Data Economy and M&A

Daniela Schoch*

ABSTRACT

In research, public, and policy debate, there is increasing interest in data accumulating firms like Google, Facebook, and Amazon. Discussions about whether the power of firms in terms of personal data concentration might harm consumers are prevalent. Since theory predicts mixed results regarding the impact of such firms' data access on social welfare and oftentimes business models with zero prices for consumers, competition authorities rarely intervene. Meanwhile, due to high scale economies and network externalities, data firms have a particularly large incentive to grow, among others through mergers and acquisitions (M&A). Adding to the up to now mainly theoretical or anecdotal discussion on data intensive firms, this paper (1) identifies data-rich firms in a systematic approach using textual analysis and (2) analyzes the relationship between firms' M&A activity and data intensity. I show that the proposed measure reflects expected characteristics of data intensive firms. I find that higher data intensity of firms corresponds to a higher probability of being acquirer or target in an M&A transaction. Controlling for market to book ratio, life cycle stage, and competition intensity does not alter the relationship. There is some indication for lower attention by competition authorities for data intensive acquirers if the target is small, but even higher attention if the target is large or public.

JEL Classification: G34, G38, L40, L80
Keywords: M&A, data intensive firms, textual analysis
Version: December 11, 2019

* Institute for Finance & Banking, Ludwig Maximilian University Munich, Ludwigstrasse 28 RGB, 80539 Munich, Germany. E-mail: schoch@bwl.lmu.de

1 Introduction

> *The world's most valuable resource is no longer oil, but data – A new commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow.* – The Economist (2017)

In research, public, and policy debate, there is increasing interest in data accumulating firms such as Google, Facebook, and Amazon. Data and particularly the capability and tools to process and exploit information gain more and more importance for businesses (e.g., Lambrecht and Tucker 2015; Hartmann and Henkel 2018). Public policy, and more specifically antitrust regulation, is discussing business models, where the mass of customers pays at least to some extent with personal data instead of money. How competition policy can grasp dynamics in such industries, is an ongoing debate. In research, most discussions are theoretical. For empirical investigations, a definition and measurement of data intensive firms is still missing. In this paper I present a method to measure the data intensity of a firm's business using textual analysis. I then analyze whether firms with higher data intensity are more attractive targets or acquire more companies. The reasons of a possibly higher activity in mergers and acquisitions (M&A) can be manifold, among others too little regulatory intervention by antitrust authorities, particularly high incentives of firms to grow their data stock and data analytics capabilities due to the specific market characteristics, or typical consolidation attempts in a growing and rather young market.

Generally, markets in which 'data' firms operate are characterized by high economies of scale and often network effects (e.g., Campbell, Goldfarb, and Tucker 2015). Network effects appear since many data accumulating firms' business models are based on platforms, i.e., two- (or multi-)sided markets, acting as an intermediary between at least two sides of customers. The attractiveness for using a platform by one side of the market is determined by the size of the other side and vice versa (e.g., Caillaud and Jullien 2003; Rochet and Tirole 2006).

According to Farboodi, Mihet, Philippon, and Veldkamp (2019), data is valuable information and through that an intangible asset that results from a firm's business activity. Companies use this data to improve their offers and internal processes, e.g., in case of two-sided markets to better match the different sides through personalized advertisement or product recommendations.

For such firms, building a large data stock is expensive in the beginning, but gets cheaper with increasing size, up to marginal costs close to zero. This pattern reflects the existence of scale economies and network externalities, which can increase entry barriers and simplify monopolization (e.g., Campbell, Goldfarb, and Tucker 2015). Some markets even have 'winner-take-all' characteristics, suggesting that natural monopolies are a probable result

of competing platforms, driving most firms out of the market (e.g., Correia-da Silva, Jullien, Lefouili, and Pinho 2019; Katz, Shapiro, et al. 1985).

As some big players appear to stand out more and more, the number of antitrust investigations of data companies is also growing. For example, both the Federal Trade Commission (FTC) and the European Commission (EC) were investigating whether Google's search algorithm discriminated against (potential) competitors. Also, investigations of merger and acquisition (M&A) deals of data accumulating firms get increasing attention. As the growing amount of data stock owned by these firms naturally evoke the question how these firms deal with their data, data privacy concerns have been raised (e.g., Casadesus-Masanell and Hervas-Drane 2015). In case of the Google/DoubleClick deal in 2007, both the FTC and the EC stated that they lack the authority to block a merger because of privacy concerns regarding increasing data accumulation resulting from the merger (Federal Trade Commission 2007). Later, e.g., in the Microsoft/LinkedIn deal in 2016, the EC acknowledged that such concerns can be relevant for competition "to the extent that consumers see it as a significant factor of quality, and the merging parties compete with each other on this factor" (European Commission 2016b).

The difficulty of antitrust authorities to evaluate possible harm to consumers due to increasing data concentration as such or due to M&A among others arises from (1) a trade-off between the benefits and costs of disclosing and analyzing customer information and (2) unclear predictions regarding the welfare effects of mergers between platforms.

Benefits from disclosure of private information are in particular lower information asymmetries, which can decrease search costs and increase quality, allocation efficiencies, and product differentiation. However, to assess the impact of growing market power in terms of data accumulation on social welfare, these benefits have to be compared with possible costs of disclosing information, i.e., the value or utility of privacy. The overall outcome of this trade-off therefore highly depends on its particular context (e.g., Acquisti, Taylor, and Wagman 2016; Posner 1981; Shy and Stenbacka 2016; Taylor and Wagman 2014).

Theory also predicts mixed effects of M&A between platforms on social welfare, depending on the extent and shape of network externalities. This holds even if the market is already highly concentrated (e.g., Correia-da Silva, Jullien, Lefouili, and Pinho 2019). Network effects might therefore intensify the two counteractive aspects in the trade-off between improving information efficiency and a possible increase in valuing privacy when personal data gets more concentrated.

These unclear predictions of M&A on competition for data intensive firms might be one reason for which they anticipate less regulatory intervention and thus are expected to invest more in M&A than rather traditional businesses.

Generally, often discussed motives for investing in M&A are among others growth-aiming or resource-driven (e.g., to foster innovation), cost-reducing (e.g., through economies of scale), risk-reducing (e.g., through better diversification of products/businesses), power-increasing (e.g., through higher market shares), or managerial (e.g., for empire building) (e.g., Berk and DeMarzo 2007). Due to the specific characteristics of data intensive firms, the combination of complementary resources, economies of scale, and enhancing market power to increase network effects appear to be particularly important. These should represent a high incentive to further grow by investing among others in M&A.

Finally, the data economy is still rather young and is characterized by high dynamics regarding firm entry, exit, and consolidation. This can simply be reflected in higher M&A activity.

In summary, data intensive firms and their markets are characterized by economies of scale, particularities of competition in two-sided markets as many data businesses are platform based, and growing privacy concerns by their consumers, especially as soon as they gain some market power. However, to the best of my knowledge there is no systematic approach to measure the data-richness of firms or categorize companies into data and non-data firms.

I fill this gap by systematically identifying and evaluating data intensive firms using textual analysis. More specifically, I combine (1) the product similarity measure of Hoberg and Phillips (2016) that is originally used to construct firm-specific network industries using (1a) 10-K business descriptions and (1b) 10-K risk factors as well as (2) a keyword-search approach. Keywords are extracted from (2a) company descriptions provided by competition authorities and (2b) 2016 risk factors in 10-K's of five case firms, namely Google, Facebook, Microsoft, LinkedIn, and CoreLogic. These five firms were subject to privacy concerns due to an increasing data concentration in investigations of M&A deals or joint ventures by the FTC, the EC, or both. Additionally, as the data intensity measure is likely to have a high power but at the same time a high type I error, I use a manual classification of a randomly selected subsample to construct a data firm dummy variable that is alternatively used in the analyses. Critical for this manual categorization of data firms is the characteristic of likely arising privacy concerns in case of a hypothetical merger or acquisition activity. This characteristic is chosen to match the selection of the five antitrust cases that define the benchmark sample in the data intensity measure.

Further, due to the unclear consequences of M&A by data intensive firms and the associated difficulties of antitrust authorities in evaluating them as well as the particularly high expected incentive of data intensive firms to invest in M&A, I analyze whether data intensive firms are more likely to be an acquirer or target in an M&A deal.

I use data on public U.S. firms for the fiscal years 2006 to 2017. The sample starts with the fiscal year 2006, as discussing risk factors in 10-K's has become mandatory since the end

of 2005, except for small reporting companies. 10-K's are retrieved from Loughran and McDonald (2019). M&A data are from Thomson Reuters SDC (Eikon) and accounting information from Compustat.

I find that the data intensity measure reflects expected characteristics. For example, firms with higher data intensity are predominantly in the computer software industry and have higher market to book ratios than firms with lower data intensity scores. Furthermore, a higher data intensity of firms corresponds to a higher probability of being acquirer or target in an M&A transaction. Some evidence suggests that data intensive firms are less regulated when they buy small targets, but get even more attention by competition authorities when they buy large, particularly public firms. This would match a pattern of pre-emtive merger activity, hence buying firms when they are still small and before they are on the radar of authorities and public. However, as only limited data on these public-private deals are available, this result must be evaluated with caution.

Controlling for competition intensity in the firms' markets and the firms' life cycle stages does not alter the effect of the data intensity measure. Industry relatedness of public transaction partners measured by (1) the pairwise business description similarity, (2) the occurrence in each other's industry network as measured by Hoberg and Phillips (2016), and (3) the occurrence in each other's primary SIC industry does not clearly depict a specific pattern. Additionally, for public targets, the deal value to sales multiples are not different for data intensive firms. Also, as combined M&A announcement returns are not different for data firms, data firms do not seem to overpay if they are acquirers and are not overpaid if they are targets compared to non-data firms.

The following Section 2 describes the methods and data used for identifying data intensive firms as well as for measuring whether data intensive firms are more involved in M&A activities. Section 3 reports the results of the analyses. Section 4 concludes.

2 Data and Method

2.1 Sample

The sample covers public U.S. firms for the fiscal years 2006 to 2017 with 10-K's published between July 2006 and June 2018. Pre-processed 10-K files are from Loughran and McDonald (2019). 10-K's are highly standardized and their product descriptions and risk factors are required to represent each firm's current fiscal year business activities and risks that might impact the firm's business activities or income in the future. This makes company business descriptions and risk factors easily comparable both, for one company over time and between companies.

Completed and withdrawn M&A deals are retrieved from Thomson Reuters SDC (Eikon) and accounting information from Compustat. The sample is restricted to firms that appear in both Compustat and CRSP, are listed at the NYSE, NASDAQ, or American Stock Exchange, and that have valid accounting values for the later analyses.

Since risk factors (item 1A's) are not required for small reporting companies (SRCs), they are excluded from the analyses. Smaller reporting companies have either a lower market value than $75 million or lower revenues than $50 million and no public float.[1]

Overall 38,652 firm-year observations match a 10-K filing that is published less than 365 days after the previous fiscal year end date. This corresponds to 5,607 unique firms. See Appendix A.1 for an overview of the filter criteria.

2.2 Data firm measure

2.2.1 Data intensity variable

For this study, the most crucial point is to systematically identify data intensive firms. I use textual analysis to determine the extent to which a firm is a data firm.

When thinking of data intensive firms, typically companies like Google, Facebook, and Amazon come to mind. Certainly, most businesses to some extent collect information on their customers. This, however, does not make them directly a data intensive firm. More specifically, companies are considered as data intensive firms, when their business models are based to a substantial part on collecting (and analyzing) their customers' or users' personal information. Using textual analysis, I identify firms that share specific characteristics with firms that were considered as data accumulating firms by antitrust authorities.

