STANFORD
COMPUTATIONAL
ANTITRUST

ARTICLE

Using Natural Language Processing to Delineate Digital Markets

Klaus Gugler*, Florian Szücs** & Ulrich Wohak***

Abstract. Delineating relevant antitrust markets poses substantial challenges, particularly so in nascent, digital markets, where data on prices, quantities, and costs often are not available. This study evaluates a complementary approach using Natural Language Processing techniques along with business descriptions of relevant firms to define markets. Applying this method to a sample of start-up acquisitions, we find considerable overlap between our approach and expert assessments by the European Commission.

* Klaus Gugler, Vienna University of Economics and Business, klaus.gugler@wu.ac.at.
** Florian Szücs, Vienna University of Economics and Business, florian.szuecs@wu.ac.at.
*** Ulrich Wohak, Vienna University of Economics and Business, ulrich.wohak@wu.ac.at.

I. Introduction

Traditionally, competition authorities define the market or markets involved in a case before assessing competitive conditions and claims about potential pro or anticompetitive effects. Defining the relevant market, encompassing both a product and a geographic dimension is a crucial initial step in any antitrust analysis. Market definition can be understood as defining the boundaries of competition between firms and constitutes the space within which competitive effects manifest. As such, it serves as the cornerstone for evaluating structural indicators of competition such as market shares and concentration indexes.[1] The central challenge in defining relevant markets lies in cleanly delineating which products are direct substitutes and, consequently, compete with each other. The most significant competitive constraint for these products is due to demand-side substitution, with supply-side substitution often playing a somewhat lesser role. Potential competition is not typically considered during the market definition phase but is factored in during the competitive assessment. There has been growing awareness among competition practitioners and academics alike that particularly for nascent, digital markets the tools of market definition are lackluster. More recently, competition authorities have begun to address structural changes like globalization and digitization, which may have altered how markets work, in response to criticisms related to both the geographic scope and product market delineation.[2]

Arguably, one of the most significant innovations in antitrust analysis over the past four decades has been the advent of the "more economics-based approach" (*see* e.g. UNCTAD, 2009). Under this approach, the conventional practice of ad hoc market definition has been replaced by the adoption of the SSNIP test, which stands for "small but significant non-transitory increase in price."[3] In essence, this approach poses the question: Would a hypothetical monopolist in a potentially relevant market find it economically viable to sustain a permanent price increase of 5-10%? If insufficient consumers would shift their consumption away to render the price hike

[1] Both the European Commission's 2004 Horizontal Merger Guidelines and the US 2010 Horizontal Merger Guidelines acknowledge that assessing levels and shifts in market shares and concentration measures can offer valuable insights when gauging the potential competitive impact of a merger or determining the presence of a dominant position. *See* Amelia Fletcher & Bruce Lyons, 'Geographic Market Definition in European Commission Merger Control: A Study for DG Competition' (Centre for Competition Policy, 2016).

[2] For example, the draft of the new FTC and DOJ merger guidelines indicates a partial abandoning of the importance of market definition and suggests looking directly at the theories of harm due to a merger. *See* U.S. Dep't of Justice & Fed. Trade Comm'n, Horizontal Merger Guidelines (2023), https://www.ftc.gov/system/files/ftc_gov/pdf/p859910draftmergerguidelines2023.pdf.

[3] *See* U.S. Dep't of Justice & Fed. Trade Comm'n, Horizontal Merger Guidelines (1992), https://www.justice.gov/sites/default/files/atr/legacy/2007/07/11/11250.pdf.

unprofitable, then the market is deemed as defined. Conversely, if a substantial number of consumers would substitute away, making the price increase unprofitable, then the candidate market is expanded (either in terms of product scope, geographic reach, or both), and the test is repeated. This process is iterated until a hypothetical monopolist would find raising prices profitable.[4] While authorities may not apply this test verbatim in every merger case, it nonetheless serves as a guiding principle in their analyses.[5]

While the SSNIP test is a sound economic tool to delineate markets in theory, its implementation remains challenging in practice for a number of reasons. First, to empirically implement the SSNIP test competent authorities must have access to appropriate product-level price and quantity data over a number of time periods to implement structural models or reduced-form reasoning. Typical analyses conducted in competition cases use such data to model consumer behavior in response to historical price changes for calculating e.g., diversion ratios.[6] Second, it is not always clear what exactly constitutes a product, such as what might be the case in technology mergers where firms tend to offer a whole ecosystem rather than just a single product.[7]

Market definition in digital markets remains a complex challenge, especially when considering the acquisition of start-up firms. Start-ups are firms in their nascent stages, characterized by rapid evolution and often not selling well-defined products. Some start-ups may primarily consist of innovative ideas or a pool of human talent with the capacity for product development. At this early juncture, pricing, outputs, or revenues may be ill-defined or nonexistent. Consequently, attempting to estimate substitution patterns for products that may not even exist yet is a challenging task. Moreover, forecasting supply-side substitution, i.e., which firms may enter the market after an SSNIP, also becomes exceedingly difficult.

---

[4] *See* George J. Stigler & Robert A. Sherwin, *The extent of the market* 28(3) J. LAW & ECON. 555 (1985); Gregory J. Werden & Luke M. Froeb, *Correlation, causality, and all that jazz: The inherent shortcomings of price tests for antitrust market delineation* 8 REV. INDUS. ORG. 329.

[5] This is reflected in the European Commission's revised Market Definition Notice where the Commission clarifies that the SSNIP test is a useful conceptual tool but that there is no obligation to apply it. Eur. Comm'n, Notice on the definition of the relevant market for purposes of Union competition law, COM(2024) (February 22, 2024), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AC_202401645.

[6] Typically, competition authorities assess the amount of (consumer) diversion from product A to product B by asking consumers a SSNIP-type question such as "If the price of product A was to increase by 5-10%, would you switch to another product/supplier?". Aggregating responses yields diversion rations that help understanding market boundaries.

[7] *See* John D. Harkrider, Operationalizing the Hypothetical Monopolist Test, (Dec. 19, 2023), DEP'T OF JUST. ARCHIVES https://www.justice.gov/archives/atr/operationalizing-hypothetical-monopolist-test (last visited Apr. 18, 2024).

