BIG DATA AND MARKET POWER

Alexandre de Streel*

I. The Big Data value chain

According to De Mauro *et al.*,[1] big data "is the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value." Thus, the difference between data and big data is the famous four V's made possible by technological progress: the *volume* of data processed, the *variety* of data aggregated, the *velocity* at which data is collected and used and the *value* of the information found in the data.

As explained by OECD[2] or Mayer-Schonberger and Cukier,[3] big data are collected, exchanged, stored and value is extracted in a complex eco-system made of many related markets which are often multi-sided: first (i) data are *collected* directly from users and from machines in many different ways or can be bought from data brokers;[4] then (ii) data are *stored* on internal

* Professor of law at the University of Namur, Director Centre Information Law and Society (CRIDS), Joint-academic director Centre on Regulation in Europe (CERRE). This paper is partly based on a policy report done in February 2017 for the Centre on Regulation in Europe (CERRE) on Big Data and Competition Policy by Marc Bourreau, Alexandre de Streel and Inge Graef.

1 A. De Mauro, M. Greco and M. Grimaldi, "A Formal Definition of Big Data Based on its Essential Features", *Library Review* 65(3), 2016, 122-135.

2 OECD, *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publications, 2015 and OECD, *Big Data: Bringing competition policy to the digital era*, DAF/COMP(2016)14.

3 V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution that will transform how we live, work and think*, Eamon Dolan/Mariner Books, 2013.

4 Data brokers have been defined by the US FTC as "*companies whose primary business is collecting personal information about consumers from a variety of sources and aggregating, analysing, and sharing that information, or information derived from it, for purposes such as marketing products, verifying an individual's identity, or detecting fraud*": Federal Trade Commission (2014), *Data Brokers: A Call for Transparency and Accountability*. The consultancy IDC has a broader concept of data marketplace defined as

servers or on external cloud computing services; and finally (iii) data are *analysed* with software analytics and the valuable information can be used to improve and personalise products' characteristics and prices as well as their marketing, to improve process and organisation or for many other purposes such as controlling epidemics or managing emergencies.

A. Data collection

A firm may collect data directly, usually having a direct contact with the person or the object from which the data is collected, or indirectly, usually by buying the data from data brokers.

Firms may collect personal and non-personal data about users, as well as machines, in many different online and offline ways. For the case of the online collection of personal information, (i) some information is publicly observed through device, operating system or IP address; (ii) other information is voluntarily provided by the consumers either with knowledge when registering to a website (such as name, date of birth, email or postal address for delivery) or often without knowledge when logging into a website (login-based data: such as products the consumer is looking for, purchases); (iii) information can also be collected by tracking the consumer online with different means, browser and device fingerprinting, history sniffing, cross-device tracking.

Firms may also procure data from third parties, even though this remains a marginal practice in Europe today. According to a recent study done for the European Commission, data trading only accounted for 4% of the companies surveyed.[5] However, the secondary market for data is not yet well

"*a third party, cloud-based software platform providing Internet access to a disparate set of external data sources for use in IT systems by business, government or non-profit organizations. The marketplace operator will manage payment mechanisms to reimburse each dataset owner/provider for data use, as necessary. Optionally, the marketplace provider may provide access to analysis tools that can operate on the data.":* IDC (2016), *Europe's Data Marketplaces – Current Status and Future Perspectives*, Report for the European Commission. IDC notes that in their simplest form, data marketplaces are online stores where firms can buy and sell data.

5 On the basis of a business models survey done by Deloitte, the European Commission observes that "*in the vast majority of cases (78% of the companies surveyed) data is generated and analysed in-house by the company or by a sub-contractor. Vertical integration remains the principal strategy in the sectors surveyed. Data stays within an organisation and is not traded with third parties. This is particularly the case in sectors with a high presence of large, technologically advanced companies, such as banks and telecom providers or automotive and machinery producers.*": Commission Staff Working Document of 10 January 2017 on the free flow of data and emerging issues of the European data economy, SWD(2017) 2, p. 15. For a broad overview of the data landscape in the European Union, see http://datalandscape.eu/.