The data intensity measure combines the business similarity measure by Hoberg and Phillips (2016) with a keyword based approach, which partly borrows from the method used by Gentzkow and Shapiro (2010).

The following steps explain the detailed procedure for identifying keywords specific for business descriptions of data intensive firms:

The first step is to identify competition cases which were (a) analyzed by antitrust authorities and (b) subject to privacy-concerns due to data accumulation. These firms are then defined as the starting sample of data intensive firms. These cases are (sources in brackets):

[1] https://www.sec.gov/corpfin/amendments-smaller-reporting-company-definition.

- Google/DoubleClick acquisition 2007 (Federal Trade Commission 2007 & European Commission 2008)
- CoreLogic/DataQuick acquisition 2013 (Federal Trade Commission 2014b)
- Facebook/WhatsApp acquisition 2014 (Federal Trade Commission 2014a & European Commission 2014)
- Microsoft/LinkedIn acquisition 2016 (European Commission 2016b)
- Sanofi/Google/DMI joint venture 2016 (European Commission 2016a)

The second step is to extract relevant keywords in these firms' business descriptions as described by the regulatory authorities. See Appendix A.2 for the case texts of the public firms in the list (Google, CoreLogic, Facebook, Microsoft, and Linkedin) and the identified keywords.

Finally, after searching for these keywords (and alternations/synonyms) in all firms' 10-K business descriptions in the overall sample, those keywords that occur more often in the business descriptions of the five case firms than in the overall sample in at least nine of 12 sample years are considered as 'informative' . Only the informative keywords then serve to identify data intensive firms. Examples for informative keywords are *advertising*, *computer*, *user*, *data*, *internet*, *network*, *platform*, and *search*. See Appendix A.3 for the list of keywords and their informativeness.

As firms like Google or Facebook write about possible harm on their business procedures and income through higher regulation of data privacy (e.g., by the European General Data Protection Regulation (GDPR)) in *10-K risk factors* (Item 1A), I additionally analyze risk factor sections. To find key terms related to data protection regulation and data breaches in risk factors, I use the 2016 10-K risk factors of the five case firms identified above and apply the same steps as for the business descriptions. 2016 is the last available year for LinkedIn, the only public target in the cases used. Informative keywords for risk factors are e.g., *data protection*, *laws regarding privacy*, *personal data*, *security breach*, and *use data*. See Appendix A.4 for a summary of keywords for analyzing risk factors.

The similarity measure by Hoberg and Phillips (2016) first requires the calculation of the occurrence of each noun in all texts of each fiscal year. Nouns that appear in less than 25% of the texts within one year are then used to estimate cosine similarities. For each text, a dummy vector P_i indicates whether a specific noun occurs (1) or not (0), e.g.,

$$P_i$$

$$\begin{matrix} noun1 \\ noun2 \\ ... \\ nounN \end{matrix} \begin{pmatrix} 0 \\ 1 \\ ... \\ 1 \end{pmatrix}.$$

The following formula shows the calculation of each pair of firms' cosine similarity, i.e., the similarity between firm i and firm j:

$$similarity_{ij} = P_i^{'} P_j (P_i^{'} P_i)^{-0.5} (P_j^{'} P_j)^{-0.5} \tag{1}$$

The similarity measure covers first, as in Hoberg and Phillips (2016), *10-K business descriptions* (Item 1) to estimate the relatedness between firm's business areas. As in Hoberg and Phillips (2016), the sample includes business descriptions only at a length of at least 1,000 characters. Second, to be included in the similarity measure using *10-K risk factors*, Item 1A texts should exceed 200 characters. The items are automatically extracted from the Loughran-McDonald Stage One 10-K files for approximately 97% of all files.

When applying the cosine similarity measure to keywords occurrence, one specificity needs to be taken into account: $similarity_{ij}$ is set to zero, if none of the informative keywords occurs in a text, as otherwise dividing by zero would make the calculation unfeasible.

The data intensity measure is the standardized sum of the four cosine similarities, i.e.,

- the mean similarity of their *10-K business descriptions* to the five case firms[2];
- the mean similarity of informative keywords in *10-K business descriptions* to the five case firms;
- the mean similarity of their *10-K risk factors* to the five case firms;
- the mean similarity of informative keywords in *10-K risk factors* to the five case firms.

The combination of cosine similarity of both, data related keywords in business/risk factors and overall texts, is used to increase the weight of data specific aspects. This procedure accounts for the fact that many firms, including the prominent case firms above, operate in different businesses, some more, some less related to data. Only using the similarity between business descriptions and risk factors might otherwise distort the measurement of the data intensity variable.

[2] Since some case firms are not in the sample for the whole observation period due to missing accounting information or their public status, the data intensity results from the mean similarity to three to five firms. While Google, CoreLogic, and Microsoft are in the sample every fiscal year, Facebook is only included in 2014, 2015, and 2017 and LinkedIn in $2011 - 2014$.

Since the similarity to a firm itself is one, the mean similarity of one of the case firms is the mean of the similarity to the other case firms.

A robustness check does not use the simple occurrence (dummy vector) of words for measuring similarity, but the frequency of noun or keyword occurrence divided by item length (relative frequency).

2.2.2 Type I error and the construction of a binary data firm variable

Due to the construction of the data intensity variable, data intensive firms are quite likely to score high. However, e.g., when counting data protection related words in risk factors, firms might also simply discuss the protection of internal data, such as employee information or trade secrets. Another example is the generally increasing public discussion of data security and regulation over the years, which might be reflected in many firm's 10-K's more or less independent of their particular businesses. In such a case, the data intensity measure necessarily captures aspects of data intensive firms, which might not be entirely related to the kind of data intensity that should be measured here. This means that, while the type II error might be low, the type I error is likely to be rather high. Hence, the measure might attribute high data intensity levels to firms, which would not be considered as data firms.

To account for this, I manually classify 1% randomly chosen firm-fiscal year observations, i.e., 400 10-K files, into data firms and non-data firms.

10-K's are categorized into data firms, if they fulfil the following criteria, which mainly fit the selection criteria of the previously used antitrust cases:

- The privacy of the firm's customer data must be a crucial aspect of their business.
- A hypothetical acquisition or merger of the company is likely to raise privacy concerns due to an increasing concentration of data within a combined firm.
- The company must be able to sell personal identifiable information or use personal identifiable information to earn money from third parties (e.g., through selling personalized advertisement).

These criteria reflect that firms which collect sensitive information, but are subject to particular data privacy regulation that prohibits using the information besides for internal processes, are not considered data firms here. Examples for such data are account transactions by bank customers or personal data of study participants in clinical approval stages. Also, firms that collect and sell non-personal data, e.g., weather or company data, firms that are pure B2B companies that facilitate data analytics and management, or companies that simply operate standard online shops or home delivery services and due

to this collect payment and address information of their customers, are not categorized as data firms.

The manual classification results in 21 (5.25%) data firms and 379 non-data firms.

In a second, more loose definition of data firms, I further classify firms as data firms that are specialised on data analytics and storage, hence providing critical capabilities for firms to actually use their data stock. Using this more loose definition, 45 (11.25%) 10-K's are categorized as data firms.

Then, a k-nearest neighbor (KNN) algorithm sorts the rest of the observations into data and non-data firms. KNN shows the highest validation accuracy (95%) among several tested algorithms. The predictors for the KNN classification are the cosine similarities of all files in the overall sample to the manually classified firms based on the vocabulary used only in these manually classified 10-K's. Again, only those words are used that appear in less than 25% of the files. Additionally, a robustness check applies a naïve Bayes classification (88% accuracy) combined with the data intensity measure.

2.3 Further variables

As in Rhodes-Kropf, Robinson, and Viswanathan (2005), who study company valuation waves and merger activity, the dependent variable represents a dummy that equals one if a firm was a transaction party in an M&A deal within the 365 days after the fiscal year end date, and zero otherwise.

The main control variable is the market to book ratio, which equals market assets over book assets. Market assets is market value of equity (fiscal year end stock price $\cdot$ current shares outstanding) + book assets − book value of equity − deferred taxes. Additionally, dummy variables indicate the benchmark firms used in the different specifications of the data firm variable, i.e., $Case firm$, $ManualData firmDummy$, and $ManualNonData firmDummy$. Furthermore, fiscal year and Fama French 49 industry fixed effects are included.

For more in-depth analyses firm age, proxied by the current fiscal year − first fiscal year of a firm occurring in Compustat, and a firm's life cycle stage are included. Dickinson (2011)'s life cycle measure attribute the stages *Introduction*, *Growth*, *Mature*, *Shake-Out*, and *Decline* to firm-fiscal year observations according to the signs of cash flows from operating, financing, and investing activities.

Three measures based on Hoberg and Phillips (2016)'s industry networks account for competition in each firm's industry, i.e., the product similarity measure (mean similarity of each firm's 10 most similar companies), the sales-based Herfindahl-Hirshman Index (HHI) within each firm's industry network, and the absolute number of firms within each firm's industry network.

2.4 Regression analysis

A logistic regression using random effects and robust standard errors estimates the probability $Pr_{i,j}$ of being involved as a transaction party in an M&A deal. The baseline regression is represented by:

$$ln\left(\frac{Pr_{i,t}(merged)}{Pr_{i,t}(nonmerged)}\right) = \beta_0 + \beta_1 Datafirm_{i,t} + \beta_2 lnMarketToBook_{i,t} + \beta_3 Casefirm_i + \sum_{t=1}^{T} \beta_{t+3} Fiscalyear_t + \sum_{k=1}^{K} \beta_{k+16} Industry_{i,t,k}. \quad (2)$$

Furthermore, a multinominal logistic regression accounts for the different possibilities of being involved in a merger or acquisition, i.e., as an acquirer or target. This can give further insights, since some control variables are expected to have different signs dependent on whether the outcome variable is to be either an acquirer or target.

The following section contains an overview and verification of the data intensity measure, the results of the binary and multinominal logistic regressions, as well as an analysis of M&A announcement returns.

3 Results

3.1 Descriptive statistics

3.1.1 Data firm measure

The data intensity scores result from adding up the four different similarity scores. The score is then standardized by the highest occurring data intensity score in the sample to yield a maximum value of 1. Table 1 shows descriptive statistics on the four similarity scores and the overall data intensity score. The standardized data intensity scores range from 0.15 to 1 with the highest density between 0.4 and 0.6, as can be observed in figure 1 depicting the density histogram of $DataIntensity$.

Adobe Inc has the highest data intensity score. Other prominent companies among the top 20 highest data intensity scores are Yahoo, Tripadvisor, and Zynga. The most represented Fama French 49 industries among the top 100 firms are Computer Software with 80 firms and Business Services with 8 firms.