While these start-ups may not presently pose direct competition to large tech firms and may even produce complementary products, they can emerge as formidable competitors developing their ecosystems and substitutable products. Consequently, the lack of well-defined market definitions in this context raises concerns about a policy void within merger control, particularly concerning start-up acquisitions by major tech companies. The fact that the majority of start-up acquisitions by GAFAM (Google, Apple, Facebook, Amazon, and Microsoft) firms have not been notified to competition authorities, only a few have been challenged and none have been prohibited, underscores the pressing need to address this issue.

Our study attempts to algorithmically identify relevant markets using text data. We do so by calculating the matrix of cosine similarities based on the business descriptions of virtually all start-up firms. In other words, we compute pairwise similarities of firms' business activities as documented by their respective business description.[8] This measure increases as two firms employ more similar terminology in their business descriptions and thus indicates the degree of similarity between specific firms and their competitors. More specifically, we vary the similarity parameter from low values (indicating little similarity) to high values (representing a high overlap of business descriptions) and derive various sets of firms that constitute candidate relevant markets. For each candidate market, we estimate treatment effects, examining how mergers impact venture capital investment while considering the size of the relevant market, which varies with the similarity parameter. In other words, we investigate how using more 'loose' (less similarity, more potential rivals) or 'tight' (more similarity, fewer potential rivals) market definitions affect the estimated treatment effects.[9] The delineation of relevant markets is a vital policy concern, and our continuous adjustment of the relevant market size allows us to shed light on this issue.

We are not the first to posit that text-based similarity measures are useful for the measurement of competition or price elasticities of demand. Hoberg and Phillips (2016) and Pellegrino (2019) measure competition by firm-by-firm pairwise similarity scores using 10-K product filings.[10]

[8] We compute similarities for each pair of firms contained in our data set. For example, consider a sample of three firms, A, B and C. The complete set of pairwise similarity scores assign a similarity score for each pair (A, A), (B, B), (C, C), (A, B), (A, C), (B, A), (B, C), (C, A) and (C, B). By definition, the similarity between identical pairs is equal to one, e.g., for the pair (A, A). The similarity score for is symmetric with respect to the firm pair, i.e. the similarity score for the pairs (A, C) is the same as for the pair (C, A). Hence, the pertinent scores are (A, B), (A, C) and (B, C).

[9] We have looked at venture capital funding as the outcome variable, which can be regarded as an important indicator of future market development. As mentioned above, data on prices and quantities are not available for these markets.

[10] *See* Gerard Hoberg & Gordon Phillips, *Text-based network industries and endogenous product differentiation* 124(5) J. POL. ECON. 1423, 1465 (2016). They show that this measure effectively captures

Hoberg and Phillips (2016) find that a 21.32% minimum similarity threshold generates 10K-based industries with 2.05% membership (i.e., in 2.05% of all firm pairs, both firms are in the same industry), which corresponds to an average SIC-3 industry.[11] Pellegrino (2019) uses the concept of product market centrality to calculate – under assumptions on the mode of competition – cross-price elasticities utilizing the extent of overlap of common characteristics.[12] He obtains own- and cross-price elasticities of demand quite similar to those found in the literature.

This paper is structured as follows. In the next section, we discuss the power and pitfalls of using natural language tools for market definition. We elaborate on the nature of text data/business descriptions and how they might be useful for market definition. In section III, we discuss a specific method, Latent Semantic Indexing. In section IV, we apply the market definition algorithm to a dataset of start-up acquisitions and show how varying the similarity-score cutoff affects the findings. Section V compares the results of the algorithmic market definition to an expert assessment by the European Commission (EC), and section VI concludes.

II. Natural Language Processing: Methods for Market Definition

In this section, we provide an overview of natural language processing (NLP) methods relevant to market definition and describe the typical pipeline required for producing inputs for such methods. Any analysis of text data starts with a sample of (raw) text from which the researcher wants to infer information. For example, consider a cross-sectional data set of firms for which we have a textual representation of their main business activity. If we let i denote a firm, then $\{D_i\}$ denotes the set of business descriptions ("documents") describing the economic activities in our sample. Together, these documents form the corpus of text C from which the researcher wants to distill information, e.g., about the competitive interactions between firms.

In contrast to numeric data, text is inherently high dimensional. Gentzkow et al. (2019) provide a simple example. Consider that each business description is 30 words long, where each word is drawn from (say) the 1,000 most frequently used words in the English language. Then, the unique representation of these business descriptions has dimension 1000^{30}. Additionally, text as a manifestation of language adds

competition among firms, in particular competition among potential rival firms that offer related products. For instance, they show that their measure of competition based on explains specific managerial discussions of high competition, rivals identified by managers as peer firms, and changes to industry competitors following exogenous industry shocks. *See* Gerard Hoberg and Vojislav Maksimovic, *Redefining financial constraints: A text-based analysis*, REV. OF FINANC. STUD. 1312, 1352 (2015), for more evidence of this measure to capture competition.

[11] Hoberg & Phillips, *supra* note 10; Bruno Pellegrino, *Product differentiation and oligopoly: a network approach* (WRDS Research Paper, 2019).

[12] Pellegrino, *supra* note 11.

additional complexities that are not (necessarily) present in conventional economic data. That is, text data is noisy along dimensions that are absent for numeric data. For example, a single word might have multiple meanings (polysemy, e.g. the word "sound") or, conversely, there are multiple words representing the same concept (synonymy, e.g. "work", "profession" and "occupation"). There is a plethora of NLP methods available to researchers, and depending on the application and the degree of sophistication, they tackle the issues described above to varying degrees.

In the discussion that follows, we consider that each document D_i is l_i words long, and we denote the complete vocabulary of the corpus as W. The first step in the analysis of text data is to reduce the dimensionality of the data referred to as *preprocessing*. This typically requires researchers to apply domain-specific knowledge about the research question as well as the text in order to remove (add) features from (to) the data that carry little (meaningful) information. In essence, this represents a feature selection process that reduces the dimensionality of the data. The next step in most applications is to represent the preprocessed text as a numerical array in order to make it workable for a wide range of computational methods. In the subsection below, we describe preprocessing steps that are commonly found in generating inputs for statistical models for text data.