understood by public authorities. In the US, a Committee of the Senate in 2013[6] and the Federal Trade Commission in 2014[7] conducted inquiries to better understand those markets. The FTC concluded that: (i) the US data broker industry is complex with multiple companies exchanging lots of data between themselves; (ii) data brokers have a vast amount of data on almost every US household and commercial transaction;[8] (iii) data brokers combine offline and online data from multiple different sources and it is very difficult for a consumer to know and determine how a data broker obtained her data; (iv) data brokers analyse the data and make inference about the consumers, placing them into categories that may be sensitive.[9]

B. Data storage and cloud computing

The storage of massive quantities of data requires large data centres consisting of big clusters of computers connected by fast local area networks. Those data centres are expensive to build and characterised by important economies of scale. However, the development of cloud computing[10] allows small firms to rent instead of owning the data centres, thereby converting

6 US Senate Committee on Commerce, Science and Transportation (2013), *A Review of the Data Broker Industry: Collection, Use and Sale of Consumer Data for Marketing Purposes*, Staff Report.

7 This report examines the following 9 data brokers: Acxiom, Corelogic, Datalogix, eBureau, ID Analytics, Intelius, PeekYou, Rapleaf, and Recorded Future.

8 According to the FTC Report, one data broker's database has information on 1.4 billion consumer transactions and over 700 billion aggregated data elements; another data broker's database covers one trillion dollars in consumer transactions; and yet another data broker adds three billion new records each month to its databases. Most importantly, data brokers hold a vast array of information on individual consumers. For example, one of the nine data brokers has 3000 data segments for nearly every U.S. consumer. Based on companies' reports, Lambrecht and Tucker note that Acxiom has multi-sourced insight into approximately 700 million consumers worldwide with over 1,600 pieces of separate data on each consumer; Datalogix asserts that its data includes almost every U.S. household; Bluekai states that it has data on 750 million unique users per month with an average of 10-15 attributes per user: A. Lambrecht and C. Tucker, "Can Big Data Protect a Firm from Competition?", 2015, available on SSRN.

9 Potentially sensitive categories include those that primarily focus on ethnicity and income levels, such as "Urban Scramble" and "Mobile Mixers," both of which include a high concentration of Latinos and African Americans with low incomes. Other potentially sensitive categories highlight a consumer's age such as "Rural Everlasting," which includes single men and women over the age of 66 with "low educational attainment and low net worths," while "Married Sophisticates" includes thirty-something couples in the "upper-middle class with no children." Yet other potentially sensitive categories highlight certain health-related topics or conditions, such as "Expectant Parent," "Diabetes Interest," and "Cholesterol Focus."

10 Cloud computing service is defined by EU law as a digital service that enables access to a scalable and elastic pool of shareable computing resources: Article 4(19) of the Directive 2016/1148 on Network Information Security.

their fixed costs into variable costs.[11] For the cloud computing market to function properly, the costs of switching between providers need not to be too high, which raises the issues of interoperability and portability in the cloud.[12] Moreover, the competition among cloud providers may be limited by data localisation restrictions which can be important for certain types of privately owned data, in particular for health, financial and gaming/gambling data as well as for publicly owned data.[13]

C. Data analytics and use

The third step in the big data value chain is the analysis of those data to extract relevant information, mainly with correlation patterns. The information found in the data can have multiple uses: they can improve products for all thanks to a better understanding of consumers' needs. Those improved products can be data-rich (mainly intangible) such as a map or data-less rich (more tangible) such as a drive-less car;[14] they can better personalise products' prices or marketing strategies; they can also improve process, marketing and organisation, thereby increasing productive and dynamic efficiencies.

Data analytics is done by applications and algorithms which are increasingly learning by themselves. The development and the improvement of those algorithms are based on many inputs such as data, skilled and creative labour force (in particular computer scientists and engineers) or capital. Thus, data are important but probably not the most important input as mentioned by Lerner.[15] Analytical applications and algorithms can be developed in-house and, for some, may require important investment in getting the best skills and volume of data. They may also be obtained from

11 S.M. Greenstein, A. Goldfarb, and C. Tucker, *The Economics of Digitization*, Edward Elgar, 2013.

12 Communication of the Commission of 19 April 2016, European Cloud Initiative – Building a competitive data and knowledge economy in Europe, COM(2016) 178; European Commission Staff Working Document of 10 January 2017 on the free flow of data and emerging issues of the European data economy, SWD(2017) 2, pp. 5-10.

