Closeness of substitution for "big data" in merger control

Norbert Maier[1]

Abstract

Recent merger cases involving the transfer of control over "big data" concluded that these mergers would not lead to competition problems because there was a sufficient number of other data sources available to the various players in the market. These merger cases implicitly accepted, but never carefully analysed, that these big datasets were close substitutes. As such big data is often not traded, one should look at the big data value chain and assess the closeness of substitution between such big datasets by evaluating to what extent the users of the insights derived from them view those insights as close substitutes. This new approach is illustrated through the examples of the Microsoft/LinkedIn and the Facebook/WhatsApp mergers.

1. INTRODUCTION

"Big data" is one of the central topics in competition policy discussions these days.[2] In particular, big data can raise issues both for antitrust and merger control as it is often a key asset for digital platforms.[3] For instance, merger control needs to assess to what extent big data (as an asset) owned by merging parties can be used to reduce competition in certain markets after a merger. One major difficulty in such competitive assessments is that big data is often not traded and, therefore, no market data is generated that could be used directly for competition analysis.

Theories of harm related to big data have been formulated In merger control for about a decade. These theories of harm were mostly of vertical foreclosure nature and the investigations focused on whether some part of the big data of the merging parties could qualify as an important or potentially essential input. In such cases the acquisition of big data could create competition problems in a downstream market if access to that big data would be limited after the merger. The related evidence presented in these cases mostly included responses to specific market investigations sent out to a selected set of players in the market, a large share of which claimed that there was a sufficient number of other data sources available to the various players in the market.[4] This was then taken to conclude that the data-related anticompetitive impact of a merger was limited.

---

[1] Norbert Maier is an economist at the economic consultancy Copenhagen Economics. This paper benefits from the author's experience as economist at the Directorate General for Competition of the European Commission. However, all the information used in this paper is taken from publicly available sources, including the public version of merger decisions. The author would like to thank Thomas Buettner, Adina Claici, Gergely Csorba, Claus Kastberg Nielsen, Gábor Koltay, Gregor Langus, Asger Lunde, Zsolt Macskási, Zoltán Marosi, Jozsef Molnar, Martin Thelle, Tommaso Valletti and Julia Wahl for helpful comments. The views expressed in this paper are exclusively those of the author.

[2] There is no unique definition of "big data". In this paper big data is used as a shorthand for potentially very large databases computed from data collected or generated from multiple sources and that can be both structured and unstructured data. For example, user data collected by a global digital platform that includes both demographic (structured) data as well as activity (non-structured) data easily qualifies as big data. The term big dataset will be used for the big data collected by individual platforms.

[3] Very often, these platforms do not sell their big data but use it to develop products and services.

[4] One could often argue that the set of players to whom market investigation questionnaires were sent could be extended to cover additional players that could potentially be affected by a transaction and that responses provided by these additional players could have potentially provided somewhat diverging views.

It must be noted, however, that this strong reliance on claims on the availability of alternative big datasets implicitly assumes that these big datasets are substitutable. Unfortunately, this assumption has not been tested and the potential implications have not been explored in merger control cases.[5] In particular, these cases avoid formally defining a separate market for big data, nor do they assess the closeness of substitution between big datasets.

This paper attempts to fill in this gap and offers an approach to assess the closeness of substitution between big datasets that are not traded. It then uses two recent mergers (Microsoft/LinkedIn and Facebook/WhatsApp) investigated by the European Commission (the "Commission") as illustrative examples on how to develop such an assessment.

The starting point of the approach is that big data is of limited value on its own and becomes valuable only when being processed and transformed into insights along the big data value chain.[6] Therefore, closeness of substitution between two big datasets should be assessed by evaluating to what extent the users of the insights derived from them view those insights as close substitutes. For example, the big data collected by Google (focusing on user search) can be viewed as being a close substitute of the big data collected by Facebook (focusing on social networking activity) for a specific group of advertisers only if these advertisers view the generated customer profiles from those two big datasets as close substitutes.

This paper looks into the issue of closeness of substitution between datasets from a purely economic point of view. Therefore, it abstracts from other big data related issues such as privacy in case of personal big data, that are also important in actual competitive assessments. Furthermore, the paper also does not aim to assess data-related theories of harm. Accordingly, for the two mergers discussed later in this paper the focus is on the closeness of substitution of big datasets rather than the substantive assessment of the merger decisions.[7]

The structure of the paper is as follows. Section 2 provides a literature review. The main analytical approach is presented in Section 3. Section 4 and 5 apply the analytical approach to the Microsoft/LinkedIn and the Facebook/WhatsApp merger cases as investigated by the Commission. Section 6 concludes, while the Annex takes one step further and presents a quantitative technique that could be applied to the Microsoft/LinkedIn setup.