A. Preprocessing

The main purpose of preprocessing text data is to manipulate the information contained in $\{D_i\}$ such that it is appropriate for further analysis. Preprocessing is the task of removing or adding information to the data in such a way that the preprocessed corpus only contains information that can be meaningfully analyzed by statistical models. In virtually all NLP applications, preprocessing involves (1) tokenization, (2) the removal of stopwords and punctuation, (3) stemming or lemmatizing words, and (4) detection and definition of n-grams. Each concept will be explained in turn.

Tokenization: involves breaking down text into smaller, meaningful chunks called tokens. A simple example is to break down a sentence into separate words, where each word is a token. Tokenization is useful in that this representation of the text allows us to numerically represent it as a matrix.

Stopwords and punctuation: the removal of stopwords involves the deletion of words that carry little semantic meaning, such as "and," "or," forms of "to be," etc. While stopwords are useful to the human reader for they provide grammatical structure, they are typically not a decisive feature that helps to distinguish one

document from another. There is no single agreed-upon list of stopwords and the researcher may add or delete words to the list of stopwords.[13] Punctuation is removed for the same reason as stopwords: they typically do not help distinguish one document from another.

Stemming and lemmatizing: Text normalization techniques that have as their goal to normalize inflected words to their roots. For example, "located" and "locates" have as their common root: "locate." Note that in a world with infinite computing power and data, stemming and lemmatizing is not necessarily useful as we are removing information from our data set by stemming. For example, "located" contains temporal information, which is removed after stemming. In practice, stemming and lemmatizing is a reduction in the dimensionality of W by trading off information for computational speed. The reason for the extensive use of stemming and lemmatizing is that in most applications, the inclusion of inflected versions of word roots does not add a lot of information. While both stemming and lemmatizing reduce inflected words to their roots, the expected output is not identical. The difference is that the output of lemmatization is always an actual word found in an (e.g., English) dictionary, while that is not necessarily the case for stemming (*see* Table I).

Raw text	Porter-stemmed	Lemmatized
compete	compet	compete
competing	compet	compete
competed	compet	compete
competitors	competitor	competitors
competition	competit	competition

Table I: Examples of stemming vs. lemmatizing

n-grams: A common issue in considering each word separately is that the meaning of single words often does not convey the same meaning as their combination. For example, considering the words "artificial" and "intelligence" separate from one another does not convey the same meaning as "artificial intelligence." In natural language processing, such phrases are referred to as n-grams. The specific example "artificial intelligence" is a 2-gram (or bigram), as it consists of two words. The concept of n-grams includes single words as 1-grams. Depending on the research question, it

[13] Most natural language processing software packages come with pre-defined lists of stopwords which may vary in the number of words included. For example, the default list of stopwords in two widely used NLP-software packages in Python include 127 and 306 words in NLKT and spaCey, respectively. Clearly, the list of stopwords will have to match the language of the corpus.

might be important to model these n-grams explicitly.[14] The inclusion of n-grams for $n > 1$ typically comes at the expense of increasing the computational cost. Most applications include n-grams of order 2 or 3, but it is best practice to start with considering 1-grams only and then compare the accuracy of the model output with corpora containing 2 or 3-grams[15].

B. Representing text as numeric data

One can broadly distinguish between two representations of text as data: *ordered* and *unordered*. In an unordered representation, each document D_i is represented by a numeric vector v_i of length $dim(W)$, where each element in v_i corresponds to the number of times word w_i occurs in D_i. The corpus can then be represented by a matrix A of dimension $D \times W$ where each row of A corresponds to a document D_i and each column corresponds to a word $w_i \in W$ in the (preprocessed) vocabulary. This particular representation is called the "Bag of Words" (BoW) representation. The name is indicative of the fact that BoW representations do not take the order of words into account; it is simply a numeric representation of the frequency count of each word for every document. At this stage, the BoW representation already provides a lot of information about the structure of $\{D_i\}$. Consider a case where the researcher is interested in clustering similar documents and suppose that there exists some w_i that occurs exactly once in every document. Clearly, this particular w_i does not contain meaningful information to distinguish one document from another. Conversely, if a word occurs only in a single document, it is not helpful in clustering similar documents together. A commonly used approach to deal with very common or rare words is to assign a weight to each word corresponding to its "term frequency-inverse document frequency" ("tf-idf"). Term frequency ($c_{i,j}$) simply counts the number of times word j (or other feature) occurs in document i. Inverse document frequency is a measure of how often word i appears *across* documents. Typically, this is expressed as a logarithmically scaled inverse fraction of the number of documents containing word i. This commonly used re-weighting procedure can expressed as:

$$tf - idf(i,j) = \underbrace{(1 + log(c_{i,j}))}_{\text{tf}} * \underbrace{log\left(1 + \frac{1+D}{1+d_W}\right)}_{\text{idf}} \quad (1)$$

where d_W is the number of documents word i appears in. An extension to representing the corpus by a (tf-idf-weighted) term-frequency matrix is to represent each word by some m-dimensional vector. Such representations are referred to as *word embeddings*. The vector representation of single words is typically learned through neural networks and computationally demanding. There are many ready-to-use software

[14] Note that adding n-grams adds information to the raw text and can be considered additional tokens.
[15] Matthew Gentzkow, Bryan Kelly & Matt Taddy, *Text as Data*, 57 J. ECON. LIT. 535 (2019).

packages that come with pre-trained word embeddings to save on computational cost.[16] A word of caution might be prudent as word embeddings are corpus-specific as the particular representation is learned *from* the corpus. As such, using pre-trained word embeddings might work well for some applications, but not others.

There are richer representations that take the order of words into account for the meaning of a word might change when used in a different context. Typically, this involves representing each sentence of a document by a $W \times w_{ki}$ matrix, where each row corresponds to a word in the vocabulary W and w_{ki} is the number of words in sentence k in document i. Each entry in this matrix represents the occurrence of word $w_i \in W$ in column-position w_{ki}. The resulting representation for each sentence is then a sequence of points. The main drawback of using representations that take the order of words into account is that it increases the dimensionality of the data, and hence the computational cost, dramatically.