13 To remove those restrictions, the Commission has recently propose a regulation to stimulate the flow of data across the Member States: Proposal from the Commission of 13 September 2017 for a Regulation on a framework for the free flow of non-personal data in the European Union, COM(2017) 495.

14 J. Prüfer and C. Schottmüller, "Competing with Big Data", CentER Discussion Paper; Vol. 2017-007 have built a model where firms can leverage their position from data-rich markets (for instance maps) to data-less rich markets (for instance drive-less cars) with products improvements based on data.

15 A. Lerner, "The Role of 'Big Data' in Online Platform Competition", 2014, available on SSRN.

a third party. In this case, like for cloud computing, the fixed development costs can be converted into variable costs.

D. The broader context

The big data eco-system is complex and involves many firms active on related markets which are often multi-sided. Therefore when assessing market power, competition authorities should keep a broad view taking into account the main characteristics of the eco-system, which are:[16] (i) the presence of direct and indirect network effects which may lead to snowball effects where the markets tip in favour of a small number of players;[17] (ii) the steep experience curve of the self-learning algorithms which may substantially increase the first-mover advantage and also lead to snowball effects and market tipping; (iii) the relationships between the different markets, often of multi-sided nature, which may, in some circumstances, ease the leverage of a dominant position from one market to another; (iv) the extensive multi-homing of customers which may be affiliated to several online platforms at the same time;[18] (v) the rate of innovation which is quick and often unpredictable and disruptive, leading to possible rapid displacement of powerful but lazy firms.

Therefore, when running an antitrust case in a big data industry, competition agencies should take into account the general characteristics of the big data eco-system which may, in some circumstances, amplify some of the effects of data control. In particular, they should take into account the direct and indirect network effects on the demand side and the multi-sidedness of the markets and leveraging possibilities on the supply side. The agencies should also adopt a dynamic view of market evolution without trying to predict or, worse, shape the future technology and market evolutions.

---

16 Autorité de la concurrence and Bundeskartellamt, *Competition law and data*, 2016, pp. 26-30 and UK House of Lords, *Online Platforms and the Digital Single Market*, Report of the Select Committee on European Union, 2016, chapter 4.

17 P. Belleflamme and M. Peitz, *Industrial Organisation: Markets and Strategies*, 2nd ed., Cambridge University Press, 2015; C. Shapiro and H. Varian, *Information Rules – A Strategic Guide in the Information Society*, Harvard Business School Press, 1999. Network effects often change the type of competition, which is often for the market (Schumpeterian) rather than in the market.

18 Multi-homing has been recognised by the General Court of the EU as a mitigating factor for finding dominance when it upheld the Commission approval of the acquisition of Skype by Microsoft: Case T-79/12, *Cisco and Messagnet v. Commission*, ECLI:EU:T:2013:635, paras. 79 et sq.

II. Dataset and market power

To determine whether the control of a dataset can be a source of market power in a big data value chain, a joint report by the French Autorité de la concurrence and the German Bundeskartellambt[19] mentions two relevant factors to analyse: (i) the scarcity (or ease of replicability) of data and (ii) whether the scale/scope of data collection matters to competitive performance. More generally, it is important to assess the effects of the possible entry barriers in the main steps of this value chain. This assessment should be done on a case-by-case basis and depends very much on the type of data and the type of use, in particular its algorithmic treatment, of such data.[20] This paper focuses on data collection and data analysis where the risks of entry barriers, because of potential non-replicability, are *a priori* higher than for data storage.

A. Data collection

1. Costs of collecting data

One of the inherent characteristics of data, as many intangible goods, is the non-rivalry which means that the same data can be collected and used many times without losing value.[21] This fundamental characteristic decreases, *ceteris paribus*, the cost of collection as data collection by one firm does not impede other firms to do the same.