2. CASE LAW AND LITERATURE REVIEW[8]

The case law involving mergers with a big data aspect is quite rich. For example, over the past ten years the Commission investigated the following 8 mergers where big datasets that were not traded were involved: (i) *Google/DoubleClick*, (ii) *Microsoft/YahooSearch*, (iii) *Telefonica/Vodafone/ Everything Everywhere JV*, (iv) *Publicis/Omnicom*, (v) *Facebook/WhatsApp*, (vi) *IMS Health/Cegedim*,

[5] In fact, very often the big data collected by various platforms are often complements as they are generated from recording different types of user activities.

[6] Note that even when big data is traded, its buyers use it as an input and develop products by processing it.

[7] Also, any closeness of substitution insight needs to be assessed in conjunction with other types of evidence collected by the competition authorities.

[8] When discussing the relevant literature, this section focuses only on the literature directly related to the closeness of substitution of big datasets and does not aim to review the literature connected to the various other aspects of the two discussed mergers or the literature on various big data-related theories of harm.

(vii) *Sanofi/Google/DMI JV*, and (viii) *Microsoft/LinkedIn*.[9] These cases mostly involved vertical foreclosure issues where the investigation focused on whether some part of the big data generated by the merging parties qualified as an essential input.[10] However, none of these merger assessments undertook a formal analysis, i.e. one going back to economics first principles, of the closeness of substitution between various big datasets.

In addition to some of these mergers (e.g. *Google/DoubleClick* and *Facebook/WhatsApp*) the FTC briefly looked too into a number of mergers where non-traded datasets were involved. It issued early termination notices, i.e. fast-track approval, for the (i) *Google/Nestlabs*, (ii) *Google/Dropcam*, (iii) *Google/Waze* and (iv) *Alliance Data Systems Corp/Conversant* mergers.[11] These early termination notices do not provide insights into the big data related theories of harm nor justification for the merger approval decisions. The third one of these mergers, the *Google/Waze* acquisition was also investigated by the UK OFT and cleared on the grounds of insufficient scale of big data accumulation rather than through a formal assessment of the closeness of substitution between big datasets. In turn, the *Alliance Data Systems Corp/Conversant* merger was investigated by the German Bundeskartellamt too, but again, with no formal assessment of closeness of substitution between big datasets.

While the majority of big data related merger cases have been linked to the FTC in the US, the DOJ also took a part in looking into such mergers by successfully challenging the *Bazaarvoice/Power-Reviews* merger. The focus in that assessment was on how big data can reinforce network effects and no closeness of substitution between big data was evaluated.

A common feature of these merger clearing decisions is that while they identify potential data-driven competition issues they do not define relevant markets for data, nor do they assess closeness of substitution between big datasets. A Joint Research Centre study by the Commission observes that this could be a problem by saying "*Data however are mostly intermediary goods that are used in production processes by other parties. There may be few substitutes available and the production process can be strongly cumulative.*"[12]

In order to improve the analysis of data-driven mergers and other antitrust cases competition authorities started publishing various competition policy related studies dealing with big data. One of the first such studies is the joint study by the Authorité de la Concurrence and the Bundeskartellamt on Competition Law and Data.[13] Similarly, the Dutch Ministry of Economic Affairs commissioned a

[9] In two additional mergers, *Thomson/Reuters* and *TomTom/TeleAtlas*, the merging parties supplied data as their core business.

[10] For a more detailed overview of the assessment of big data related theories of harm in these mergers see M Kadar and M Bogdan, ""Big data" and EU merger control – An overview", *Journal of European Competition Law & Practice*, 8 (8), 2017.

[11] Some other data-related mergers investigated by the FTC and that included traded data included: *Hearst Trust/First Databank*, *Reed Elsevier/ChoicePoint*, *Fidelity National Financial/Chicago Title Corporation*, *CCC Holdings/Aurora Equity Partners*, *Nielsen/Arbitron* and *Dun&Bradstreet/Quality Education Data*. All of these cases included horizontal data-related theories of harm. For a brief discussion see D Feinstein, "Big Data in a Competition Environment", *CPI Antitrust Chronicle*, May (2), 2015, https://www.competitionpolicyinternational.com/assets/Uploads/FeinsteinMay-152.pdf.