III. Methods for text-based market delineation

A. Overview

The drastic reduction in costs associated with computation has spurred research in, and subsequently increased the number of statistical methods available for, text data. The narrow application of defining antitrust markets from text data is not amenable to every type of model. Therefore, we refrain from providing an exhaustive overview of NLP methods and focus on applications in the literature as well as indicate which methods could be useful for such tasks.

In order to guide the discussion, we consider a situation where a researcher or practitioner is confronted with a collection $\{D_i\}$ of N texts describing the business activity of a sample of N firms. This may include the description of products, services, geographic coverage, etc. Throughout the analysis, we consider that there are M possible antitrust markets in the sample. As such, the main objective is to map (business descriptions of) firms D_i to markets M. In most applications, neither the number of markets M nor the number of firms within each market is known. Since M is not observed, a variety of methods are unsuitable a priori.

For example, text-based regression functions, such as penalized regressions or regression trees, typically require formulating a conditional expectation function. In the context of market definition, it is conceivable to model the probability of a firm operating in markets m_i as some function of its textual business description, $E[m_i|x_i]$,

[16] For example, word2vec, fastText and GloVe.

where x_i is some known transformation of v_i, the numeric vector representation of firm i's business description (*see* Section IIB). Clearly, in the setting described above m_i is not observed. While text-regressions have seen various applications in the literature, they are not particularly useful without information on m_i since $E[m_i|x_i]$ cannot be estimated.

A similar issue arises for the class of supervised machine learning algorithms. A necessary (first) step in the application of supervised algorithms is to split A into a training set A^{train} and test set A^{test} (with information on m_i suitably included). In the context of market definition, one could then fit a model $f_\theta(v_i;m_i)$ using A^{train} only and validate the (predictive) accuracy of the model using estimates of θ and A^{test}. Again, this model requires information on m_i, which is not observed. The class of algorithms that is suitable for the setting described above are "unsupervised" methods. These algorithms allow to infer (latent) m_i from A by, for example, clustering together similar business descriptions. Researchers might be familiar with k-means clustering algorithms, but the computational linguistics literature has produced a wide array of applicable methods that are suitable for inferring markets from a corpus. Instead of providing an exhaustive overview, we present one unsupervised method in detail with an application to tech markets.[17]

B. Latent Semantic Indexing

In this section, we describe Latent Semantic Indexing (LSI), an unsupervised machine learning algorithm, that allows us to map firms to markets based on their business description. Consider a situation in which the researcher has data on business descriptions $\{D_i\}$ for each firm $i = 1,...,N$. The goal of the exercise is to infer competitive relationships between firms using $\{D_i\}$ data only. Following the pipeline described in Section II, the first task is to represent (tokenize) the corpus into a document-term matrix $\boldsymbol{A}$ where each row corresponds to a document $d \in D$ and each column corresponds to a word $w \in W$, where D and W correspond to the set of all documents ("corpus") and the complete vocabulary of the corpus, respectively.[18] Each element in $\boldsymbol{A}$ corresponds to the number of times word w occurs in document d:

$$\underset{D\times W}{A} = \begin{pmatrix} c_{1,1} & \cdots & c_{1,W} \\ \vdots & \ddots & \vdots \\ c_{D,1} & \cdots & c_{D,W} \end{pmatrix}$$

[17] In addition to the Latent Semantic Indexing (*see* section IIIA), topic modelling and clustering algorithms are generally useful for allocating firms to markets based on their business description.

[18] We assume that there is exactly one business description for each firm. Hence, $N=D$.

The matrix $\boldsymbol{A}$ can then be re-weighted by a tf-idf transformation (*see* section IIA) to weigh tokens according to their term frequency-inverse document frequency. Let $\boldsymbol{B}$ denote the re-weighted matrix:

$$\underset{D\times W}{B} = \begin{pmatrix} tfidf(c_{1,1}) & \dots & tfidf(c_{1,W}) \\ \vdots & \ddots & \vdots \\ tfidf(c_{D,1}) & \dots & tfidf(c_{D,W}) \end{pmatrix}$$

This re-weighted matrix $\boldsymbol{B}$ is then decomposed using singular value decomposition (SVD). SVD transforms B of rank r into three matrices.

$$\underset{D\times W}{B} = U\Sigma M^T$$

where $\boldsymbol{U}$ is a $D \times r$ orthogonal matrix, $\boldsymbol{\Sigma}$ is a $r \times r$ diagonal matrix and $\boldsymbol{M}^T$ is an $r \times W$ orthogonal matrix. The final step in LSI is the elimination of $(r - C)$ rows and columns in $\boldsymbol{\Sigma}$ corresponding to the smallest eigenvalues of $\boldsymbol{\Sigma}$, where C is a scalar chosen by the researcher. There is no optimal scalar C, but it is recommended to choose $100 \leq C \leq 1000$ (Martin and Berry, 2007). In our application, $C = 500$.[19] This truncation process results in the best rank-C approximation $\boldsymbol{B_C} = \boldsymbol{U_C\Sigma_C M_C}^T$ to the input matrix $\boldsymbol{B}$. Each document d_i is represented as a $1\times C$ vector v_i in the document-component matrix $\boldsymbol{U_C\Sigma_C}$. The last step is to calculate the similarity between documents. To that end, we use a widely-used measure of similarity that considers two documents (vectors) i and j to be similar if the angle between the two vectors is small. This measure is referred to as *cosine similarity* and can be expressed as:

$$S(v_i, v_j) = \frac{\sum_{c=1}^{C}(v_i \cdot v_j)}{\sqrt{\sum_{c=1}^{C} v_i}\sqrt{\sum_{c=1}^{C} v_j}}$$

where $S(v_i,v_j) \in [-1,1]$ increases in the similarity of documents. Note that this is a pair-wise measure of similarity and, for the sake of market definition, must be computed between any two vectors v_i and v_j. By definition, $S(v_i,v_i) = 1$. The result of this exercise is a symmetric $D \times D$ *similarity matrix* where each entry along the diagonal is equal to unity and the off-diagonal elements represent the cosine similarity between document i and j, $j \neq i$. This matrix is useful since it allows the researcher to retrieve documents that are "similar" to, say, firm i. The notion of similarity is now formalized by defining a particular cosine similarity cutoff. For example, in the application below we consider that firms (and their respective business description) are sufficiently similar, and hence compete against one another

[19] A robustness exercise with C=400 and C=600 suggests that our results are not sensitive to the choice of C.

if the cosine similarity of their business description is greater than 0.3. Since there is no objective criterion for choosing the similarity cutoff, researchers must proceed at their own discretion by, for example, visually assessing the similarity of the business description at the candidate cutoff.