However, technical, legal or contractual restrictions may weaken or even remove the inherent non-rivalry of data to make them exclusive:[22] (i) technical barriers, such as encryption techniques, can make the collection of data more difficult or even impossible; (ii) some of the legal rules applicable to personal or non-personal data[23] increase the costs of collection, in particular

19 Autorité de la concurrence and Bundeskartellamt, *Competition law and data*, 2016, p. 35.

20 D.L. Rubinfeld and M.S. Gal, "Access Barriers to Big Data", *Arizona Law Review* 59, 2017, 339. For instance, the 2014 DoJ's action against the merger of *Bazaarvoice* and its leading rival *Power-Reviews* established that data can serve as an entry barrier in the market for rating and review platforms: DOJ, Antitrust Division, Competitive Impact Statement of 8 May 2014, 13-cv-00133 WHO, http://www.justice.gov/atr/case-document/file/488826/download.

21 This is why the often used analogy between data and oil is misleading.

22 See I. Graef, *EU Competition Law, Data Protection and Online Platforms: Data as Essential Facility*, Kluwer Law International, 2016.

23 For a description of those rules, see European Commission Staff Working Document of 10 January 2017 on the free flow of data and emerging issues of the European data economy, SWD(2017) 2, pp. 19-22; M. Bourreau, A. de Streel, I. Graef, Big Data and Competition Policy, CERRE Report, 2017, pp. 15-28.

for personal data whose means of collection are limited by the general data protection rules and even more by the telecom specific data protection rules; (iii) contractual barriers such as exclusivity clauses, on the transfer of data.[24]

In some cases, data are collected for their own sake and the collecting firm is ready to invest solely to gather data. This was the traditional business model of polling institutes and market research firms such as GFK or TNS or financial information companies such as Bloomberg or Reuters. This is also the new business model of the (often) Internet firms offering products which are supposedly free because they are not paid with money but with data (whose value has increased with the development of big data). Interestingly, this new business model exhibits in general more network and experience effects than the traditional model and may lead to large platforms offering supposedly free services and collecting many (often personal) data.[25] In other cases, data are collected as by-product of the selling of another goods or services and the collecting firm does not invest specifically to gather the data. The standard example is customers' lists that firms are building over time as they sell their products. This data collection as by-product increases with proliferation of connected devices and the diminishing costs of information storage. The cost of collection is obviously higher when data are collected for their own sake than as by-product. To be sure, those are two extreme cases and reality often lies in-between. Indeed, there are some cases where a firm may want to invest in improving its products just to get better data as by-products.

2. Antitrust assessment of data availability and replicability

On the basis of the availability of data and their collection costs, the competition agencies have determined in several cases whether datasets were replicable.

In some merger cases involving data-rich Internet firms, the Commission concluded that datasets of the merging parties were replicable, hence the combination of those data would not significantly impede competition. In *Google/DoubleClick*,[26] the Commission considered that the combination of

[24] For an analysis of some of those contractual clauses, see Osborne Clark, *Legal study on ownership and access to data*, Study for the European Commission, 2016.

[25] See also Autorité de la concurrence and Bundeskartellamt, *Competition law and data*, 2016, p. 38.

[26] Commission Decision of 11 March 2008, Case M.4731 *Google/ DoubleClick*, paras. 364-366.

the information on search behaviour from Google and web-browsing behaviour from DoubleClick was already available to a number of Google's competitors, hence would not give the merged entity a competitive advantage. In *Facebook/WhatsApp*,[27] the Commission considered that a large amount of Internet user data that are valuable for online advertising are not within the exclusive control of Facebook. Therefore, even if Facebook would use WhatsApp as a potential source for user data to improve Facebook's target advertising, this would not significantly impede competition on the online advertising market.[28] In *Microsoft/LinkedIn* the Commission considered that no competition concerns on the market for online advertising arose from the concentration of the parties' user data that can be used for advertising purposes because a large amount of such user data would continue to be available after the transaction.[29]

However, in abuse of dominance cases, some national competition authorities decided that customer lists gathered by firms enjoying a legal monopoly may not be reproducible by competitors[30] and cannot be used to launch other services which are under competition.