[12] N Duch-Brown, B Martens and F Mueller-Langer, "The economics of ownership, access and trade in digital data", *JRC Digital Economy Working Paper* 2017-01, https://ec.europa.eu/jrc/sites/jrcsh/files/jrc104756.pdf

[13] Authorite de la Concurrence and Bundeskartellamt: "Competition law and data", 2016, http://www.autoritedelaconcurrence.fr/doc/reportcompetitionlawanddatafinal.pdf

study on Big Data and Competition.[14] Both studies look at big data from a competition policy angle but do not discuss explicitly the issue of closeness of substitution. A discussion paper by the Canadian Competition Authority, supporting its report summarizing key competition policy and enforcement themes related to big data, manages to take a further step.[15] For the case when data is traded it notes that "*the closeness of competition between two firms selling data will depend on the extent to which customers view their products as substitutable*", then it draws a wider conclusion when claiming that "*two sources of data are more likely to be viewed by customers as substitutable when they provide the same or similar information (e.g. similar financial data)*". Finally, the Directorate General for Competition of the European Commission has launched a consultation on competition policy in the era of digitization that may receive input on these matters.[16]

Finally, some countries started launching sector inquiries in online advertising where the closeness of substitution between various non-traded big datasets is particularly relevant. In early 2018 the German Bundeskartellamt launched its sector inquiry[17], the UK House of Lords Select Committee on Communications recommended the UK Competition and Markets Authority to undertake a market study of the digital advertising market[18] and the French Authorité de la Concurrence concluded its sector inquiry.[19] The sector inquiry by the Authorité de la Concurrence discusses the importance and value of the user data collected by Facebook and Google for their advertising services but does not explicitly analyse the closeness of substitution between the two datasets. It is not yet clear at the time of writing this paper to what extent the other two countries, Germany and the UK will look into the issue.

Academic studies offer limited help too. The most insightful paper that is also closest to the topic of this paper is Graef (2015) that discusses in detail market definition for big data as well as substitutability of different types of data.[20] First, in relation to market definition the paper observes that in order to determine the boundaries of a potential relevant market for data the key issue is the substitutability of different types of data. Second, the paper also realizes that the data collected by various types of digital platforms, like a search engine, a social network and an e-commerce platform, may belong to different market segments within a defined big data market, e.g. data market for online advertisers. Third, the paper also notes that the substitutability of various types of big data can be assessed by looking at the functionality of the service offered by the company (platform) collecting the data. However, the paper does not explicitly use the big data value chain (the core element of the approach developed in this paper) to discuss substitution between various big datasets in general.

---

[14] Dutch Ministry of Economic Affairs: "Big data and competition", *ECORYS* 2017

[15] Canada Competition Bureau: "Big data and innovation - Implications for competition policy in Canada", 2018 http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/vwapj/Big-Data-e.pdf/$file/Big-Data-e.pdf

[16] European Commission Directorate General for Competition, Call for contributions: Shaping competition policy in the era of digitisation, http://ec.europa.eu/competition/scp19/.

[17] See https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2018/01_02_2018_SU_Online_Werbung.html.

[18] See https://www.parliament.uk/HLComms-advertising-industry.

[19] See http://www.autoritedelaconcurrence.fr/pdf/avis/18a03.pdf.

[20] I Graef, "Market definition and market power in data: The case of online platforms", *World Competition* 38(4), 2015, pp.473-506, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2657732.

Some of these insights are further accentuated in Bourreau, de Streel and Graef (2017), a study explicitly claiming that "*The point is that data about one consumer collected by one company may not be a substitute for data about the same consumer collected by another: they may concern different aspects of the consumer and not have the same value on the advertising market.*"[21]

Stucke and Grunes (2016) also touches upon the substitutability of big datasets, however, the main related point of the book is that such big datasets are very often complements and they explain how that could lead to a horizontal theory of harm for mergers.[22]

Finally, substitutability of big datasets is also discussed, although in a different context, by Lambrecht and Tucker (2015).[23] This paper evaluates the substitutability of big datasets from the point of view of companies that do not have such big data but are active in or want to enter markets where players disposing of big data operate. By citing several such examples, the authors find that companies not having access to accumulated big datasets could still enter many of such industries and conclude that big data is not non-substitutable.[24]

3. CLOSENESS OF SUBSTITUTION OF BIG DATA

The Commission's Horizontal Merger Guidelines, the US Horizontal Merger Guidelines and the UK Merger Assessment Guidelines offer many ways to assess closeness of substitution between products. For example, paragraph 29. of the Commission's Horizontal Merger Guidelines recommends that "*(w)hen data are available, the degree of substitutability may be evaluated through customer preference surveys, analysis of purchasing patterns, estimation of the cross-price elasticities of the products involved, or diversion ratios.*"

These recommendations, however, can only be applied directly to products that are, or can be, traded and for which, therefore, there is a market. They are of limited use for products for which currently there exist no market such as big data recorded by some digital platforms.[25] In such cases, no market data is generated that could be used for competition analysis.

Although big data recorded by some digitals platform are not traded, it does not mean that they are not valuable. In fact, such big data should be viewed as an input that is not traded and for which there is no market, and the value of which is coming from its contribution to a certain product or service. A simplified version of the big data value chain, shown below, helps to understand how big data can generate value once it is processed.