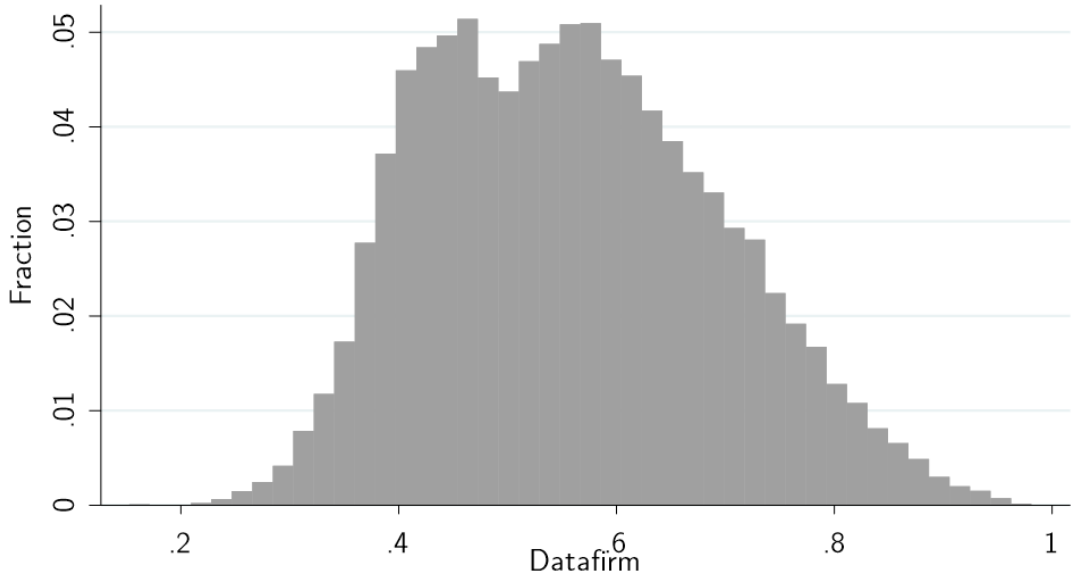
J.W. Mays Inc., a real estate company, has the lowest data intensity score. Financial firms represent about a third of the 100 firms with the lowest data intensity scores, with 19 firms in Trading, 7 in Real Estate, and 5 in Banking. Further frequently appearing

Table 1: Cosine similarities

The table reports the four mean cosine similarities and the overall standardized data intensity scores. Business similarity corresponds to the cosine similarity of Item 1 and risk factor similarity to the cosine similarity of Item 1A in 10-K's to five distinct case firms: Google, Facebook, Microsoft, LinkedIn, and CoreLogic. Keyword similarity covers the mean cosine similarity to the five case firms based on selected keywords, instead of the whole body of nouns. Data intensity score is the sum of the four similarities, divided by the maximum data intensity score in the sample. *Corr* is the correlation between each of the four similarity scores with the overall data intensity score.

	obs	**mean**	**median**	**s.d.**	**min**	**max**	**corr**
Business Similarity	38,686	0.10	0.09	0.04	0.00	0.27	0.50
Risk Factor Similarity	38,686	0.10	0.10	0.04	0.00	0.30	0.63
Business Keyword Similarity	38,686	0.74	0.75	0.09	0.26	0.94	0.69
Risk Factor Keyword Similarity	38,686	0.24	0.27	0.20	0.00	0.73	0.88
Data Intensity Score	38,686	0.56	0.55	0.13	0.15	1.00	1.00

Figure 1: Density histogram of data intensity scores

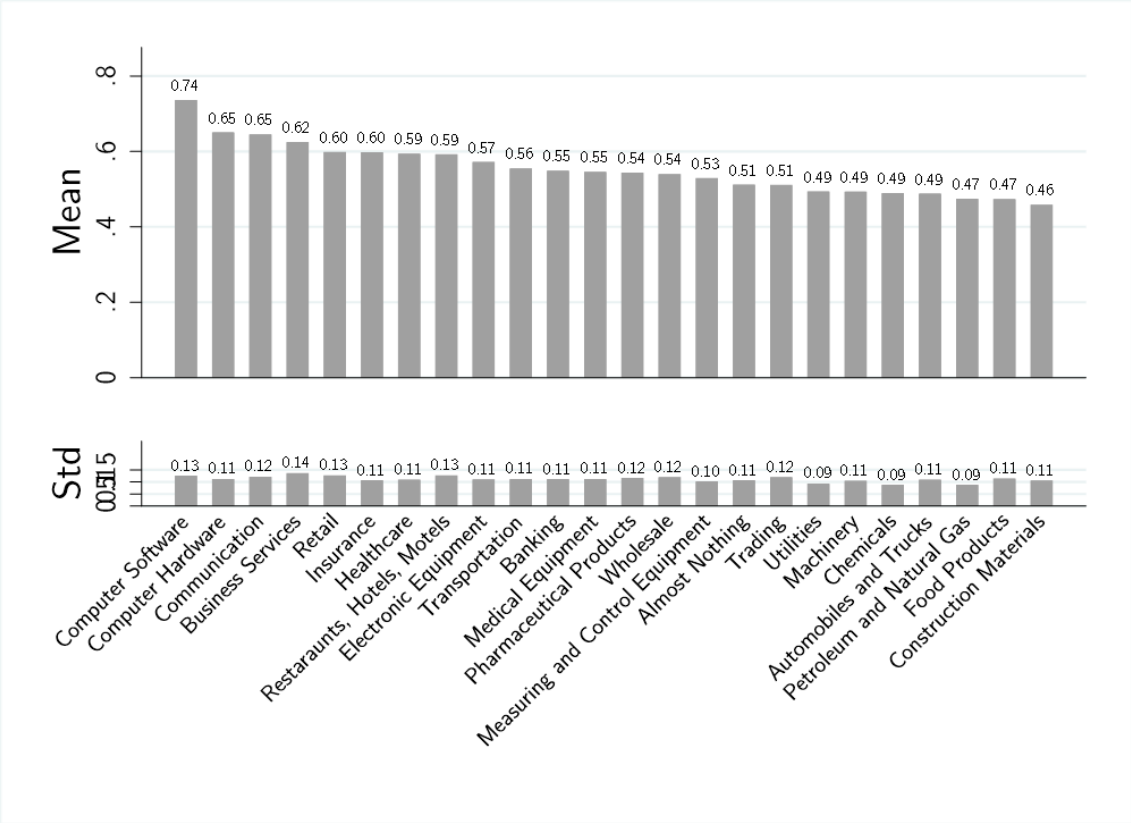


The figure shows the density histogram of the data intensity scores for all firm-fiscal year observations in the sample. The maximum value is standardized to one. Number of observations: 38,686.

industries are Petroleum & Natural Gas (10), Utilities (6), Construction Materials (5), Food Products (5), and Precious Metals (5).

Furthermore, mean data intensity scores in the largest Fama French 49 industries (industries with $\geq$500 firms) are shown in figure 2. The most data intensive industry is Computer Software, the least intensive Construction Materials. This also matches natural expectations and indicates that the measure reflects data intensity of firms. Additionally,

Figure 2: Mean and standard deviation of data intensity score per Fama French 49 industry (industries ≥500 firms)



The figure shows the mean and standard deviation of data intensity scores for firms in the largest Fama French 49 industries. Industries are included if they contain at least 500 firm-fiscal year observations. Number of observations: 31,398.

standard deviations of data firm intensity are very similar within each industry, indicating that the data intensity measure does not only reflect standard industry categorizations.

The five case firms appear in the 7^{th} decile in the fiscal years $2006 - 2007$, in the 8^{th} decile in the fiscal years $2008 - 2009$, afterwards they are all in the 10^{th} decile.

Amazon, which would be probably thought of as being a data firm, is mostly in the 10^{th} decile (besides in the fiscal years 2008 (8^{th}) and 2016 (9^{th})) with an overall mean data intensity score of 0.76. EBay is in the 10^{th} decile with a mean data score of 0.79.

In April 2019 Deutsche Bank issued a Cloud and Big Data Index Certificate[3], covering 20 firms including 13 U.S. based companies. Nine[4] of them appear in the final sample of this study with overall 61 firm-fiscal year observations and a mean data intensity score of 0.83. One of the 61 observations is in the 8^{th}, 6 in the 9^{th}, and 54 in the 10^{th} decile. This

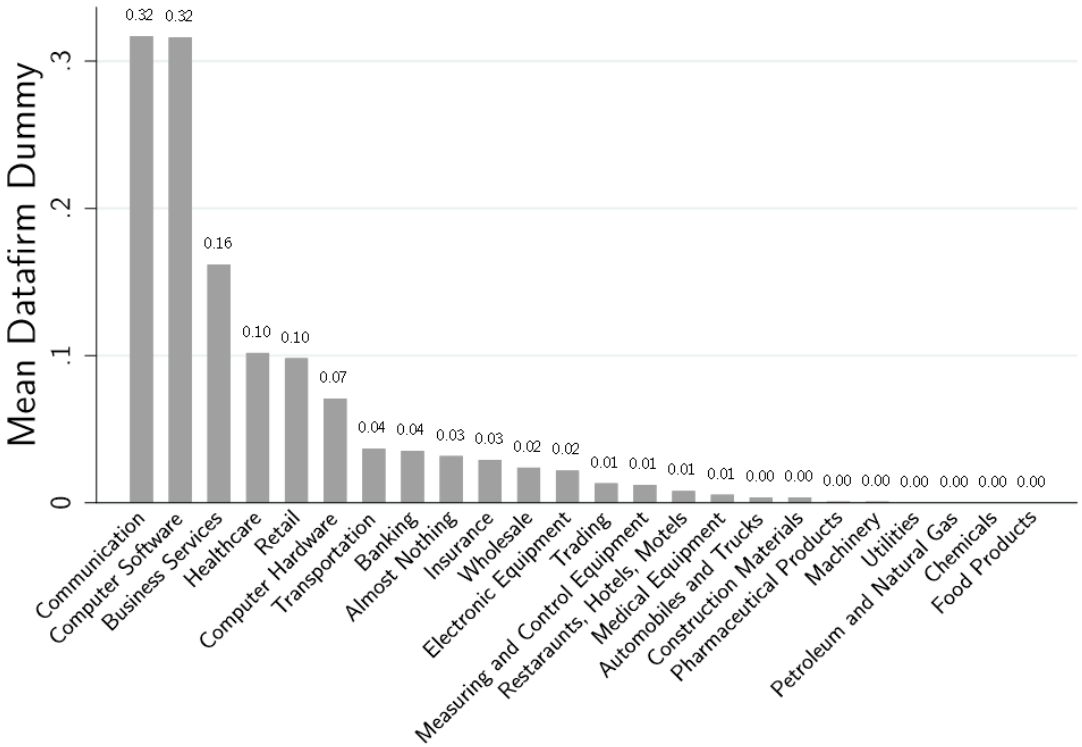
[3] https://www.xmarkets.db.com/DE/EN/KID/DE000DC8CLD2.

[4] These companies are Citrix, Paycom Software, Salesforce.com, ServiceNow, Splunk, Tableau Software, Ultimate Software, VMware, and Workday.

result further indicates that the data intensity measure proposed here reflects expected characteristics of data intensive firms.

The additional construction of the data firm dummy as explained in section 2.2.2 should account for a possibly inflated data intensity measure. The KNN classification results in 2,260 (5.85%) data firm observations. Using this dummy shows a similar ordering of Fama French 49 industries as before, but a much more conservative picture if interpreting the mean data firm intensity scores or the share of observations turning to one in the data firm dummy as the probability of being a data firm in a specific industry (see figure 3).

Figure 3: Share of data firms (mean data firm dummy) per Fama French 49 industry (industries ≥500 firms)



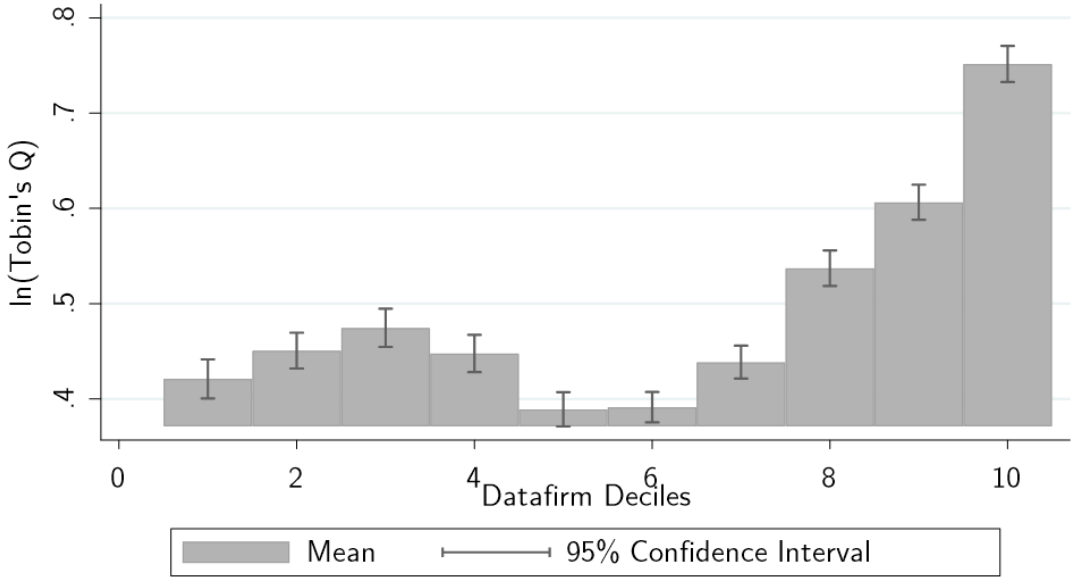
The figure shows the mean of the strict specification of the data firm dummy for the largest Fama French 49 industries. Industries are included if they contain at least 500 firm-fiscal year observations. Number of observations: 31,398.