IV. Natural Language Processing: Application

In a 2023 working paper, we apply LSI (*see* Section IIIA) to define antitrust markets for technology start-ups and subsequently estimate the effects of acquisitions by Google, Apple, Facebook, Amazon, and Microsoft on aggregate investment activity in these markets.[20]

In this section, we discuss how the results of such an approach vary depending on subjective choices made by the researcher. As illustrated in section II, the process of converting a body of text to a dataset of numeric vectors involves several steps of pre-processing. For example, one must decide whether the words in the dataset are standardized through stemming or lemmatizing the text. Similarly, after the conversion to a numerical format, a weighting scheme for the terms must be chosen. These choices obviously affect the final dataset obtained.

While defining antitrust markets for start-up firms, we find that the effect of choices made during the pre-processing stage and the numerical conversion of the data on the outcome is somewhat limited.[21] Thus, our findings are largely robust with regard to the NLP parametrization. Instead, the most important parameter in our example is the choice of similarity-score cutoff: the size of the treatment group varies substantially depending on how similar we require affected start-ups to be to the GAFAM acquisition target.

Figure I shows that the number of start-ups that we consider to be affected by a GAFAM acquisition changes strongly depending on the similarity-score cutoff. While requiring a low cosine similarity of 0.2 yields a mean (median) of around 3500 (2000) affected firms, these numbers drop sharply for more stringent similarity scores. Increasing the cutoff to 0.7 (0.6) leads to a mean (median) treatment group size very close to zero. While these numbers constitute a logical upper bound for the similarity we can impose, it is not clear which cutoff in the interval [0.2,0.7] should be

[20] Klaus Gugler, Florian Szücs & Ulrich Wohak, *Start-up acquisitions, Venture Capital and Innovation: A Comparative Study of Google, Apple, Facebook, Amazon and Microsoft* (Department of Economics Working Paper Series 340, WU Vienna University of Economics and Business, 2023).

[21] While we did not conduct a comprehensive grid search across all parameter values for reasons of computational feasibility, we tried out a number of different approaches that seemed reasonable to us, obtaining similar results.

chosen. The choice reflects the bias/variance trade-off often encountered in applied statistics: should we impose a low similarity cutoff resulting in a large sample size at the risk of including unaffected start-ups (bias), or should we rather only look at very similar start-ups, risking a low sample size (variance)?

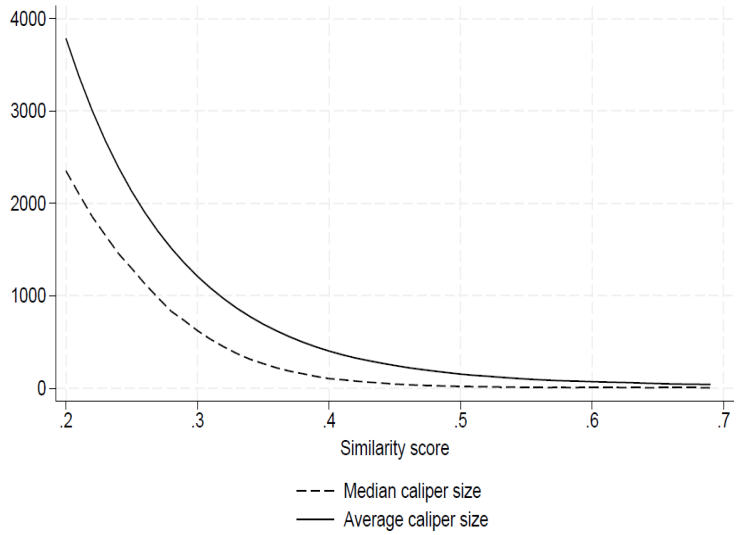

Figure I: Mean and median caliper size as a function of similarity score

We illustrate this trade-off in Figure II, where we plot the treatment effect (the decrease of investment rounds in similar firms) as a function of the similarity score used to define the treatment group. The estimated treatment effect for a similarity score of 0.2 is around -20%, suggesting that the frequency of investment in similar start-up firms declines by 20% after a GAFAM acquisition. Increasing the required similarity cut-off to be affected by the takeover increases the estimated effect (i.e., the coefficient size decreases). If we increase the required similarity to around 0.35, the frequency of investments declines by 30% and for a similarity cutoff of 0.6 we obtain an effect of -40%. Thus, as the similarity of start-ups in the treatment group (relative to the GAFAM target) increases, so does the estimated effect.

At the same time, we observe a substantial increase in the observed confidence intervals. Increasing the required similarity lowers sample sizes, as illustrated above where the efficiency of estimation declines and the confidence bands widen. In the empirical exercise in Gugler et al. (2023), we opted for a similarity-score cutoff value of 0.3, which appeared to be a reasonable compromise on the bias/variance trade-off.[22] Figure II seems to corroborate this choice: the estimated treatment effect

[22] Klaus Gugler et al., *Start-up Acquisitions, Venture Capital and Innovation: A Comparative Study of Google, Apple, Facebook, Amazon and Microsoft*, Technical Report (2023), https://research.wu.ac.at/ws/portalfiles/portal/44832243/WP340.pdf (last visited Apr. 18, 2024).

becomes more and more pronounced as we increase the similarity score from 0.2 to around 0.35. It then enters a somewhat flat region, before becoming increasingly numerically unstable at around 0.6. Thus, in this example, imposing a similarity score between 0.3 and 0.4 seems to yield good results.