In September 2014, the Autorité de la concurrence adopted an interim decision in which it found GDF Suez (now Engie) capable of taking advantage of its dominant position in the market for natural gas by using customer files it had inherited from its former monopoly status to launch offers at market prices outside the scope of its public service obligation. As regards the reproducibility of the database, the French agency considered that it was not reasonably possible for the competitors to reproduce the advantage held by GDF Suez or to rely on other databases from which information could be retrieved that was effective for prospecting new customers in the market for the supply of gas. In March 2017, the French authority condemned Engie and imposed a fine of 100m €.[31]

---

27 Commission Decision of 3 October 2014, Case M.7217 *Facebook/WhatsApp*, paras. 167-189.

28 Note that at the time of the merger in August 2014, Facebook indicated to the Commission that it was unable to establish a reliable automated matching between Facebook and WhatsApp users' accounts. In August 2016, WhatsApp announced the possibility of linking WhatsApp user phone with Facebook user identities and it appears that technical possibilities of automatic matching existed already in 2014 contrary to Facebook allegation. On that basis, the Commission fined Facebook €110 million for providing misleading information: Commission Decision of 18 May 2017, Case M.8228 *Facebook/WhatsApp*.

29 Commission Decision of 6 December 2016, Case M. 8124 *Microsoft/LinkedIn*.

30 For a more elaborate analysis of indispensability of data, see I. Graef, *op. cit.*, pp. 270-273.

31 Autorité de la concurrence, Décision 14-MC-02 du 9 septembre 2014 relative à une demande de mesures conservatoires présentée par la société Direct Energie dans les secteurs du gaz et de l'électricité, paras. 147-154; Décision 17-D-06 du 21 mars 2017.

In September 2015, the Belgian competition authority imposed a fine on the National Lottery for having abused its dominance in the Belgian market for public lotteries in which it has a legal monopoly. When entering the competitive market for sports betting in 2013, the National Lottery used the contact details of individuals contained in a database that it had established in the context of its legal monopoly in order to send a one-off promotional email for the launch of its new sports betting product Scooore!.[32] Given its nature and size, the Belgian competition authority concluded that the contact details could not have been reproduced by competitors in the market at reasonable financial conditions and within a reasonable period of time.[33]

B. Data analysis

The relationship between, on the one hand, the volume, the variety and the velocity of the data and, on the other hand, the quality of the analytical tools (often a computer algorithm) and the value of the inferred information is complex and changing with rapid progress in artificial intelligence. It mainly depends on the type of data and analysis performed.

1. Volume of data: the economies of scale

The marginal return of having more data depends very much on the type of data and the type of analysis which is done. Economies of scale in data use may be low when data are used for inference purpose, but higher for other usages. For search services, the economies of scale are lower for head queries which are frequently entered by users than for tail queries[34] which are rarer. As noted by the German Monopolkommission: "While the added value of a frequently searched term can thus be very low, seldom-made search queries may make a major contribution towards improving search results. Such infrequent search queries are likely to particularly include those search queries concerning for instance current events with regard to

[32] Belgian Competition Authority, Beslissing BMA-2015-P/K-27-AUD van 22 September 2015, Zaken nr. MEDE-P/K-13/0012 en CONC-P/K-13/0013, *Stanleybet Belgium NV/Stanley International Betting Ltd en Sagevas S.A./World Football Association S.P.R.L./Samenwerkende Nevenmaatschappij Belgische PMU S.C.R.L. t. Nationale Loterij NV*, paras. 44-48.

[33] *Ibidem*, paras. 69-70.

[34] A. Lerner, *op. cit.* defines tail queries as including misspelled queries, addresses, specific product descriptions or model numbers, and detailed queries composed of multiple terms.

which there is as yet no information on users' conduct, and search queries consisting of several terms, "long-tail queries".[35]

However, the level of those scale economies is not clear. Some authors like Lerner,[36] Lambrecht and Trucker,[37] or Sokol and Comerford[38] submit that the scale economies are low even for tail queries and that there is a diminishing marginal return of data both for head and tail queries.[39] Others like Mc Afee find that more data matters for tail queries. In addition, in the context of the *Microsoft/Yahoo! Search Business* merger decision, Microsoft argued that with larger scale a search engine can run tests on how to improve the algorithm and that it is possible to experiment more and faster as traffic volume increases because experimental traffic will take up a smaller proportion of overall traffic.[40] The extent of the economies of scale is thus an empirical question which should be tested in each case on the basis of the type of data and application at hand. In particular, the necessity of having more data to improve the quality of the application and the algorithm should be carefully analysed.