In this binary classification three of the five case firms, namely LinkedIn, Google, and Facebook, are always classified as data firms. However, Corelogic is classified as a data firm in one out of twelve fiscal years, while Microsoft is categorized not at all as a data firm. As Microsoft's business is highly diverse, it might not be surprising that other firms are more similar to Microsoft than other data firms. Therefore, an additional more loose definition of the data firm dummy includes firms that deliver capabilities and tools to

analyze and store data. Then, 4,174 (10.80%) observations are classified as data firms and Corelogic appears as a data firm in two and Microsoft in ten fiscal years.[5]

Since data firms are said to be highly valued without many tangible assets, market to book ratios should be higher for higher data intensity scores. Figure 4 supports this expectation, depicting the mean market to book ratios for *DataIntensity* deciles. This holds also when using the data firm dummy, as shown in table 9 in appendix A.5.

Figure 4: Mean market to book ratio per data intensity score decile



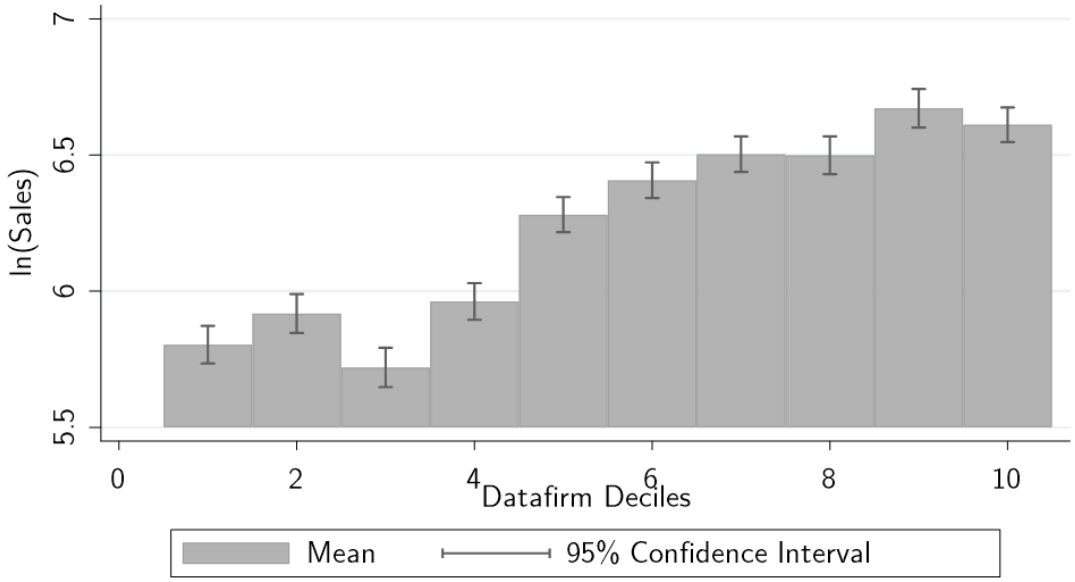
The figure shows the mean of the natural logarithms of market to book ratios within each data intensity score decile including 95% confidence intervals. Number of observations: 38,583.

Additionally, famous data firms have been struggling often to meet revenue expectations. The following figure 5 does not suggest this pattern (also see table 9 in appendix A.5 for the data firm dummy). A more detailed look into the data however reveals that this result holds only for the second half of the sample period. In the first half of the sample period, the relationship appears inverse U-shaped.

Furthermore, data firms are expected to be comparably younger than more traditional firms with a lower data intensity. Figure 6 supports this expectation, showing a mean age of approximately 28 (median = 24) in the first decile and a mean of 16 (median = 14) in the 10^{th} decile. Table 9 in appendix A.5 shows a similar result for the data firm dummy.

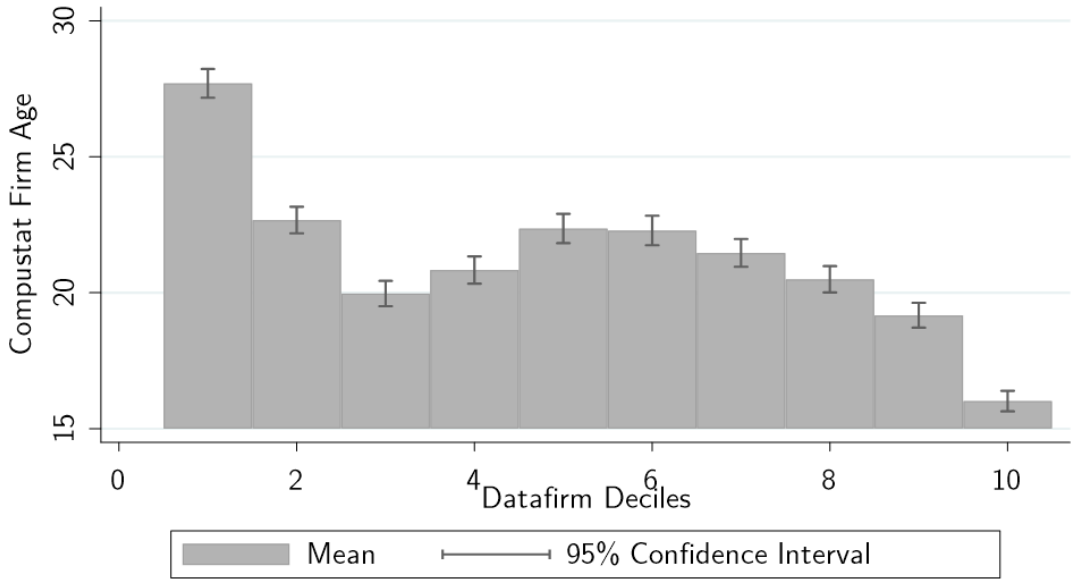
[5] For robustness checks, a further specification applies a naïve Bayes algorithm. See section 3.2 for more details.

Figure 5: Mean sales per data intensity score decile



The figure shows the mean of ln(sales) within each data intensity score decile including 95% confidence intervals. Number of observations: 36,876.

Figure 6: Compustat firm age per data intensity score decile



The figure shows the mean of Compustat age, i.e., the number of years listed in Compustat, within each data intensity score decile including 95% confidence intervals. Number of observations: 38,652.

3.1.2 Further firm characteristics

The sample comprises overall 38,649 observations with mean market assets of 11.8 billion U.S. Dollars (see table 9 in appendix A.5 for more details). Data firms report on average

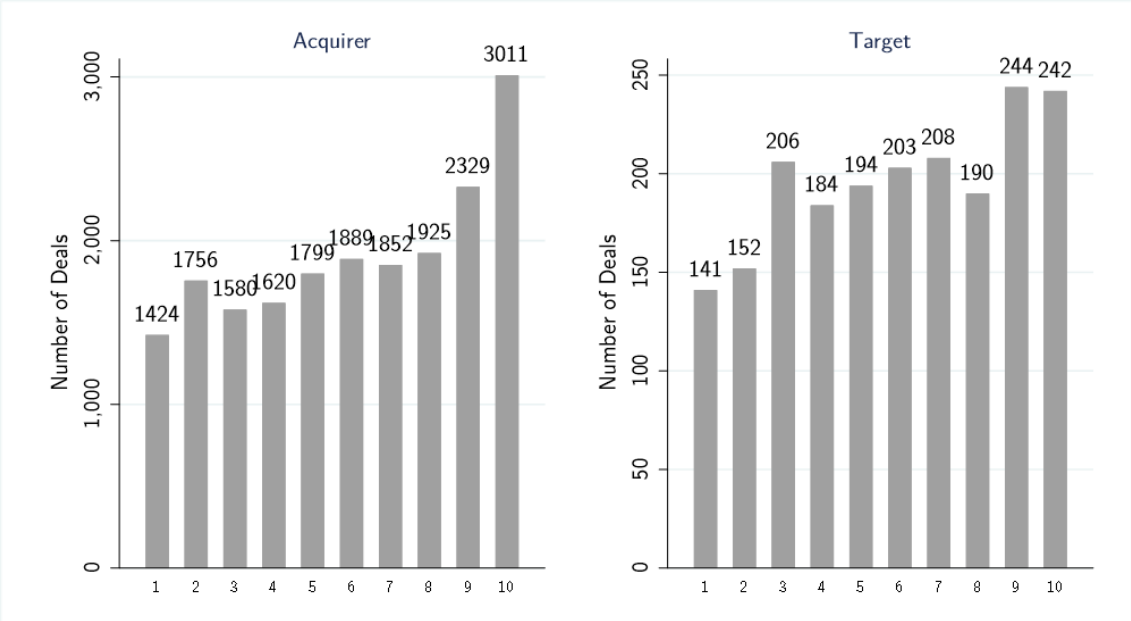
significantly higher market assets, market values of equity, net income, long term debt, and market to book ratios as well as significantly lower plant, property, and equipment, firm age, and sales per employee.

Over 12,000 firm-fiscal year observations report at least one merger within the next 12 months. Firms that invest in mergers seem to be overall larger than non-merged firms. Among merged firms, targets are overall smaller and less successful than acquirers.

3.1.3 M&A of data firms

The sample covers overall 21,148 deals by 3,282 unique acquirers and 1,741 targets. Figure 7 already gives an indication on M&A activity by data intensive firms, depicting an increasing number of deals for higher data intensity score deciles.

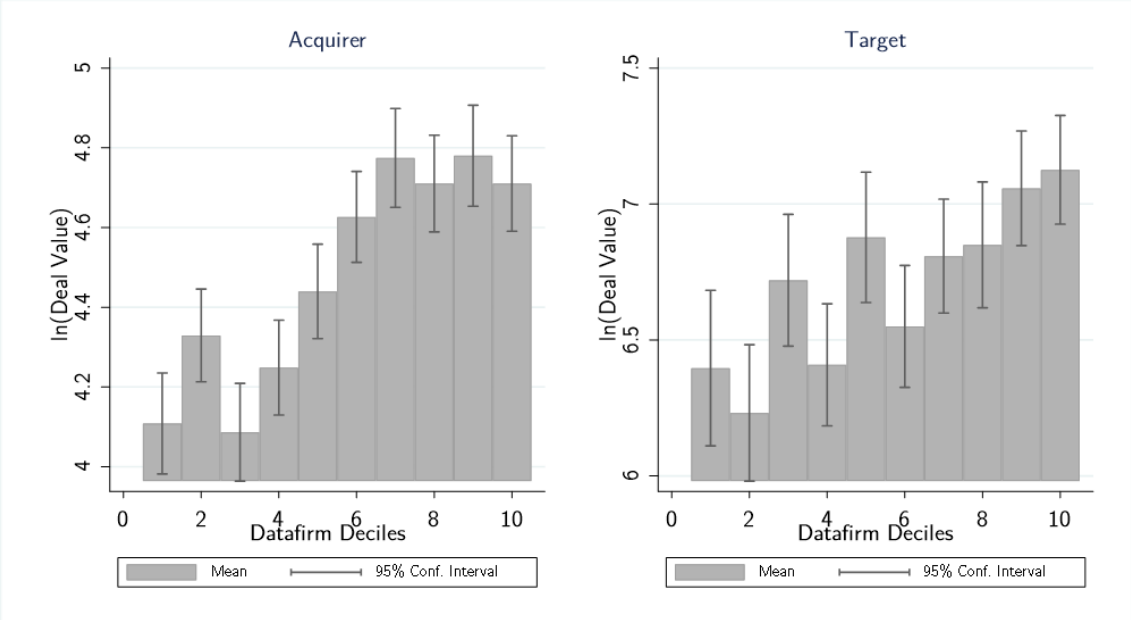
Figure 7: Number of deals per data intensity decile



The figure shows the absolute number of deals within each data intensity score decile, separately for acquirers and targets. Number of observations: acquirer: 19,185; target: 1,964.