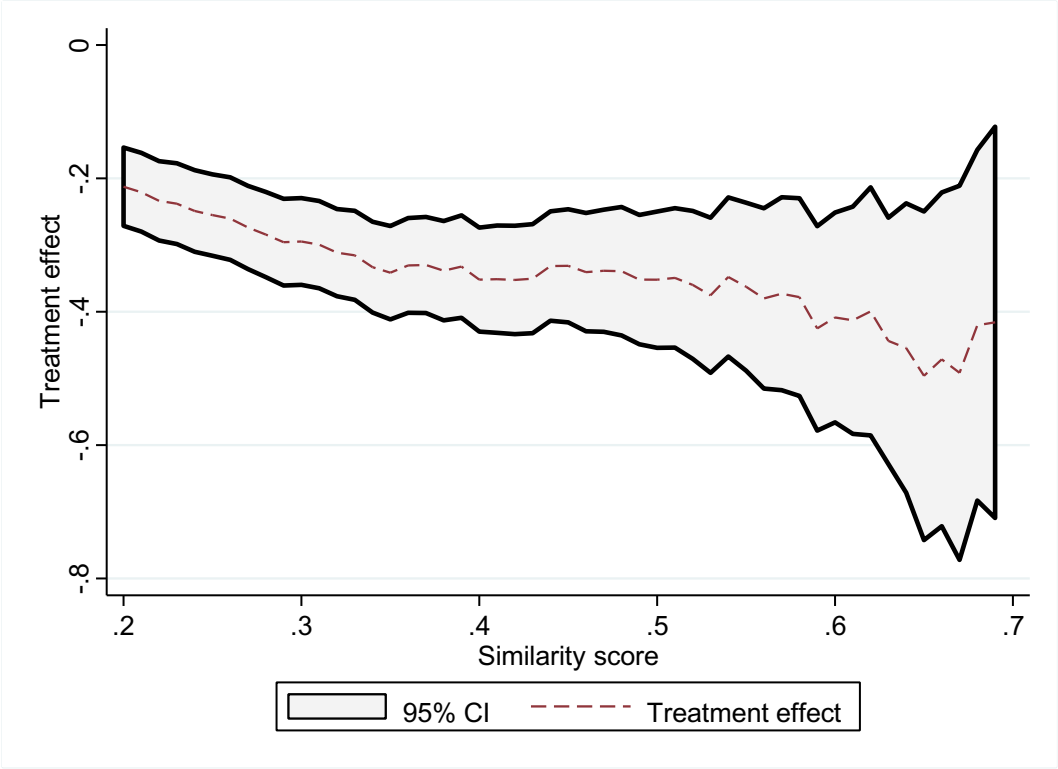

Figure II: Treatment effect as a function of similarity score

V. Algorithmic and Expert Market Definition

In this section, we conduct a reality check on the algorithmic approach to market definition proposed in the previous section. For a small set of start-up acquisitions, we have data in both i) our GAFAM acquisition dataset and, ii) competitive assessments by the EC's DG Competition. Because few start-up acquisitions meet the EC's notification thresholds, the overlap is small. Specifically, the cases *Apple/Beats*, *Facebook/WhatsApp*, *Google/DoubleClick* and *Microsoft/Skype* were both scrutinized by the EC and in Gugler et al. (2023).[23]

Table 2 reports how many rivals the EC identified in each case and how many of those were also identified by our NLP algorithm. The table shows that, on average, half of the competitors identified by the EC were also found by the relevant market

[23] *Id.*

algorithm.[24] While the majority of EC rivals are discovered in the *Apple/Beats* and *Google/DoubleClick* cases, only slightly more than a third are identified for the other two cases.

Figure III shows that the share of EC rivals found is a function of the stringency of the similarity cutoff. As we increase the required similarity, the number of EC rivals found decreases. While the most EC rivals are found for low similarity values, those values also yield very large overall treatment group sizes (*see* Figure I). On the other hand, similarity cutoffs beyond 0.5 discard almost all EC-identified rivals.

Case	EC rivals	NLP rivals	Share found
Apple/Beats	9	7	78%
Facebook/WhatsApp	14	5	36%
Google/Doubleclick	16	9	56%
Microsoft/Skype	13	5	38%
Total (Average)	52	26	(50%)

Table II: Comparison of algorithmic and expert assessment

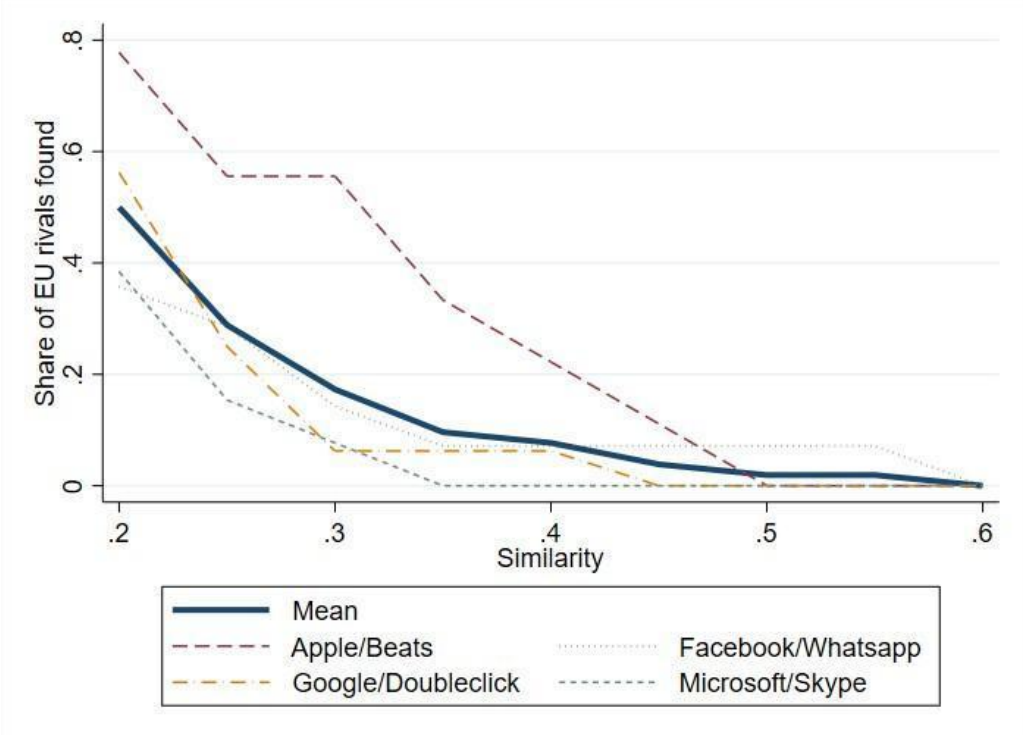

Figure III: Overlap

[24] We excluded a few EC rivals that are not contained in Crunchbase, and thus, the algorithm had no chance of identifying them. Including those rivals decreases the total share found to 44%.