---

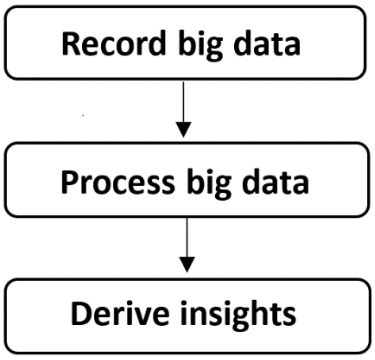
[21] See M Bourreau, A de Streel and I Graef, "Big Data and Competition Policy: Market power, personalised pricing and advertising", *CERRE Project Report*, 2017, http://cerre.eu/sites/cerre/files/170216_CERRE_CompData_FinalReport.pdf.

[22] See chapter 8.C of ME Stucke and AP Grunnes, *Big Data and Competition Policy*, Oxford University Press, 2016.

[23] See A Lambrecht and C Tucker, "Can Big Data Protect a Firm from Competition?", *CPI Chronicle*, January, 2017, https://www.competitionpolicyinternational.com/wp-content/uploads/2017/01/CPI-Lambrecht-Tucker.pdf.

[24] This conclusion is based on the presented examples and not developed from a complementary formal assessment. The more general conclusion could be that big data is not always non-substitutable.

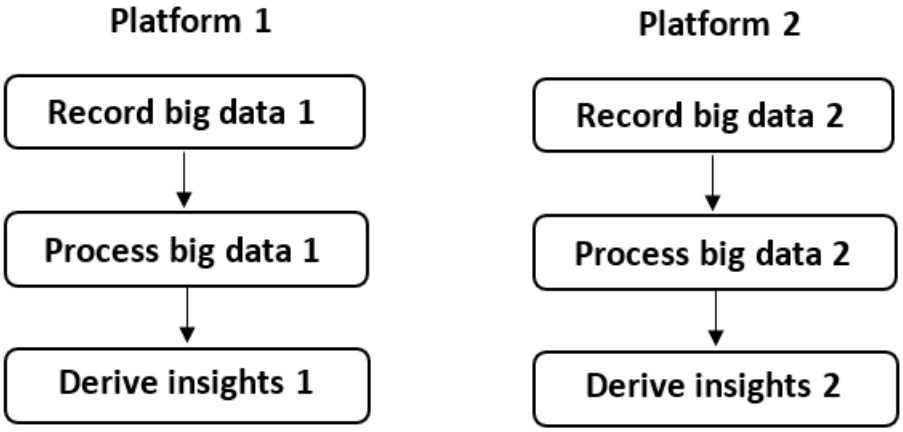
[25] The market investigation might provide some information on what the likely alternative uses are or what inputs are likely to be close alternative for a given use. This fits very well in the current practice of competition authorities to support theories of harm with evidence collected from various sources.



In particular, big data is collected, potentially from different sources, in the *Record* stage. After storing and integrating the data from various sources, big data is processed through traditional statistical methods or machine learning algorithms in the *Process* step.[26,27] It is this processing that leads to the *Derive(d) insights* that can then be used by various customers.

It must be noted here that the insights derived through processing a certain big dataset also depend on the usage objective of the insights. For example, LinkedIn data can be processed to derive insights for recruiters on whom to target with their job advertisements. Or, LinkedIn data can be processed to derive insights for sales and marketing support purposes, helping sales and marketing professionals to find the right person to contact within a sales target company.

This interpretation of the big data value chain and the fact that customers of the insights derived from big data cannot make use of the big data without processing it implies that these customers will think about big datasets collected by two platforms as substitutes if the insights derived from these big datasets are substitutes for them. The following figure illustrates this idea.



[26] The steps of storage and integration between recording and processing big data were left out from the representation as they are not highly relevant for the current discussion.

[27] For the discussion in this paper the most useful definition of machine learning is provided by *techradar.com*, an online publication focused on technology that says: "*Machine learning is the branch of computing that incorporates algorithms to analyze data which is inputted, and via statistical analysis can make a prediction on an output, while incorporating new data as it becomes available, to update the predicted output.*"(https://www.techradar.com/news/what-is-machine-learning). This definition highlights a key feature of machine learning that is relevant for the current discussion: machine learning can be looked at as a sophisticated statistical tool and while it can efficiently analyse big datasets it cannot invent insights that are based on information not included in the analysed big dataset.

To understand the point, one needs to understand the limitations of the *Processing* step, even if it includes machine learning. In other words, one cannot derive insights that would require information not included in the original dataset. For example, assume that in the figure above *Platform 1* is a professional social network and *Platform 2* is a retail platform and both are recording large amounts of user data. Then, a company selling video games and thinking about advertising its product on one of the two platforms, e.g. through targeted advertising, is highly unlikely to view the two big datasets as close substitutes. The reason for this is that one platform is collecting data on users' job characteristics and professional connections whereas the other one is collecting information on users' searches of certain consumer products. The insights derived from these data, regardless of how efficient the processing of the data is, are likely to be different and not be viewed as close substitutes from the video games seller's point of view.