Deal values are overall increasing over the data intensity score deciles for both acquirers and targets (see figure 8). The confidence intervals additionally indicate that the differences between the highest to lowest deciles are statistically significant. Especially for targets this result is not surprising as it matches overall increasing market to book ratios for higher data intensity scores.

Figure 8: Deal value per data intensity decile



The figure shows the mean of ln(deal values) within each data intensity score decile, separately for acquirers and targets. Number of observations: acquirer: 9,861; target: 1,895.

3.2 Logistic regression

Table 2 shows the results of the logit regressions. Firms with higher data intensity scores appear to increase the probability of being a transaction partner in an M&A deal (see column (2)). More specifically, an increase in the data intensity score by 0.1 (recall that $DataIntensity$ ranged between 0.15 and 1) increases the probability of being a transaction partner by approximately two percentage points. This result holds also, if the more loose definition of the data firm dummy variable is used (see column (3)). Then, the probability of becoming a transaction partner is on average 2.7 percentage points higher for data firms than for non-data firms. The data firm dummy in the more strict specification is however not significant. Using a specification with three categories, i.e., (1) never data firms, (2) strict definition data firms, and (3) those companies that become a data firm when using the loose definition, reveals that the result of the more loose specification is driven by those observations that are data firms only in the more loose specification, hence by data analytics or storage companies and not by data accumulating firms.

Results are robust to using the relative frequency of terms instead of simple occurrence (1/0) for the continuous data intensity measure, to excluding those firms that switch[6] their data firm dummy status more than twice during the observation period, and to using the predicted probabilities for being a data firm resulting from the KNN algorithm.

[6] There are 1,111 observations by switching firms using the strict data firm dummy specification and 1,983 observations using the more loose definition.

Table 2: Logit regression results

The table reports the marginal effects of a logit regression using random effects and robust standard errors. The dependent variable is $Merged$, turning to 1 if a firm was a transaction partner in the following fiscal year. Model (1) reports the baseline regression; Model (2) includes the continuous data intensity variable; Model (3) uses a dummy variable for data firms as explained in section 2.2.2; Model (4) uses an alternative, less restrictive, definition of the data firm dummy variable than Model (3). $Casefirm$ is a dummy variable, indicating the firms Google, Facebook, Microsoft, LinkedIn, and CoreLogic, which were the reference firms for measuring the cosine similarities for the data intensity scores. $ManualDatafirmDummy$ and $ManualNonDatafirmDummy$ represent those firms that were manually clustered into data and non-data firms. Z scores are reported in parentheses. Significance levels are indicated by: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Dependent Variable: Merged (0/1)	**(1)**	**(2)**	**(3)**	**(4)**
DataIntensity		0.197***		
		[5.996]		
DatafirmDummyStrictDef			-0.001	
			[-0.076]	
DatafirmDummyLooseDef				0.027**
				[2.604]
lnMarketToBook	0.024***	0.021**	0.023***	0.022**
	[3.398]	[2.999]	[3.338]	[3.207]
Casefirm		0.409***	0.420***	0.415***
		[3.564]	[3.702]	[3.709]
ManualDatafirmDummy			0.017	0.010
			[0.205]	[0.122]
ManualNonDatafirmDummy			-0.004	-0.004
			[-0.182]	[-0.159]
Constant	yes	yes	yes	yes
Year/Industry Dummies	yes	yes	yes	yes
# Observations	36,754	36,754	36,754	36,754
# Firms	5,387	5,387	5,387	5,387

A positive relation to merger activity shows also another robustness check using a naïve Bayes algorithm for the two specifications of the dummy variable. Then, approximately one fourth of all observations are categorized as data firms. As this reflects a much higher share of data firms than the manual classification of the randomly chosen subsample would suggest, the dummy only turns to one for data intensity scores in the top 30% quantile. Then, 16.4% (17.1%) categorize as data firms using the strict (loose) definition.

3.3 Multinominal logit regression

Furthermore, I apply a multinominal logistic regression using $TransactionParty$ as a dependent variable to distinguish not only between not merged / merged, but also within the category merged between *acquirer* and *target*. For this analysis I exclude the 300 firm-fiscal year observations in which a firm was both an acquirer and target within the following 12 months after fiscal year end.

Columns (1)-(3) in table 3 show the results of the multinominal logit regression using the same independent variables as in the binary logit model. The results for the data intensity and data firm dummy variables match the previous ones.

The market to book ratio shows different signs for acquirers and targets, indicating that firms with higher Tobin's Q are significantly more likely to acquire firms and significantly less likely to be acquired.

There are several further determinants of whether a firm becomes an acquirer or target. Firm age should account for the expectation that firms are more likely acquired in younger ages. Similarly, firms are expected to be more mature when investing in M&A, which should cover a firm's life cycle stage. The measure introduced by Dickinson (2011) distinguishes between the five life cycle stages *Introduction*, *Growth*, *Mature*, *Shake-Out*, and *Decline* and is based on the signs of a firm's cash flows from operating, financing, and investing activities.

The results are reported in Models (4)-(6) in table 3 and are similar to the reduced model. As expected, older firms are significantly more likely to acquire and younger firms are more likely to be acquired. Furthermore, there is no significant difference for firms in different life cycle stages to become a target, while firms in the *Growth*, *Mature*, and *Shake-Out* phase are more likely to acquire than firms in the *Introduction* phase.

Table 3: Multinominal logit regression results

The table reports the marginal effects of a multinominal logit regression using random effects and robust standard errors. The dependent variable is $TransactionParty$, which equals 1 for non-transaction parties, 2 if a firm was an acquirer and 3 if a firm was a target in the following fiscal year. Model (1) reports the baseline regression including the continuous data intensity variable; Models (2) and (3) use the two different dummy variable specifications for data firms; Models (4)-(6) add the further control variables firm age and life cycle stage. The life cycle stages report differences to the base category *Introduction* and are constructed based on Dickinson (2011). $BenchmarkDum$ corresponds to the dummy variables $Casefirm$, $ManualDatafirmDummy$ and $ManualNonDatafirmDummy$, representing those firms that were used as benchmark firms for the data intensity and data dummy variable construction. Z scores are reported in parentheses. Significance levels are indicated by: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

DV: TransactionParty	*Acquirer*						*Target*					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
DataIntensity	0.164***			0.191***			0.036**			0.027*		
	[5.290]			[6.092]			[2.910]			[2.226]		
DatafirmDummy		-0.000			0.004			0.002			0.000	
StrictDef		[-0.035]			[0.294]			[0.327]			[0.099]	
DatafirmDummy-			0.018*			0.022*			0.011**			0.010**
LooseDef			[2.009]			[2.391]			[2.871]			[2.635]
lnMarketToBook	0.046***	0.048***	0.047***	0.047***	0.049***	0.049***	-0.019***	-0.019***	-0.019***	-0.019***	-0.018***	-0.018***
	[6.973]	[7.223]	[7.134]	[6.999]	[7.260]	[7.167]	[-7.523]	[-7.292]	[-7.438]	[-7.547]	[-7.387]	[-7.521]
Firm Age				0.002***	0.002***	0.002***				-0.000**	-0.000***	-0.000***
				[7.422]	[6.669]	[6.756]				[-4.517]	[-4.893]	[-4.766]
Life Cycle Stage												
Growth				0.097***	0.098***	0.098***				-0.002	-0.002	-0.002
				[10.031]	[10.116]	[10.106]				[-0.387]	[-0.352]	[-0.339]
Mature				0.099***	0.100***	0.100***				0.000	0.000	0.001
				[10.166]	[10.237]	[10.236]				[0.038]	[0.069]	[0.111]
Shake-Out				0.043***	0.043***	0.043***				0.005	0.005	0.005
				[4.058]	[4.095]	[4.089]				[0.910]	[0.911]	[0.913]
Decline				-0.008	-0.008	-0.008				0.001	0.001	0.001
				[-0.697]	[-0.709]	[-0.704]				[0.221]	[0.200]	[0.206]
BenchmarkDum	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Constant	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Year/Industry FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
# Observations	36,461	36,461	36,461	36,461	36,461	36,461	36,461	36,461	36,461	36,461	36,461	36,461
# Firms	5,352	5,352	5,352	5,352	5,352	5,352	5,352	5,352	5,352	5,352	5,352	5,352

3.4 Alternative explanations

As explained before, there are several possible explanations for higher M&A activity by data intensive firms. One is that data firms simply have a high incentive to merge because of the particular market characteristics and the assumption that data is a valuable resource.

Another explanation is that the data economy is still young with a large number of firms and fierce competition, possibly resulting in typical consolidation dynamics including higher M&A activity in early industry life cycles.

The first approach to address this is to include industry size and competition intensity, here based on Hoberg and Phillips (2016)'s industry networks using their product similarity measure (mean similarity of each firm's 10 most similar companies), the sales-based Herfindahl-Hirshman Index (HHI) within each firm's industry network, and the logarithmized absolute number of firms within each firm's industry network in the regression. Table 4 reports a persistent positive influence of data intensity, while the competition measures have no influence in case of the product similarity and the industry size measure, but significant negative influence on both becoming an acquirer when using the HHI measure. This result indicates that higher industry concentration relates to a decrease in the probability of acquirers to invest in M&A, but does not change the positive and significant marginal effect of data intensity.

