, Ecorys, & IViR. (2015). *Digital platforms: an analytical framework for identifying and evaluating policy options.* The Hague: TNO, commissioned by teh Dutch Ministry of Economic Affairs.

TNO, Ecorys, & IViR. (2015). *Digital Platforms: an analytical framework for identifying and evaluating policy options.* Den Haag: TNO, in opdracht van het Ministerie van Economische Zaken.

U.K. Parliament. (2019). *Disinformation and 'fake news': Final Report.* London: the Digital, Culture, Media and Sport Committee of the British Parliament.
U.S. Department of Justice. (2010, February 18). *Statement of the DoJ Antitrust Division on Its Decision to Close Its Investigation of the Internet Search and Paid Search Advertising Agreement Between Microsoft Corporation and Yahoo! Inc.* Retrieved from United States Department of Justice: https://www.justice.gov/opa/pr/statement-department-justice-antitrust-division-its-decision-close-its-investigation-internet
UK CMA. (2019). Appendix E: The role of data. In *Online platforms and digital advertising market study* (pp. E1-E44). London: UK CMA. Retrieved from https://assets.publishing.service.gov.uk/media/5df9ecc040f0b609402e2838/Appendix_E_The_role_of_data.pdf
United Nations Statistical Commission and Economic Commission for Europe. (2000). Terminology on Statistical Metadata. *CONFERENCE OF EUROPEAN STATISTICIANS STATISTICAL STANDARDS AND STUDIES – No. 53.* Geneva: United Nations. Retrieved from https://ec.europa.eu/eurostat/ramon/coded_files/UNECE_TERMINOLOGY_STAT_METADATA_2000_EN.pdf
Van Gorp, N. (2018). Data en marktmacht: privacy is de oplossing, niet het probleem. *Tijdschrift voor Toezicht*(nr. 2-3), 88-89.
Van Gorp, N. (2018). Data en marktmacht: privacy is de oplossing, niet het probleem. *Tijdschrift voor Toezicht*(nr. 2-3), 88-89.
Van Gorp, N., & Batura, O. (2015). *Challenges for competition policy in the digitalised economy.* Rotterdam: Ecorys, i.o.v. het Europese Parlement.
Van Gorp, N., & De Bijl, P. (2019). *Digital Gatekeepers, assessing exclusionary conduct.* Zevenhuizen: e-Conomics, in opdracht van het Ministerie van Economische Zaken en Klimaat.
Van Gorp, N., Graef, I., Lemstra, W., Stokking, H., Nooren, P., Billeskov, K., . . . Valcke, P. (2016). *Study on future trends and business models in communication services.* Rotterdam: Ecorys, commissioned by the European Commission, DG Connect.
Van Gorp, N., Graef, I., Lemstra, W., Stokking, H., Nooren, P., Billeskov, K., . . . Valcke, P. (2016). *Study on future trends and business models in communication services.* Rotterdam: Ecorys, commissioned by the European Commission, DG Connect.
Van Gorp, N., Van Til, H., & Price, K. (2017). *Big Data and Competition.* Rotterdam: Ecorys, in opdracht van het Ministerie van Economische Zaken.
Van Til, H., Van Gorp, N., & Price, K. (2017). *Big Data and Competition.* Rotterdam: Ecorys, in opdracht van het Ministerie van Economische Zaken. Retrieved from https://zoek.officielebekendmakingen.nl/blg-813928.pdf
Verdonck, Klooster & Associates, Berenschot & RAND Corporation. (2016). *Mogelijkheden om de afhankelijkheid van ICT-leveranciers te verminderen.* Verdonck, Klooster & Associates.
World Economic Forum. (2014, Mei). *Rethinking Personal Data: A New Lens for Strengthening Trust.* Retrieved from weforum.org: http://www3.weforum.org/docs/WEF_RethinkingPersonalData_ANewLens_Report_2014.pdf
Wright, J. (2004). One-sided Logic in Two-sided Markets. *Review of Network Economics,* (3(1)), 44-64.

Postbus 4175
3006 AD Rotterdam
Nederland

Watermanweg 44
3067 GG Rotterdam
Nederland

T 010 453 88 00
F 010 453 07 68
E netherlands@ecorys.com
K.v.K. nr. 24316726

W www.ecorys.nl

Sound analysis, inspiring ideas