This approach also relies on the assumption that the processing of the big data is efficient and that the algorithms and their developers in the two processing phases are equally or almost equally efficient. This means that the big data processing team of, say, platform 2 would be able to derive insights of similar value (on the downstream market) from the big data of platform 1 as the big data processing team of platform 1.[28] This can be a reasonable assumption for many digital platforms, especially the large global ones coming under the scrutiny of the Commission, the FTC or the DOJ. These digital platforms have large big data processing teams with top talent and extensive expertise. For example, it would be reasonable to think that the algorithms and the big data processing team of Google could derive similar insights from the data collected by Facebook as the algorithms and the big data processing team of Facebook itself.

The following example illustrates how the big data value chain approach presented above can be used to assess closeness of substitution for big datasets from the point of view of users of the derived insights. Let's look again at Facebook and Google. A sportswear and gear manufacturer may find the user big data collected by Facebook and Google close substitutes as users leave quite a few traces on their interest in certain sports (football, running, sailing etc.) on these platforms and therefore both platforms can develop quite accurate user profiles on this dimension. In turn, a household appliance manufacturer may not find the user big data collected by the two platforms to be close substitutes because users leave fewer traces on Facebook about their intention of buying a new household appliance, e.g. a washing machine, than they do on Google, which can record all such related searches and activities (e.g. reading reviews) of its users.

In the implementation of this approach competition authorities can consult the users of the various insights derived from the combination of various datasets. As part of this consultation these users could be asked, for example, to describe the service (insights) that they are using, test their knowledge of the underlying big datasets and discuss their views on other insights, some of which could or would be derived from different existing big datasets. In some cases, one could design some quantitative exercises to support this discussion.[29] This approach is justified by the fact that the users

[28] To adjust to the specificities of merger control, it is enough to assume that the big data processing team of platform 2 would be able to derive insights of similar value (on the downstream market) from the big data of platform 1 as the big data processing team of platform 1 within the two years, after an initial adjustment period.

[29] The Annex will include such an exercise for the Microsoft/LinkedIn merger.

of insights derived from various big datasets are very often large, sophisticated economic actors who are able to discuss such questions.

The next two sections further illustrate this approach through the examples of the two largest recent big data related mergers, the Microsoft/LinkedIn merger and the Facebook/WhatsApp merger.

4. THE MICROSOFT/LINKEDIN MERGER

The Commission investigated the proposed Microsoft/LinkedIn merger in 2016 and discussed three theories of harm in its Phase I clearance decision ("Decision" for the rest of this section). This section focuses on the theory of harm involving big data aspects.

This theory of harm focuses on Microsoft as a global provider of customer relationship management (CRM) software solutions. CRM software solutions help companies to manage their interactions with customers, including marketing, sales and after sales customer support.[30] Microsoft competed with several other software providers in the CRM software solution market, including, among others, Salesforce, SAP and Oracle, all of them with higher market shares than Microsoft[31]. A key feature of the market was the introduction of machine learning functionalities in CRM software solutions around the time when the deal was announced.[32] This allowed the efficient processing of business big data for CRM purposes.

The other merging party, LinkedIn, the provider of a global professional social network service, offers sales intelligence solutions that could be used as input in Microsoft's and its competitors' CRM software solutions. Such sales intelligence solutions provide sales professionals with background and contact information about individuals (name, address, place of employment, title and position, contact details, etc.) and firms (product and service portfolio, financial information, organizational structure, industry background, etc.) and they can be used to identify new leads and efficiently reach out to potential customers and relevant decision makers. LinkedIn competed with several other suppliers in the sales intelligence solution market, including, Dun & Bradstreet, Zoominfo, InsideView, Avention, InsideSales and others.[33] The market investigation indicated that the market for sales intelligence solutions was highly fragmented and that customers used different solutions depending on their needs, often multisourcing such solutions.[34] The Decision does not report market shares of various players in the sales intelligence solution market.

One of the key elements of the theory of harm was the LinkedIn full data database.[35] This database included the collection of all the data that LinkedIn gathered about its users, including professional details, posts, connections, interests, endorsements, etc.[36] Through the analysis of this LinkedIn full data database, e.g. using machine learning, one could derive important insights for a CRM software

[30] The theory of harm presented in this paper deals with the marketing and sales phases of the CRM service.
[31] See Table 2 of the Decision.
[32] See paragraphs (196)-(199) of the Decision.
[33] See paragraph (203) of the Decision for more details on these companies.
[34] See paragraph (204) of the Decision.
[35] See footnote 54 of the Decision.
[36] Connections in this paper refers to social connections, i.e. connections initiated by one person and accepted/ confirmed by the other person – two individuals known to have been attending the same school at the same time would not be taken to be connected if there is no accepted/confirmed connection between them. These social connections could vary in intensity depending on the interaction between the two people after the establishment of the connection.