Table 4: Multinominal logit regression results including competition measures

The table reports the marginal effects of a multinominal logit regression using random effects and robust standard errors. The dependent variable is $TransactionParty$, which equals 1 for non-transaction parties, turning to 2 if a firm was an acquirer and to 3 if a firm was a target in the following fiscal year. Model (1) includes the product similarity measure, i.e., the mean similarity score of the 10 closest competitors; Model (2) uses a sales-based Herfindahl-Hirshman index (HHI) within a firm's industry network; Model (3) applies the logarithmized number of firms within a firm's industry network ($lnIndustrySize$). $Casefirm$ is a dummy variable, indicating the firms Google, Facebook, Microsoft, LinkedIn, and CoreLogic, which were the reference firms for measuring the cosine similarities for the data intensity scores. The life cycle stages report differences to the base category $Introduction$ and are constructed based on Dickinson (2011). Z scores are reported in parentheses. Significance levels are indicated by: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

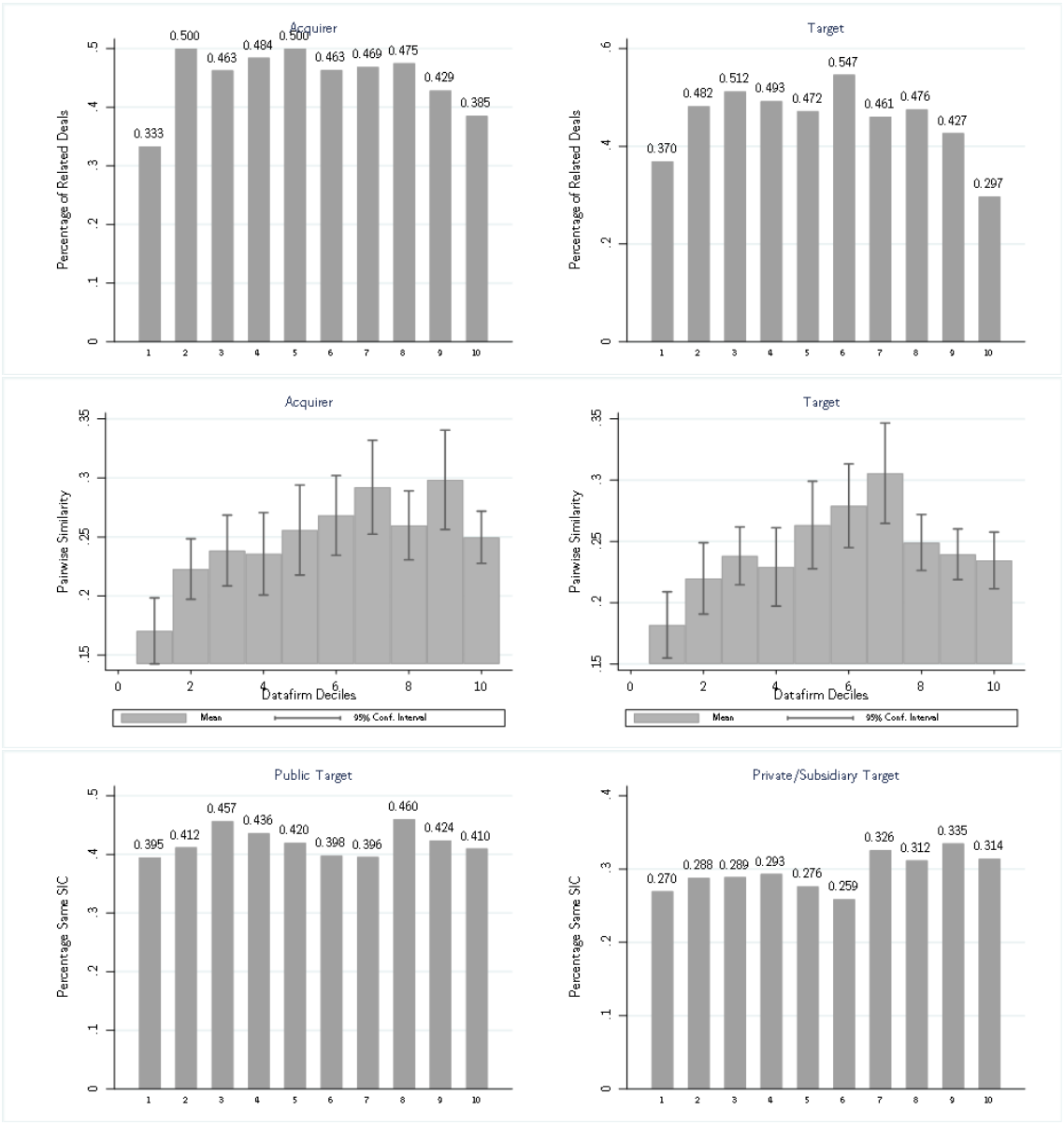
DV: TransactionParty	*Acquirer*			*Target*		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
DataIntensity	0.194***	0.188***	0.193***	0.029*	0.028*	0.027*
	[6.013]	[5.814]	[5.939]	[2.256]	[2.106]	[2.050]
lnMarketToBook	0.044***	0.045***	0.045***	-0.019***	-0.019***	-0.019***
	[6.500]	[5.609]	[6.565]	[-7.480]	[-7.510]	[-7.524]
Casefirm	0.559***	0.556***	0.557***	-0.783***	-0.784***	-0.783***
	[5.629]	[3.606]	[5.600]	[-22.450]	[-22.523]	[-22.467]
Firm Age	0.002***	0.002***	0.002***	-0.000*	-0.000**	-0.000***
	[6.844]	[7.228]	[7.055]	[-3.903]	[-4.002]	[-3.893]
ProductSimilarity	-0.090			0.020		
	[-1.632]			[0.896]		
HHI		-0.028+			-0.010	
		[-1.825]			[-1.451]	
lnIndustrySize			-0.000			-0.001
			[-0.047]			[-1.309]
Life Cycle Stage						
Growth	0.101***	0.101***	0.101***	-0.002	-0.002	-0.002
	[10.326]	[10.299]	[10.357]	[-0.451]	[-0.501]	[-0.440]
Mature	0.103***	0.104***	0.104***	-0.001	-0.001	-0.001
	[10.498]	[10.533]	[10.569]	[-0.148]	[-0.209]	[-0.120]
Shake-Out	0.044***	0.044***	0.044***	0.004	0.004	0.004
	[4.127]	[4.110]	[4.140]	[0.746]	[0.711]	[0.757]
Decline	-0.003	-0.004	-0.004	0.001	0.001	0.001
	[-0.277]	[-0.337]	[-0.313]	[0.197]	[0.190]	[0.177]
Constant	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes
Industry FE	yes	yes	yes	yes	yes	yes
# Observations	34,704	34,704	34,704	34,704	34,704	34,704
# Firms	5,296	5,296	5,296	5,296	5,296	5,296

The second approach is to evaluate the extent of industry relatedness of a deal measured by (a) the occurrence of the transaction parties in each other's industry networks, (b) the

pairwise similarity between transaction parties, and (c) the occurrence of the transaction parties in each other's primary SIC industries.[7] Closer relatedness should on the one hand indicate more fierce competition between transaction parties. On the other hand, less relatedness for high data intensity scores can also indicate that data is always valuable, both as a complement and supplement, and therefore also attractive transaction parties for more diverse firms. These two contrary predictions could offset each other, which can be one explanation for the depiction in figure 9. For both, acquirer and target, the relatedness increases in the beginning and slightly decreases for higher data intensities (see upper and middle graphs). As these two measures allow only the inclusion of public – public deals, analyzing SIC-Code relatedness provides also information on public – private/subsidiary deals (see lower graphs). Generally, the share of same-SIC Code deals is lower when acquirers buy private or subsidiary targets, but there is no systematic difference for different data intensity levels. Overall, these graphs do not show a clear pattern of transaction party relatedness for different data intensity levels.

[7] Data for industry networks and HHI are retrieved from Hoberg and Phillips (2016), primary SIC industries from Thomson Reuters SDC.

Figure 9: Relatedness of transaction parties

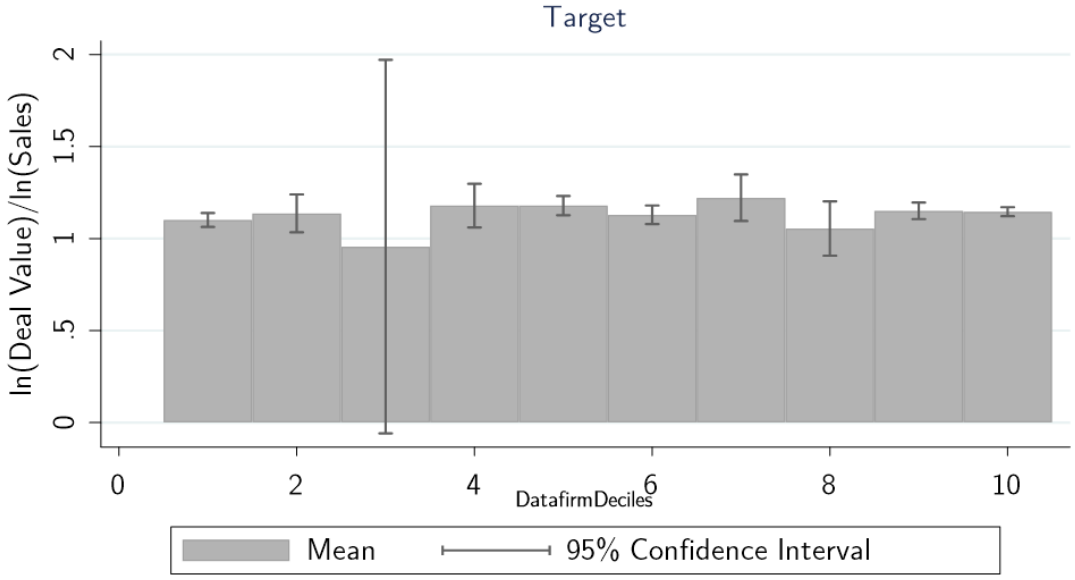


The two upper graphs show the relative frequency of transaction parties occurring in each others' network industries within each data intensity score decile, separately for acquirers and targets. Number of observations: 668. The two middle graphs show the mean and confidence interval of pairwise similarity between transaction parties within each data intensity score decile, separately for acquirers and targets. Number of observations: 730. The two lower graphs shows the relative frequency of transaction parties occurring in each others' primary SIC industries within each acquirer data intensity score decile, separately for public and private/subsidiary targets. Number of observations: public target: 1,633; private / subsidiary target: 17,403.

Furthermore, prominent examples as the Facebook/WhatsApp deal were characterized by extremely high deal value to sales multiples. Facebook paid $19.5 billion for WhatsApp earning a revenue of $10.2 million in the pre-acquisition year (New York Times 2014). As

WhatsApp did not cross revenue thresholds for being likely to significantly change revenue-based competition indicators, data firm deals gave rise to lacking regulatory attention. To look into this, the deal value to sales multiple is depicted in figure 10. However, the graph does not show a specific pattern over the data intensity score deciles.

Figure 10: Deal value over sales per data intensity decile



The figure shows the mean of ln(deal value) over ln(sales) within each data intensity score decile. Number of observations: 1,851 (between 131 (1^{st}) and 234 (10^{th}) observations within the deciles).

Regressing a dummy variable indicating whether a deal had to be approved by some competition authority on data intensity reveals that data intensive acquirers are overall not more exposed to attention from competition authorities (see table 5, models (1)-(3)). Nevertheless, higher data intensity increases the probability to get attention by competition authorities when the target firm is public, which holds also for the loose definition of the data firm dummy, but not for the strict one (see models (4)-(6)).

As deal values are only available for half of the deals, I include a dummy indicating whether a deal value is reported or not ($isDealValue$). Since deal values are expected to be rather reported for larger deals, the dummy variable proxies for deal size, independent of the firm's public status. Deal values are reported for 8,164 (47%) public – non-public and for 1,600 (98%) public – public deals. Again, for the strict definition of the data firm dummy there is no significant result. Data intensity, however, shows significant negative marginal effects which are offset if a deal value is reported. Using the loose definition of the data firm dummy again shows a positive and significant interaction effect, but no significant effect for the data firm dummy itself. There is no significant difference between only-loose-definition data firms and both-definitions data firms, indicating no countable difference in regulation for data accumulators and data analyzers or storage companies. If reporting a

deal value in Thomson Reuters SDC is a good proxy for deal size, the result indicates that there is less attention by competition authorities for data intensive acquirers when buying small targets. However, more evidence is needed to evaluate this effect thoroughly.