As evident from Table II, the algorithmic approach to market delineation performs well in some cases (*Apple/Beats* and *Google/Doubleclick*) and relatively poor in others (*Facebook/Whatsapp* and *Microsoft/Skype*). Naturally, the quality of the results is a direct function of the underlying information fed into the algorithm. We exemplify this using one of the two weak results from Table II. For the *Microsoft/Skype* merger, our algorithm only identified 38% of the rivals identified by the EC. Among the rivals not identified are ICQ, AOL, and Yahoo, i.e. a number of relatively well-known providers of messaging applications. Recall that the approach uses Skype (and its business description) as the focal firm and then assesses the similarity of the business descriptions of potential rivals. To illustrate why some rivals were not found, consider first the business description of Skype:

Skype is for doing things together, whenever you're apart. Skype's text, voice and video make it simple to share experiences with the people that matter to you, wherever they are. With Skype, you can share a story, celebrate a birthday, learn a language, hold a meeting, work with colleagues (...). [25]

While the EC's experts—having detailed knowledge of the industry and the ability to request information from the merging parties—may have correctly identified Skype, IBM, and Yahoo to exert competitive pressure on each other, this is not evident from the companies' business description. [26]

Consider the business descriptions of two of the rivals identified by the EC but not by the algorithm, IBM and Yahoo:

IBM is a global technology and innovation company. It is the largest technology and consulting employer in the world, with more than 400,000 employees serving clients in 170 countries. IBM offers a wide range of technology and consulting services; a broad portfolio of middleware for collaboration, predictive analytics, software development, and systems management; and the world's most advanced servers and supercomputers.[27]

Yahoo is a technology company that is known for its web services and applications. It is focused on creating personal digital experiences that keep users connected to what matters most to them, across devices, and around the globe. Yahoo features Yahoo Sports, Yahoo Finance, Yahoo Fantasy, Yahoo Mail, and more in the works every day. The company was founded in 1994 and is headquartered in Sunnyvale, California.[28]

[25] Skype, *Crunchbase*, https://www.crunchbase.com/organization/skype (last visited Apr. 18, 2024).
[26] We provide a full list of business descriptions of Skype rivals that were identified by the EC but not by the LSI algorithm in the appendix.
[27] IBM, *Crunchbase*, https://www.crunchbase.com/organization/ibm (last visited Apr. 18, 2024).
[28] Yahoo, *Crunchbase*, https://www.crunchbase.com/organization/yahoo (last visited Apr. 18, 2024).

This example highlights a limitation of our approach: while the algorithm generally does well in identifying small firms with rather specific business descriptions (e.g., the algorithm identified the messenger services Fring and Oovoo.com), it is unable to link rather specific target business descriptions (such as that of Skype) to more generic business descriptions of larger tech companies, such as those of IBM or Yahoo. One potential way to mitigate this constraint would be to train the similarity algorithm on larger bodies of text describing the activities of firms. Using, e.g., full-text business reports or product filings of firms would very likely lead to improved matches, although at the cost of a substantially increased computational burden.

Conversely, our algorithm found several firms similar to Skype that were not listed in the EC's decision. The two closest business descriptions of firms that existed prior to the transaction were Telkom UMeetMe[29] and SayType.[30] It is possible that the EC experts did not include these firms in the relevant markets for a variety of reasons. For example, these two firms might be outside the geographical scope or they do not exert a sufficient competitive constraint to be considered part of the relevant market. While the former start-up is located in Indonesia, the latter appears to be at a rather early stage of development and has a speech-to-text feature. Thus, they may not (yet) impose a competitive constraint on *Microsoft/Skype* and, therefore, were not considered competitors by the EC. However, based on their business descriptions, they seem to offer very similar services. For example, Telkom UMeetMe's description includes "Video calls, instant messages, voice calls and video UMeetMe [sic] easier for you to share your experiences with the important people in your life,

[29] UMeetMe's full business description as recorded in Crunchbase: "Video calls, instant messages, voice calls and video UMeetMe easier for you to share your experiences with the important people in your life, no matter where they are. Solutions that accelerate the sharing of knowledge and overcome the challenges of distance and time zones. Make a call, see each other, send messages and share them with others. In the business world, this means that you can unite the entire ecosystem of employees, partners, and customers to complete your job. Try UMeetMe today and begin to add friends, family, and colleagues. With UMeetMe, you can hold meetings, working with colleagues to share stories, learn the language - everything you need to do together every day. You can use UMeetMe on the device that best suits you - on your phone or computer or a TV with UMeetMe installed. Start using UMeetMe to try group video call, talk, see, and send instant messages to anyone else on UMeetMe."

[30] SayType's full business description as recorded in Crunchbase: "Hands Free Texting, E-mail, and Internet Search Services – License Plate Messaging SayType Is the Most Effective and Complete Alternative to Touch Texting Available, Connecting All Drivers Free of Charge Innovative Driving Solutions That Work Instantly From Any Phone Using Standard Voice and Text Messaging Services SayType.com is giving everyone with phone service something to talk about; an innovative service that uses voicemail to instantly turn talk into text and send accurately converted messages to any phone, e-mail address or license plate. SayType also searches the internet and translates messages into selected languages. Call or text 916-947-7325 or visit www.saytype.com for details. "Our solutions are easy to use – leave a voicemail up to 2 minutes – receive or send a text message – arrive faster and safer every time." John Kent, SayType inventor."

no matter where they are" (UMeetMe business description, Crunchbase), which is functionally equivalent to the second sentence of Skype's description, cited above.[31]

The examples above highlight what NLP algorithms can and cannot do when it comes to market delineation. First, if the business descriptions of firms do not meaningfully capture the activities in which competition authorities are interested for a particular case, one can hardly expect it to perform well. Second, it appears that smaller or single-product firms improve the algorithmic performance, given that they have more targeted and specific business descriptions.