user, including recommendations for best next actions.[37] For example, a user (a salesman) could find the most efficient path to get to another user (a certain procurer in charge in a sales target company).[38] An interesting feature of the case is that at the time of the merger assessment LinkedIn did not market its full data database. LinkedIn's only similar product on the market was the LinkedIn Sales Navigator, a query tool drawing from LinkedIn's user database and displaying a subset of the database. The data collected through LinkedIn's Sales Navigator can be used to identify and create new customer leads and sales opportunities but it is not "big" enough for more complex machine learning analysis.[39]

According to the theory of harm formulated by the Commission, Microsoft would have restricted access to LinkedIn full data, an important input within the meaning of paragraph 34 of the Commission's Non-Horizontal Merger Guidelines, for the purposes of machine learning for competing CRM software solutions, making it harder for them to compete in the CRM software solutions market.[40] The following figure provides an illustration.

It can be seen from this figure that CRM software solutions used data from and solutions from multiple sales intelligence solutions providers to develop their own services.

The Commission dismissed this theory of harm on the following grounds:

i. LinkedIn full data was not sold or licensed in the market and the company's internal documents did not reveal any such plans, therefore there was uncertainty as to whether in the near future LinkedIn full data could become an important input within the meaning of paragraph 34 of the Commission's Non-Horizontal Merger Guidelines.[41]

ii. All major CRM players in the CRM software solutions market launched their own machine learning product around the time of the announcement of the merger.[42]

iii. Respondents to the market investigation indicated that even if the LinkedIn full data has some specific qualities, there is a sufficient number of alternative solutions in the upstream market

[37] See footnote 230 of the Decision.
[38] The efficiency of this use is affected by user data privacy issues, a consideration not discussed in this paper.
[39] See paragraph (58) of the Decision.
[40] See paragraph (246) of the Decision.
[41] See paragraph (250) of the Decision.
[42] See footnote 32.

of sales intelligence solutions that CRM software solutions providers can use in their machine learning based service functionalities.[43]

While these are all valid arguments and evidence to support the dismissal of the theory of harm, the Commission did not undertake a more formal assessment to what extent the available alternative datasets (see condition iii. above) were relevant alternatives to the LinkedIn full data as inputs into CRM software solutions. In particular, the Decision does not include an explicit assessment of the closeness of substitution between these big datasets.[44]

The closeness of substitution between the LinkedIn full data and the data provided by the other sales intelligence solutions suppliers can be analysed by applying the framework presented in Section 3. Using that approach, the recording of big data is done by the sales intelligence solutions providers whereas the processing of the big data (through machine learning) is done by the CRM software solutions providers.[45]

In this framework, the LinkedIn full data and the sales intelligence data provided by the other sales intelligence solutions suppliers could be viewed as close substitutes if the insights derived from processing the two big datasets are close substitutes from the CRM software solution customers point of view.

It must be noted at this stage that the market investigation in the Decision only includes the views of the CRM software solutions providers who are the processors of the sales intelligence big data (i.e. the customers of LinkedIn and its competitors). To implement the approach described in this paper one would have needed to consult the customers of the insights derived from processing that big data, i.e. the large and sophisticated customers of CRM software solutions services, which would have meant that one would have needed to go one further level downstream on the market. In such a consultation, one could focus on that part of the LinkedIn full data that describes the connections among users, i.e. the social network type of information. As such or similar type of information was not identified in the datasets provided by the other sales intelligence solution providers, one could assess explicitly to what extent CRM insights derived by using such specific information are close substitutes for CRM insights derived without using such specific information.[46] Such an assessment could be highly relevant when one wants to find the most efficient way to get in active contact with the relevant procurement person in a potential client company.

It must be noted that this is a formal, yet qualitative assessment. An idea towards a more quantitative assessment of the closeness of substitution between the LinkedIn full data and the other

[43] See paragraphs (274) and (276) of the Decision. It must be noted, for completeness, that this was not an unanimous view among CRM software solutions providers as some of them expressed concerns regarding the existence of relevant alternative data sources (see paragraphs (264) and (273) of the Decision).

[44] At the same time, the Decision also does not include an explicit distinction between the LinkedIn full data as a data asset and the provision of sales intelligence solutions from a relevant market definition point of view.

[45] It must be noted here that some of this sales intelligence data was traded. It is the LinkedIn full data that was not traded and could have been chosen not to be traded after the merger. Furthermore, the data procession by the CRM software solutions providers also used some of their own data collected during their business operation, a detail important in the actual investigation but not for using this merger for the illustration of the approach presented in Section 3.

[46] This argument remains valid even if one compares the LinkedIn full data with the combined data provided by the other sales intelligence solutions providers.

sales intelligence datasets with no information on connections between professionals is presented in the Annex.

5. THE FACEBOOK/WHATSAPP MERGER

The Facebook/WhatsApp merger (2014), provides yet another opportunity to apply the framework presented in Section 3 to assess closeness of substitution between big datasets. The Commission cleared the merger unconditionally in a Phase I decision ("Decision" for the rest of this section).