Table 5: Logistic regression explaining competition authority investigation

The table reports the marginal effects of a logistic regression using robust standard errors clustered on the firm level. The dependent variable is $CompetitionAuthority$, which equals 1 for deals that were investigated by some competition authority and else 0. Model (1) reports the regression including the continuous data intensity variable; Model (2) and (3) use the two different dummy variables for data firms; Models (4)-(6) include interaction terms of the data firm variables with the $PublicTargetDummy$. $BenchmarkDum$ corresponds to the dummy variables $Casefirm$, $ManualDatafirmDummy$ and $ManualNonDatafirmDummy$, representing those firms that were used as benchmark firms for the data intensity and data dummy variable construction. Z scores are reported in parentheses. Significance levels are indicated by: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

DV: CompetitionAuthority	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Datafirm Variable	Intensity	StrictDef	LooseDef	Intensity	StrictDef	LooseDef	Intensity	StrictDef	LooseDef
DataIntensity	-0.007			-0.041			-0.167***		
	[-0.267]			[-1.537]			[-3.614]		
DatafirmDummy		0.002	0.003		-0.005	-0.006		-0.031	-0.033
		[0.280]	[0.409]		[-0.477]	[-0.626]		[-1.040]	[-1.553]
PublicTargetDummy	0.138***	0.138***	0.137***	0.059**	0.135***	0.133***	0.056**	0.088***	0.087***
	[29.351]	[29.331]	[29.290]	[3.029]	[27.872]	[26.864]	[2.845]	[20.243]	[19.422]
DataIntensityxPublicTarget				0.132***			0.054		
				[4.037]			[1.621]		
DatafirmDummyxPublicTarget					0.025	0.028*		0.008	0.011
					[1.541]	[2.296]		[0.505]	[0.924]
isDealValue							0.034	0.131***	0.129***
							[1.218]	[16.797]	[16.228]
DatafirmxisDealValue							0.170***		
							[3.627]		
DatafirmDummyxisDealValue								0.034	0.035+
								[1.234]	[1.758]
AcquirorlnSales	0.009***	0.009***	0.009***	0.010***	0.009***	0.009***	0.014***	0.014***	0.014***
	[7.272]	[7.577]	[7.620]	[7.336]	[7.586]	[7.641]	[11.306]	[11.621]	[11.670]
Constant, BenchmarkDum, Year/Industry FE	yes	yes	yes	yes	yes	yes	yes	yes	yes
# Observations	18,407	18,407	18,407	18,407	18,407	18,407	18,407	18,407	18,407

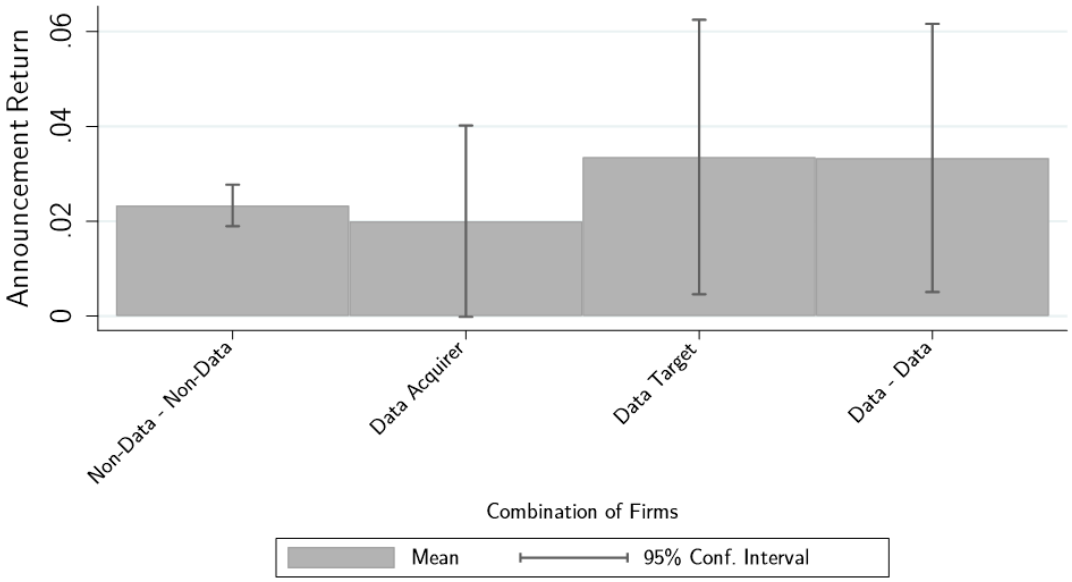
3.5 Merger announcement returns of data firms

An additional analysis evaluates merger announcement effects to analyze whether acquirers overpay when being a data intensive firm or when buying data intensive firms. To estimate abnormal returns at merger announcement, I use an event study with a 252 day calibration period and a 2 day event window covering the announcement day and the day after. Daily firm returns (r_i) and the CRSP value weighted index (r_m) are collected from CRSP. The risk free return (r_f) is the three-month treasury bill from the Federal Reserve Bank.

An overall t-test reveals that combined (value-weighted) abnormal returns of acquirer and target are overall significantly positive ($mean = 2.4\%$, $t - stat = 15.73$).

This result is not dependent on data intensity as there is no clear pattern of combined abnormal returns for the different data intensity deciles. This holds also when using both specifications of the data firm dummy and is independent of the combination of data and non-data firms (see figure 11). This result suggests that being a data (intensive) firm does not lead to different patterns regarding value creation through a merger.

Figure 11: Combined abnormal announcement return per firm combination



The figure shows the mean and confidence interval of combined abnormal returns within each data intensity score decile, separately for acquirers and targets. Number of deals: in each category: non-data – non-data: 685; data acquirer: 37; data target: 39; data – data: 29.

Acquirer and target abnormal returns separately show no differences for the data intensity deciles. Using the dummies reports significantly higher data firm target abnormal returns ($t - stats = -1.75$ (strict); -1.99 (loose)). Acquirer abnormal returns are significantly lower for data firms, but only when using the loose dummy definition ($t - stat = 2.10$).

Furthermore, looking at the reaction of the acquirers' closest competitors reveals that there is only a significant difference between data acquirers and non-data acquirers when using the strict definition of data firms ($t - stats = 2.14$ (strict); 0.82 (loose)). The result for the strict specification is driven by significant negative announcement returns for data acquirer competitors and zero announcement returns for non-data acquirer competitors. An acquirers closest competitors are defined as the up to ten most similar firms with a minimum cosine similarity score of 0.01. This result indicates that particularly data accumulators fear deals by other market participants. A possible explanation for this reaction are the market characteristics as described before, i.e., high economies of scale and network effects, with a natural monopoly as a probable outcome.

4 Discussion and conclusions

The data economy has become an integral part in everyday life. Research, public, and policy makers are discussing on data intensive firms and their impact on welfare and society. Among others, the market power and competitive dynamics of data accumulating firms and their markets are subject to more and more regulatory attention. However, antitrust authorities appeared to have difficulties to assess a possible harm to consumers due to a higher data concentration within one firm, among others caused by M&A deals. Adding to the up to now mainly theoretical or anecdotal discussion on data intensive firms, this study identifies the data intensity of firms using textual analysis and applies this measure to analyze whether data intensity corresponds to more M&A activity.

The data intensity measure proposed here appears to reflect typical data firm characteristics, as higher scores of the measure correspond to higher market to book values, lower firm age, and the most frequent occurrence in the Fama French 49 industries Computer Software and Communication. Additionally, firms that are expected to score high actually do so, such as Amazon, Adobe, Yahoo, the five case firms Google, Facebook, Microsoft, CoreLogic and LinkedIn, as well as firms that are listed in the Deutsche Bank Cloud and Big Data Index Certificate.

A limitation of this data intensity measure that needs to be addressed is a possibly high Type I error, hence firms scoring high even though they would not be considered as typical data firms. I address this by categorizing a randomly collected subsample manually into data and non-data firms using one very strict and a second more loose definition of data firms. Then, a k-nearest neighbor algorithm sorts the rest of the sample into this binary classification.

Regarding M&A activity, the data intensity of firms corresponds to a higher probability of acquiring or being acquired.

Controlling for competition intensity in the firms' markets, firm age, and the firms' life cycle stages does not alter the effect of the data intensity measure. This means that even though higher industry concentration negatively relates to M&A activity, industry consolidation mechanisms do not seem to drive higher M&A activity of data intensive firms.

Industry relatedness of public transaction partners measured by (a) the pairwise business description similarity, (b) occurrence in each other's industry network as measured by Hoberg and Phillips (2016), and (c) occurrence in each other's SIC industries overall do not depict a clear pattern. This could mean that a possibly lower relatedness, which could mean that data assets are valuable also for more diverse transaction partners, might offset a possibly higher relatedness, indicating higher competition between data intensive firms in early industry consolidation stages.

Furthermore, the deal value to sales multiples are not different for data intensive firms. Therefore, at least for public – public transactions, anecdotal evidence of huge multiples of, e.g., almost 2000 in case of the Facebook/WhatsApp deal, cannot be confirmed in the present sample.

There is some indication that data intensive acquirers need less often approval by competition authorities if the target is small, but significantly more often if the target is large or a publicly listed firm. However, as data availability is limited for non-public targets, this result needs further investigation.

Additionally, as combined M&A announcement returns not different for data firms, data firms do not seem to overpay if they are acquirers and are not overpaid if they are targets.

The main limitation of this study is the sample selection, as (1) small reporting companies are excluded because of missing risk factors in 10-K's and (2) only public acquirers and targets could be included. This means that information on for example the competition cases used for extracting keywords had to be omitted for all but one target (LinkedIn). Therefore, the famous cases like Google/DoubleClick (DoubleClick was a subsidiary) and Facebook/WhatsApp (WhatsApp was a private firm) only appeared in the acquirer, but not in the target statistics. This means that especially the acquisition of early stage businesses had to be omitted, as they are often not public (yet).

A Appendices

A.1 Data filters

Table 6: Data filters

The table reports the applied filters resulting in the final sample for the analyses in this paper.

filter	**# obs**
all *crsp/compustat merged (ccm)* firm-*fyears* 2006 - 2017	67,924
listed at NYSE, NASDAQ or AMEX	60,920
valid accounting values	50,635
merge of historical CIKs via *comphist* (*ccm*)	50,566
excluding small reporting companies	46,779
10-K filings published $<$ 365 days after *compustat* fiscal year end date	40,491
Item 1 (1A) with at least 1000 (200) characters	38,652

A.2 Case texts - business descriptions

Keywords are in *italic*.

Google/DoubleClick 2008

Google operates an *Internet search engine* and provides *online advertising* space on its own *websites* as well as on partner websites (*affiliated* to the Google 'AdSense' network). More recently, especially via the acquisition of YouTube, Google started to *provide content*. Google derives almost all of its revenues from online advertising.

CoreLogic/DataQuick 2013

CoreLogic, a publicly-traded company headquartered in Irvine, California, provides real property *information*, *analytics*, and *services* through a host of *products* tailored to the needs of customers in the lending, investment, and real estate industries. As part of its *Data* and Analytics segment, CoreLogic *collects*, *maintains*, and offers *licenses* for national assessor and recorder bulk data.

Facebook/WhatsApp 2014

Facebook (hereinafter also referred to as the 'Notifying Party') is a provider of *websites* and *applications* for *mobile devices* ('apps') offering *social networking*, *consumer communications* and photo/video *sharing* functionalities. Facebook also provides *online advertising* space. In particular, Facebook offers the social networking *platform* 'Facebook', the

consumer communications app 'Facebook Messenger' and the photo and video-sharing platform 'Instagram'.

Microsoft/LinkedIn 2016

Microsoft is a global *technology* company, whose product offering includes *operating systems* ('OSs') for *personal computers* ('PCs'), servers and *mobile devices*, related services, cross-device productivity *applications* and other *software* solutions, *hardware* devices, *cloud*-based solutions, *online advertising* (primarily with its *web search engine*, Bing).