Thus, we conclude that an algorithmic approach to market definition, based on natural language processing and business descriptions, can—to a certain degree—approximate the findings of an expert assessment. Further, it appears that manual and algorithmic market definitions are complementary approaches. While human experts are good at identifying large and well-established rival firms, an algorithm can more easily uncover small and unknown firms that may be potential competitors based on the descriptions of their business activities.

VI. Conclusions

This paper tackles the problem of market definition in digital markets. Similar to Hoberg and Phillips (2016), we measure (potential) competition by firm-by-firm pairwise similarity scores using their business descriptions.[32] We form candidate relevant markets by varying the similarity parameter. For each candidate market, we estimate treatment effects of real-world mergers on venture capital investment. While we cannot determine a clearly defined cutoff value of similarity that warrants the conclusion that the respective firms are in the same relevant market, we view similarity score values between 0.3 and 0.4 as a good compromise in our example. Moreover, we find that the NLP algorithms perform well with smaller or single-product firms, given their more targeted business descriptions.

Our paper has important implications for market definition in merger cases involving digital markets. Traditional market definition usually involves multiproduct firms and actual competitors. In digital markets, however, traditional market definition yields misleading results as substitution patterns between potential or nascent competitors may not (yet) be traceable.

[31] Telkom UMeetMe, *Crunchbase*, https://www.crunchbase.com/organization/telkom-umeetme (last visited Apr. 18, 2024).
[32] Hoberg & Phillips, *supra* note 10.

This study illustrates that text data is a useful novel tool for merger analysis and suggests a number of actionable insights. First, NLP methods can add value to merger analysis since they enable practitioners to assess competitive interactions at scale and low cost. Our results suggest that algorithmic assessment of business descriptions to assess competitive closeness is not a standalone tool and requires careful consideration of competition practitioners. However, NLP methods allow authorities to consider a much broader set of firms that would otherwise be prohibitively expensive to assess by human effort alone. While our study does not identify all relevant competitors identified by the EC, our approach spotted a number of firms engaging in similar business activities that the EC did not consider. Our paper also highlights limitations: NLP methods perform well for single-product firms and are sensitive to insufficiently precise business descriptions.

Moreover, NLP methods can be a useful tool for competencies of competition authorities beyond merger control. Across many work streams of competition authorities, it is pivotal to identify relevant documents among thousands of irrelevant documents. This is a costly and time-intensive task to do for human agents. The method presented in this study can identify similar documents in a fraction of the time.

In addition, competition authorities frequently investigate markets in which they have grounds to believe that competition is not working in the best interest of consumers. The assessment of competitive conditions is a central element in market investigations and is frequently hampered by a lack of conventional market and firm-level data.

Appendix: Business descriptions of rivals of Skype identified by the EC but not the LSI algorithm

Firm	Business description
ICQ	ICQ was developed in 1996 by Mirabilis, the creators of the first fully functional Internet-wide Instant messenger comprising presence, buddy list, and rapid messaging. Since its early days, ICQ's mission has been not only to provide the world with quick and rich communication tools; the ICQ idea is to provide people, individuals, and groups around the world with the most complete means to find each other and communicate better.
AOL	AOL Lifestream is a web-based application that enables users to keep track of all their comments on social networking sites. AOL users can publish their statuses, reply to comments on networking sites from their Lifestream tab, and more. It was founded in 1985 and headquartered in New York, United States
IBM	IBM is a global technology and innovation company. It is the largest technology and consulting employer in the world, with more than 400,000 employees serving clients in 170 countries. IBM offers a wide range of technology and consulting services; a broad portfolio of middleware for collaboration, predictive analytics, software development, and systems management; and the world's most advanced servers and supercomputers. (...).
Yahoo!	Yahoo is a technology company that is known for its web services and applications. It is focused on creating personal digital experiences that keep users connected to what matters most to them, across devices, and around the globe. Yahoo features Yahoo Sports, Yahoo Finance, Yahoo Fantasy, Yahoo Mail, and more in the works every day. The company was founded in 1994 and is headquartered in Sunnyvale, California. (...).
Aastra Technolgies	Aastra Technologies Limited makes products and systems for accessing communication networks including the Internet. Its products include residential and business telephone terminals, screen telephones, Enterprise Private branch exchanges (PBX), network access terminals and high quality digital video encoders, decoders and gateways. (...).
eBuddy	eBuddy offers web and mobile messaging solutions. It created the world's first, independent, web browser-based instant messaging service as e-Messenger in 2003. The company was rebranded in 2006 from e-Messenger to eBuddy. eBuddy currently offers two products: eBuddy Chat and XMS. eBuddy Chat is an online messenger that enables global users with MSN, Yahoo, Gtalk, Facebook, ICQ, and AIM accounts to chat free of charge in one aggregated interface. It is currently available on the web and offers free mobile solutions for iOS, Android, J2ME, and mobile web-enabled devices. XMS is a free, real-time messaging app for smartphones. (...).
Avaya	Avaya is a global leader in communication systems, applications, and services. Learn about where they've been, where they're headed, and the leaders that are getting them there. The company's communication equipment and software integrate voice and data services for customers including large corporations, government agencies, and small businesses. Its office phone systems incorporate Internet protocol (IP) and Session Initiation protocol (SIP) telephony, messaging, Web access, and interactive voice response. Avaya also offers a wide array of consulting, integration, and other managed IT services. The company sells directly and through distributors, resellers, systems integrators, and telecommunications service providers.
Miranda Technologies	Miranda Technologies Inc. develops, manufactures, and markets high-performance hardware and software for the television broadcast industry. Its solutions are purchased by content creators, broadcasters, specialty channels, and television service providers to enable and enhance the transition to a complex multi-channel digital and HD broadcast environment. This equipment allows customers to generate additional revenue while reducing costs through more efficient distribution and management of content as well as the automation of previously manual processes.