This was a merger between Facebook, a global provider (both through a website and a mobile application) of social networking, consumer communication and photo and video-sharing services and WhatsApp, a global provider of consumer communication services through a mobile application ("WhatsApp"). Facebook provided its services to users for free and collected revenues from selling advertising spaces. It used data collected from the provision of its consumer services to develop consumer profiles and improve the accuracy of targeted and individualised ads. WhatsApp had a different business model: it provided its consumer services for free in most countries but did not sell advertising spaces.[47]

The Commission examined the following two theories of harm in relation to the proposed merger:

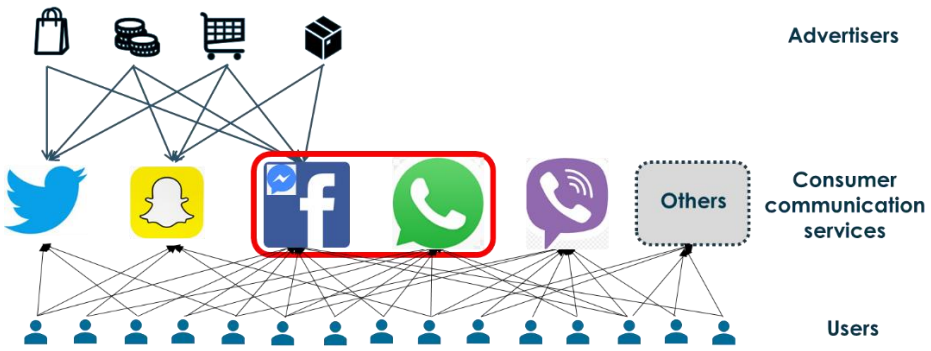
i. The merged entity could introduce targeted advertising on WhatsApp by analysing the data collected from WhatsApp users and use that to reinforce Facebook's position in the online advertising market or some of its sub-segments.[48]

ii. The merged entity could start collecting data from WhatsApp users (by keeping the WhatsApp service ad-free), integrate that use data with the Facebook user data and improve the accuracy of targeted ads served on the latter's social networking platform.[49]

While not stated explicitly in the second theory of harm, improving the accuracy of targeted ads served on Facebook's social networking platform would also lead to reinforcing Facebook's position in the online advertising market or some of its sub-segments. This reinforcing of Facebook's position in the online advertising market would have happened through offering a new platform for advertisers in the first theory of harm and through improving the accuracy of targeted ads placed on its platform in the second theory of harm.

The focus in this section is an assessment of Facebook's position in the online advertising market or some of its sub-segments in the view of the second theory of harm as that is more directly linked to the presentation of big data in Section 3. The following figure provides an illustration of the case.

[47] WhatsApp charged an annual subscription fee of around EUR 0.89 in Italy, the UK, Canada and the US.
[48] See paragraph (168) of the Decision.
[49] See paragraph (180) of the Decision.



It is important to see that consumer communication service providers followed different business models at the time of the merger: some of them (Facebook, Twitter and Snapchat) were funded by advertisers whereas others (WhatsApp and Viber) were not. This way the theory of harm stated that by acquiring and integrating user data from a consumer communication platform that was not financed by advertisers could still be used to strengthen the position in the advertising market for an advertiser-funded consumer communication platform.

The Commission dismissed the theory of harm on the following grounds:

i. The technical integration of the Facebook and WhatsApp user identities and networks is technically difficult.[50]

ii. Facebook had a small market share in the online advertising market where it also faced many competitors.[51]

While it later turned out that the first condition did not hold,[52] the Commission's approval of the proposed deal could have been justified even based on the second condition alone as the clearance decision explained and presented evidence that Facebook did not have a strong position in the online advertising market in the first place.[53]

A key point of the Decision is that it defines a single market for online advertising and uses the responses to the market investigation as supporting evidence.[54] Furthermore, by making reference to Facebook's low market share in the online advertising market the Decision implicitly relies on the assumption that the online advertising market is a market with homogenous services. This assumption, however, also implies that there are no well-defined segments of the online advertising market where Facebook's position could be particularly strong and subsequently reinforced by the merger.[55]

A more formal analysis could provide a test for this assumption. In particular, the online advertising market could also be seen as including players offering differentiated products, leading to a situation

[50] See paragraph (138) and (139) of the Decision.
[51] See paragraph (188) of the Decision.
[52] See case M.8228 – Facebook/WhatsApp by the Commission.
[53] See paragraph (187) and (189) of the Decision.
[54] See paragraphs (74)-(79) of the Decision.
[55] This view is further supported by the responses of Facebook's advertising customers to the market investigation claiming that there are a sufficient number of alternative providers of advertising services that compete with Facebook (see paragraph (177) of the Decision).

where some players are stronger in certain segments of the market, even if their overall market share is not particularly high.