LinkedIn operates a professional *social network* ('PSN') and generates revenues through the following product lines: (i) 'Talent Solutions' (63% of its revenue), which include *recruiting tools* and *online education courses*; (ii) 'Marketing Solutions' (19% of its revenue), which allow individuals and enterprises to *advertise* to LinkedIn's PSN members; and (iii) '*Premium* Subscriptions' for both consumer and businesses (18% of its revenue).

Sanofi/Google/DMI 2016

(2) Google is a multinational *technology* company specialising in Internet-related search and *products*. It operates an Internet search engine and provides online advertising space on its own websites and partner websites. Google also offers a number of other online *services* and *software* products.

A.3 Keywords business description

Table 7: Summary of keywords in business descriptions

The table reports the keywords for the business description keyword search. The keywords were collected from business descriptions of five case firms, i.e., Google, Facebook, Microsoft, LinkedIn, and CoreLogic, by the European Commission and/or the Federal Trade Commission, as shown in Appendix A.2 *Informative* are those keywords that appeared more often in the case firm business descriptions than in all business descriptions in the sample.

informative		**not informative**
advertising	analytics	affiliated
application	cloud	collect
communication	computer/PC	course
consumer/user	content	education
data	device	engine
hardware	information	maintain
internet	license	premium
mobile/phone	network	product
online	operating	recruiting
personal	platform	sharing
provide	search	system
server	service	
social	software	
technology	tool	
website		

A.4 Keywords risk factors

Table 8: Summary of keywords in Item 1A Risk Factors

The table reports a summary of the over 300 keywords for the risk factor keyword search. The large amount of synonyms for the words / terms listed below are omitted for better comprehensibility. The keywords were collected from Item 1A risk factors in 10-K's for the fiscal year 2016 of five case firms, i.e., Google, Facebook, Microsoft, LinkedIn, and CoreLogic. *Informative* are those keywords that appeared more often in the case firm risk factors than in all risk factors in the sample.

informative	**not informative**
(cyber) attack	California's Information Practices Act
datacenter	consumer/customer protection
data collection	consumer/customer protection regulation
data protection	data access
data protection regulation	data acquisition
data security	data deletion
data storage	data encryption
data disclosure	data localization laws
laws regarding privacy	data registration
personal data	data transfer
privacy	do-not-track
privacy concern	General Data Protection Regulation
privacy invasion	information processing
privacy policy	provide data
privacy protection	Safe Harbor
security breach	self-regulatory principles for online behavioral advertising
security control	targeted advertising
use data	

A.5 Firm characteristics

Table 9: Accounting characteristics

The table reports accounting values for non-merged and merged, and in particular acquirer, target, and both, firm-fiscal year observations. The market to book ratio equals market assets over book assets. Market assets is market value of equity (fiscal year end stock price · current shares outstanding) + book assets − book value of equity − deferred taxes. Age is proxied by the current fiscal year − first fiscal year of a firm (*gvkey*) occurring in Compustat. *Data/non − data* represent firm-fiscal year observations classified as data firms as explained in section 2.2.2. *T-test* represents the t-statistics for testing the significance of the difference between the means of each, no data firm / data firm and non-merged / merged.

(all whole numbers are in million U.S. Dollars)	all firm-fiscal years							only merged		
	all	**non-data**	**data**	*t-test*	**non-merged**	**merged**	*t-test*	**acquirer**	**target**	**both**
Sample size	38,649	36,389	2,260		26,598	12,051		10,189	1,562	300
Size measures										
Market Assets	11,770	11,553	15,262	-2.40	9,152	17,548	-10.75	19,897	4,193	7,296
Book Assets	8,705	8,702	8,748	-0.03	7,234	11,952	-6.47	13,513	3,133	4,864
Market Equity	5,393	5,137	9,511	-8.71	3,702	9,124	-21.42	10,360	1,982	4,347
Book Equity	2,070	2,042	2,522	-2.31	1,580	3,152	-14.97	3,556	795	1,708
Plant, Property, & Equip.	1,532	1,558	1,133	2.64	1,352	1,934	-6.97	2,150	708	1,178
Long Term Debt	1,598	1,564	2,156	-3.01	1,298	2,262	-9.70	2,504	848	1,370
Capital Expenditure	232	229	277	-1.94	196	311	-9.29	345	113	195
Net Income	256	248	380	-3.84	162	464	-17.45	536	55	167
Sales	3,636	3,618	3,928	-0.92	2,669	5,730	-18.09	6,429	1,595	3,523
Sales per Employee	0.866	0.895	0.417	2.53	0.803	1.002	-2.06	1.048	0.730	0.862
Performance measures										
Return on Assets	0.018	0.020	-0.017	0.67	0.018	0.018	-0.02	0.029	-0.053	0.026
Return on Equity	0.253	0.271	-0.046	1.15	0.300	0.149	1.11	0.168	0.043	0.078
ln(Market to Book Assets)	0.491	0.480	0.667	-14.38	0.485	0.505	-3.07	0.516	0.435	0.489
Leverage measures										
Book Leverage	0.574	0.575	0.556	2.41	0.578	0.564	3.67	0.560	0.585	0.576
Market Leverage	0.396	0.401	0.317	14.40	0.407	0.373	11.70	0.366	0.415	0.377
Age measure										
Compustat Age	21.30	21.71	14.75	20.36	20.83	22.33	-8.65	22.93	18.99	19.44

A.6 Regression analysis robustness checks

Table 10: Logit regression robustness checks results

The table reports the marginal effects of a logit regression using random effects and robust standard errors. The dependent variable is *Merged*, turning to 1 if a firm was a transaction partner in the following fiscal year. Model (1) uses the continuous data intensity variable measured using relative word frequencies instead of dummies; Models (2) and (3) use two different specifications of the dummy variable for data firms turning to one if (a) an observation is sorted into the data firm category by a naïve Bayes algorithm and (b) has a data intensity score in the upper 30% quantile. Z scores are reported in parentheses. Significance levels are indicated by: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

Dependent Variable:	**(1)**	**(2)**	**(3)**
Datafirm Rel. Freq.	0.158***		
	[5.317]		
DatafirmDummyStrictDef Naïve Bayes		0.026**	
		[2.752]	
DatafirmDummyLooseDef Naïve Bayes			0.016^{+}
			[1.777]
lnMarketToBook	0.022**	0.022**	0.022**
	[3.157]	[3.147]	[3.199]
Casefirm	0.416***	0.414***	0.417***
	[3.653]	[3.640]	[3.674]
ManualDatafirmDummy		0.013	0.014
		[0.154]	[0.171]
ManualNonDatafirmDummy		-0.002	-0.003
		[-0.070]	[-0.127]
Constant	yes	yes	yes
Year/Industry Dummies	yes	yes	yes
# Observations	36,754	36,754	36,754
# Firms	5,387	5,387	5,387

References

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman, 2016, The economics of privacy, *Journal of Economic Literature* 54, 442–92.

Berk, Jonathan B, and Peter M DeMarzo, 2007, *Corporate Finance* (Pearson Education).

Caillaud, Bernard, and Bruno Jullien, 2003, Chicken & egg: competition among intermediation service providers, *RAND Journal of Economics* 34, 309–328.

Campbell, James, Avi Goldfarb, and Catherine Tucker, 2015, Privacy regulation and market structure, *Journal of Economics & Management Strategy* 24, 47–73.

Casadesus-Masanell, Ramon, and Andres Hervas-Drane, 2015, Competing with privacy, *Management Science* 61, 229–246.

Correia-da Silva, Joao, Bruno Jullien, Yassine Lefouili, and Joana Pinho, 2019, Horizontal mergers between multisided platforms: insights from Cournot competition, *Journal of Economics & Management Strategy* 28, 109–124.

Dickinson, Victoria, 2011, Cash flow patterns as a proxy for firm life cycle, *The Accounting Review* 86, 1969–1994.

European Commission, 2008, Summary of Commission decision of 11 March 2008 declaring a concentration compatible with the common market and the functioning of the EEA Agreement (Case COMP/M.4731 - Google/DoubleClick), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2008.184.01.0010.01.ENG&toc=OJ:C:2008:184:TOC, last access date: 15 Jan 2019.

European Commission, 2014, Mergers: Commission approves acquisition of WhatsApp by Facebook, http://europa.eu/rapid/press-release_IP-14-1088_en.htm, last access date: 15 Jan 2019.

European Commission, 2016a, Commission decision of 23/02/2016 declaring a concentration to be compatible with the common market (Case No COMP/M.7813 - SANOFI / GOOGLE / DMI JV) according to Council Regulation (EC) No 139/2004, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1549876772849&uri=CELEX:32016M7813, last access date: 15 Jan 2019.

European Commission, 2016b, Mergers: Commission approves acquisition of LinkedIn by Microsoft, subject to conditions, http://europa.eu/rapid/press-release_IP-16-4284_en.htm, last access date: 17 Jan 2019.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp, 2019, Big data and firm dynamics, in *AEA Papers and Proceedings*, volume 109, 38–42.

Federal Trade Commission, 2007, Federal Trade Commission closes Google/DoubleClick investigation, https://www.ftc.gov/news-events/press-releases/2007/12/federal-trade-commission-closes-googledoubleclick-investigation, last access date: 15 Jan 2019.

Federal Trade Commission, 2014a, FTC notifies Facebook, WhatsApp of privacy obligations in light of proposed acquisition, https://www.ftc.gov/news-events/press-releases/2014/04/ftc-notifies-facebook-whatsapp-privacy-obligations-light-proposed, last access date: 15 Jan 2019.

Federal Trade Commission, 2014b, FTC puts conditions on CoreLogic, Inc.'s proposed acquisition of DataQuick Information Systems, https://www.ftc.gov/news-events/press-releases/2014/03/ftc-puts-conditions-corelogic-incs-proposed-acquisition-dataquick, last access date: 15 Jan 2019.

Gentzkow, Matthew, and Jesse M Shapiro, 2010, What drives media slant? Evidence from US daily newspapers, *Econometrica* 78, 35–71.

Hartmann, Philipp, and Joachim Henkel, 2018, Really the new oil? a resource-based perspective on data-driven innovation, *Academy of Management Global Proceedings* 142.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of Political Economy* 124, 1423–1465.

Katz, Michael L, Carl Shapiro, et al., 1985, Network externalities, competition, and compatibility, *American Economic Review* 75, 424–440.

Lambrecht, Anja, and Catherine E Tucker, 2015, Can big data protect a firm from competition?, *Available at SSRN 2705530* .

Loughran, Tim, and Bill McDonald, 2019, Stage one 10-x parse data, https://sraf.nd.edu/data/stage-one-10-x-parse-data/, last access date: 15 May 2019.

New York Times, 2014, Facebook's $21.8 billion WhatsApp acquisition lost $138 million last year, https://dealbook.nytimes.com/2014/10/28/facebooks-21-8-billion-acquisition-lost-138-million-last-year/, last access date: 20 Jan 2019.

Posner, Richard A, 1981, The economics of privacy, *The American Economic Review* 71, 405–409.

Rhodes-Kropf, Matthew, David T Robinson, and Sean Viswanathan, 2005, Valuation waves and merger activity: the empirical evidence, *Journal of Financial Economics* 77, 561–603.

Rochet, Jean-Charles, and Jean Tirole, 2006, Two-sided markets: a progress report, *The RAND Journal of Economics* 37, 645–667.

Shy, Oz, and Rune Stenbacka, 2016, Customer privacy and competition, *Journal of Economics & Management Strategy* 25, 539–562.

Taylor, Curtis, and Liad Wagman, 2014, Consumer privacy in oligopolistic markets: winners, losers, and welfare, *International Journal of Industrial Organization* 34, 80–84.

The Economist, 2017, The world's most valuable resource is no longer oil, but data, https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, last access date: 08 Feb 2019.