2. Variety of data: the economies of scope

Another characteristic of the big data resides in the capacity and the importance of combining different types of data, which points to the economies of scope. As noted by the US Executive Office of the President,[41] the combination of data from different sources "may uncover new meanings. In particular, data fusion can result in the identification of individual people, the creation of profiles of an individual, and the tracking of an individual's activities."[42] Similarly, the European Commission states, in its *Google/DoubleClick* merger decision, that "competition based on the quality of collected data thus is not only decided by virtue of the sheer size

35 Monopolkommission, *Competition policy: The challenge of digital markets*, Special Report 68, 2015, para. 202.

36 A. Lerner, *op. cit.*, p. 37.

37 A. Lambrecht and C. Trucker, *op. cit.*, p. 10.

38 D. Sokol and R. Comerford, "Antitrust and Regulating Big Data", *Georges Mason Law Review*, 2016, 1129-1161.

39 See also Z. Dou, R. Song and JR Wen, "A Large-scale Evaluation and Analysis of Personalizes Search Strategies", Paper presented at the IW3C Conference, 2007.

40 Commission Decision of 18 February 2010, Case M. 5727 *Microsoft/Yahoo! Search Business*, paras. 162 and 223.

41 US Executive Office of the President, *Big Data and Privacy: a Technological Perspective*, 2014.

42 See also the UK Information Commissioner's Office, *Big data and data protection*, 2014, para. 25 observing that variety is the most important characteristic of big data.

of the respective databases, but also determined by the different types of data the competitors have access to and the question which type eventually will prove to be the most useful for internet advertising purposes."[43] This is also an empirical question which should be tested in each case on the basis of the type of data and application at hand. In particular, the necessity of having more variety of data to improve the quality of the application and the algorithm should be carefully analysed.

3. Depreciation value of the data and velocity of the analysis

Many types of data are transient in value and only relevant over a short period of time, hence their depreciation rate is very high.[44] As noted by Autorité de la concurrence and Bundeskartellambt,[45] "historical data, while useful for analysing trends in advertising markets, may have comparatively little value for instant decision making such as the choice of which ad to display in real-time bidding. Moreover historical data may be of relatively low value for some actors like search engines in view of the high rate of new search queries: as reported by Google, 15% of every day people's searches are new, implying that algorithms continuously need new data to be effective in providing the most relevant ranking of results to those new queries." Thus, the control over these types of data may not in itself give rise to a sustainable competitive advantage.[46] However, those data may be used to improve existing applications or algorithms or to develop new applications or algorithms and those improvements or creations will have more permanent value. In other words, the transient value of data may be capitalised and transformed into more permanent value of applications' improvements or developments.

Other data have more permanent value, such as names, gender, date of birth, address and their depreciation rate is much slower. Therefore, the control of those data gives a more permanent benefit than the control of transient data.

---

43 Commission Decision of 11 March 2008, Case M.4731 *Google/ DoubleClick*, para. 273.

44 N.P. Schepp and A. Wambach, "On Big Data and Its Relevance for Market Power Assessment", *Journal of European Competition Law and Practice* 7, 2016; Sokol and Comerford, *op. cit.*; UK Competition & Markets Authority (2015, para. 3.6).

45 Autorité de la concurrence and Bundeskartellambt, *op. cit*, p. 49.

46 As noted by Competition Commissioner Vestager: "*It might not be easy to build a strong market position using data that quickly goes out of date. So we need to look at the type of data, to see if it stays valuable*": Competition Commissioner Vestager, "Competition in a big data world", DLD 16 Munich, Speech 17 January 2016.

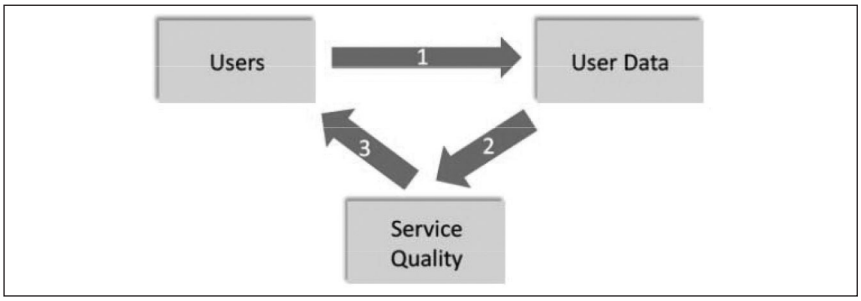
4. Artificial intelligence

Finally, data analytics tools, like any other tools or workers, improve with experience. However, with the development of artificial intelligence and self-learning algorithm, the experience curve may become much steeper as computers can learn some tasks much faster than humans.[47] This may, in some circumstances, increase the first-mover advantage and make the entry of late-comer competitors more difficult. Their entry strategies may then be limited to acquiring the algorithms or investing in different algorithms trying to provide similar services to end-users, albeit in a different manner.