The potential differentiation between the advertising space services offered by various platforms can be best understood by looking at the vast diversity of user data collected by various digital platforms and how such data is valued by various advertisers. For example, a video games producer as advertiser would prefer to advertise on Amazon rather than on LinkedIn, whereas a company advertising a job vacancy would prefer to advertise on LinkedIn rather than on Amazon. Furthermore, a bank advertising its new loan product would prefer to advertise on Google or a national news portal, whereas a restaurant in Stockholm would prefer to do so on TripAdvisor.

More generally, the economic power of some players in the advertising market, or some segments of it, can be assessed by looking into the closeness of substitution between the services provided by them. As these players rely on user data to derive user profiles, the closeness of substitution between their advertising services can be assessed by looking into the closeness of substitution between the underlying user data. The framework introduced in Section 3 can guide this assessment.

In that framework the recording of big data is done by the various digital platforms and so is the processing of the big data in form of consumer profiling. The final consumer profiles developed by individual digital platforms correspond to the insights generated from the big data. The view of the advertising market taken in the Decision assumes that the consumer profiles developed by different digital platforms are in fact close substitutes and therefore the market shares of these platforms, e.g. that of Facebook, in the online advertising market provide an accurate representation of their overall market power in this market.

This, however, may turn out to be an extreme view as the consumer profiles developed by various digital platforms tend to be fairly differentiated. For example, activity on LinkedIn and on Amazon would reveal different features of a certain user's personality and interest and would be of interest to different types of advertisers. Going back to our example used earlier in this section, a producer of video games would be willing to pay different amounts of money for a targeted advertising space on LinkedIn or Amazon. This suggests that it is rather likely that the online advertising market could be further divided into narrower segments. Even if these narrower segments could not be separated by a relevant market definition test they could confer different market position for the various digital platforms providing advertising spaces in those segments.

Based on this logic, a more in-depth assessment of the online advertising market, allowing for potential heterogeneity of the consumer profiles developed by various digital platforms from their own user data and involving a larger set of advertisers than just those of one player (Facebook) into the consultation, could provide more nuanced insights on whether the online advertising market is segmented with various players being stronger in different segments. If such a more in-depth analysis of the online advertising market concluded that the advertising market is segmented, it could have also highlighted that Facebook might have a strong market position in some segments. Clearly, if that was the case, complementing the Facebook user data with the WhatsApp user data could have further strengthened the former's position in that particular online advertising market segment.

6. CONCLUSION

This paper looks into the issue of how to assess closeness substitution between big datasets, a question that is at the heart of merger control dealing with transactions involving big data. As such big data is often not traded, a direct assessment of the closeness of substitution cannot be developed directly, in line with the recommendations formulated in the Commission's Horizontal Merger Guidelines.

The value of these big datasets is coming from the insights derived after processing them and it is these insights and their substitutability that can be used to assess the closeness of substitution of the underlying big datasets. Such an analysis fits perfectly into the existing competition law framework.

ANNEX

This annex provides a high-level presentation of a quantitative technique that could be used to assess the closeness of substitution between LinkedIn full data and the data provided by competing sales intelligence solution providers.

The starting point of the current analysis is that the potential specific feature of the LinkedIn full data lies in the fact that it includes information on the professional connections between people as well as the strength of these connections.

Next, one could think of the following two big datasets:

i. the LinkedIn full data, as defined earlier in this paper;
ii. the LinkedIn reduced data, including the LinkedIn full data, with all the information referring to the existence and nature/intensity of connections between people taken out.

The second dataset could be argued to be very similar to other available sales intelligence datasets as it contains up-to-date information about professionals as well as their professional background. The first dataset is a richer dataset as it adds information on professional social connections to the second dataset.

The main step of this approach would be to consult a number, say 20, of the large and sophisticated customers of the CRM software solutions services about their views on certain sales and marketing insights derived by data scientists and economists of competition authorities from a subset of the LinkedIn full data as well as the corresponding subset of the LinkedIn reduced data.[56] Ideally, these insights should be prepared based on the needs of the CRM software solutions customers.

If a large fraction of these customers would view the insights derived from the two databases as close substitutes that could be taken as evidence of the LinkedIn data being a close substitute for the data provided by the alternative sales intelligence data providers. If, on the contrary, a large fraction of these customers would view the insights derived from the two databases as not so close substitutes that could be taken as evidence of the LinkedIn data not being a close substitute for the data provided by the alternative sales intelligence data providers.

On a final note, a main benefit of this technique is that it does not rely on extensive (big) data collection from alternative sales intelligence data providers. Instead, it can be developed relying on the data of a merging party, LinkedIn, only.

[56] This corresponding subset of the LinkedIn reduced data would be obtained by taking out the social connection type of information from the subset of the LinkedIn full data.