C. Relationships between data collection and analysis

As the different parts of the big data value chain are closely related, they are be some feedback loops between data collection and data analysis which decrease the cost of the former. A first feedback loop, which constitutes a network effect, is linked to the *number of users* as depicted in Figure 1 below and runs as follows: (1) more users means more data; (2) which in turn, means better quality of the service in a general way (on the basis of general indicators such as language, location etc.) as well as in a personalised way (on the basis of the profile that has been built of a specific user); (3) which in turn, attracts even more users to the service.

Figure 1. The user feedback loop



Source: Lerner, 2014, p. 19

[47] OCDE, *op. cit.*, p. 11 notes another reason why the experience curve can be steeper in big data: the lack of physical bounds to the quantity and variety of data that can be collected in a digital world and the unlimited knowledge that can be obtained by running data mining algorithms on a variety of datasets, or using data-fusion.

If this feedback loop takes place, the cost of data collection is higher for a new and small platform than for a larger one.[48] However, the existence of the feedback loop depends on the relationship between the data and the service quality which in turn depends on the type of data and the type of application at hand. Balto and Lane,[49] Lerner,[50] Lambrecht and Trucker[51] or Sokol and Comerford[52] submit that in most cases, the service quality depends only marginally – if at all – of the user data, hence the feedback loop is rarely existent.[53] Moreover, even when the feedback loop exists, if the data collection cost is very small, the effects of the feedback loop in decreasing this cost will be very small as well.

A second feedback loop is linked to the *monetisation possibilities* as depicted in Figure 2 below and runs as follows: (1) more users means more data; (2) which in turn, means better targeting possibilities for advertisement when additional data are necessary to improve the targeting algorithms; (3) which in turn, raises the likelihood that users click on the ads that are displayed to them, hence increases monetisation under the commonly used pay-per-click model; (4) and which attracts more advertisers because of the higher probability that a user buys the advertised product; (5) which in turn raises again the revenues of the provider; (6) those increased possibilities of monetisation increase in turn the possibility of investment to improve the service and attract more users; (7) which also contributes to the increase of advertisers.

48 F.A. Pasquale, "Privacy, Antitrust and Power", *Georges Mason Law Review* 20(4), 2013, 1009-1024; A. Ezrachi and M.E. Stucke, *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy*, Harvard University Press, 2016.

49 D.A. Balto and M.C. Lane, "Monopolizing water in a tsunami: Finding sensible antitrust rules for Big Data", *Competition Policy International*, 2016.

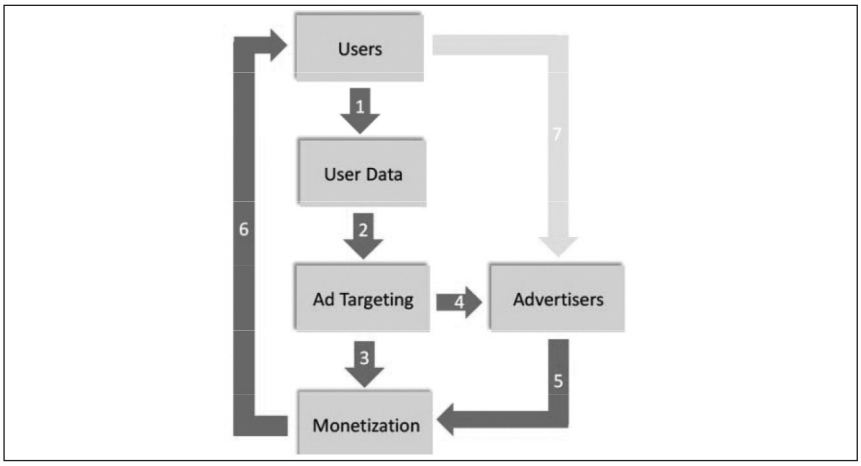
50 A. Lerner, *op. cit.*

51 A. Lambrecht and C. Trucker, *op. cit.*

52 D. Sokol and R. Cromerford; *op. cit.*

53 According to Lambrecht and Trucker, *op. cit.*, p. 11, data being very cheap relative to processing skills suggests that processing skills are more important than data in creating value for a firm.

Figure 2: The monetisation feedback loop



Source: Lerner, 2014, p. 40

If this second feedback loop takes place, a new and small platform with less users and less data will have more difficulty to target its ads and attract advertisers than large platforms, hence less money to improve its products. Here again, the existence and the intensity of this feedback loop needs to be tested on a case-by-case basis and depends on (i) the relationship between the quantity of data and the improvement of ad targeting algorithm, (ii) the relationship between the quality of ad targeting and the attraction of advertisers, and (iii) how the platform invests advertisement revenues and finances service improvement. Lerner[54] and Sokol and Comerford[55] submit that empirical evidence does not show that a higher number of users necessarily leads to a better monetisation. Moreover, even when the feedback loop exists, if the data collection cost is very small, the effects of the feedback loop will be very small as well.

Finally, it is important to note that the development of artificial intelligence may in some circumstances re-inforce those feedback loops as it may strengthen the relationship between user data and service quality in the first loop and the relationship between user data and ad targeting in the second loop.[56] Again, this effect should be assessed on a case-by-case

54 A. Lerner, *op. cit.*, p. 42.
55 D. Sokol and R. Cromerford; *op. cit.*
56 As explained in Domingos, "*the power of machine learning is perhaps best explained by a low tech analogy. For example in farming, we plant the seeds, make sure they have enough water and nutrients, and reap the grown crops. When it comes to artificial intelligence,*

basis and very much depends on the type of data and on the type of self-learning algorithms at hand.

III. Recommendations for competition agencies

To assess the importance of data in determining market power in a big data value chain, I recommend an analytical framework based on three principles and two questions.

The first principle is that *data are one input, which is important but not unique*, to develop successful applications and algorithms. Other inputs are also important such as skilled and creative labour force (in particular computer scientists and engineers), efficient hardware, capital and distribution channels. Above all, the skills and creativity of the labour force will make the success of the applications. The second principle is that big data value chains (data collection, storage and analysis) exhibit *many direct and indirect network effects that need to be captured* by the competition authorities. That requires an understanding and an analysis of each part of the value chain but also the interaction and possible feedback loops between its different parts instead of analysing one part of the value chain in isolation. The third principle is that each big data application and algorithm is different and should be analysed on a *case-by-case basis*. Therefore, it would be inappropriate to propose detailed recommendations at a general level beyond a broad framework for analysis.

With those principles in mind, a competition authority trying to assess market power in the big data value chain should answer two main questions:

The first question relates to the *value of the data* under examination for the applications and the algorithms under examination. That question requires determining: (i) the extent of the economies of scale in the data, in particular what is the marginal benefit of having more data under examination to improve the quality of the application under examination; (ii) the extent of the economies of scope in the data, in particular how important it is to combine different types of data to improve the quality of the application under examination; (iii) the time depreciation value of the data, in

learning algorithms are the seeds, data is the soil, and the learned programs are the grown plants. This means that the more data is available to a learning algorithm, the more it can learn. In short: No data? Nothing to learn. Big data? Lots to learn.": P. Domingos, *The Master Algorithm, The Machine Learning Revolution*, Basic Civitas Books, 2015.

particular the relationship between the age of the data and its relevance to develop or improve the application under examination.

The second question relates to the *availability of the data* under examination for the applications and the algorithms under examination. This question requires determining: (i) the possibility and the costs for an application developer to collect data directly from the users, machines, sensors …; (ii) the possibility and the costs for the application developer to buy the data from data broker and in a data market place.

Such data availability is to a large extent influenced by the legal framework regarding data collection and use. As this framework is different in the EU than in other parts of the world (and within the EU, in some Member States than in others), as well as for firms offering some services (such as traditional telecommunications services) than competitors offering other services (such as the communication app services), those legal differentiations should be factored in the competition analysis.