Concentration and Internet Advertising: The Rise of Buyer Power *

Francesco Decarolis Gabriele Rovigatti

- Preliminary Draft -

Abstract

The effects of increased concentration in an industry fundamentally depend on how other industries trading with it respond. By looking at the market of advertising space on search pages, which Google firmly dominates, we document how buyers of ad space responded through technological innovation and increased concentration. By combining data on advertisers' affiliation to marketing agencies with data on their bidding in Google' sponsored search auctions, we analyze how changes in agency networks' concentration are associated with changes in Google' revenues. While concentration can lead to less aggressive bidding through increased buyer power, it can also allow a more efficient targeting of keywords through the use of superior information. We first use a machine learning algorithm to cluster the sample of keywords by thematic groups to define the relevant markets. Then, using an instrumental variable strategy, we assess the impact of changes in networks' concentration and find evidence of a negative effect on the search engine's revenues. The choice of the keywords on which to advertise appears to drive this result.

Keywords: Online Advertising, Internet Auctions, Marketing Agency, Ad Network, Agency Trading Desk.

JEL Classification: C72, D44, L81.

*Francesco Decarolis: Bocconi University, Department of Economics and IGIER. Via Reontgen 1, Milan, Italy. E-mail address: *francesco.decarolis@unibocconi.it.*. Gabriele Rovigatti: Bank of Italy, Research Department. Via Nazionale 91, Rome, Italy. E-mail address: *gabriele.rovigatti@bancaditalia.it*. The authors thank the NET Institute for financial support.

I Introduction

In economics, as well as in the media, new evidence has been hotly debated regarding the increased concentration in important European and US industries.[1] The profound implications of industry concentration on firms' competition, workers' salaries and, ultimately, consumers' welfare explain this revived interest in a classical industrial organization topic. The debate so far has centred around two key issues. The first is the quantification of concentration increases, with an emphasis on what is the proper use of industry data to identify markets. The second regards the impacts of concentration on competition, workers and consumers.

This study contributes to this latter aspect of the debate by looking at a feature that has been so far overlooked: how other industries trading with the industry that experiences high concentration respond to it. This is the old, but powerful idea of countervailing power. Galbraith [1952] notoriously remarked that "the best and established answer to economic power is the building of countervailing power: the trade union remains an equalizing force in the labor markets, and the chain store is the best answer to the market power of big food companies." A more recent, but no less egregious example in the case of the US healthcare is the insurers' introduction of HMOs and PPOs which is credited to have dramatically rebalanced power in favor of insurers after decades of hospitals' increased concentration.

But how and to what extent countervailing power can emerge? This study offers answers for the industry of internet advertising, which is both of major economic relevance and very clearly dominated by a single firm: Google. Thanks to its technological innovations, Google has come to dominate this market becoming a leading example of those "superstar firms" at the center of the academic debate. In particular, we analyze the sale of ad space on search pages (*sponsored search*), which represents about half of all of internet advertising revenues, or about $40 billion dollars in 2017.[2] This is a market that for nearly twenty years has been highly concentrated with Google earning a share between 75% and 80% of the total US

[1] See, among others, Autor et al. [2017], De Loecker and Eeckhout [2017], Werden and Froeb [2018], Gutierrez and Philippon [2017] and Weche and Wambach [2018], as well as the Obama administration's CEA [2016] and the press coverage by Economist [Economist, 2016*a*,*b*] and Guardian [Stiglitz, 2016].

[2] The Internet Advertising Board evaluates in $88 billion dollars the revenues in 2017 of the US internet ad industry, with the main tiers being sponsored search (46%), banner (31%) and video (14%) [IAB, 2018].

search ad revenues in the period 2016-2018 [eMarketer, 2018].

The sponsored search's demand side is constituted by advertisers seeking to capture the attention of users querying search engines. Advertisers buy ad space through auctions in which they compete for the adjudication of one of a given number of 'slots' available in the search engine result pages. Our analysis focuses on how buyers' concentration changed thanks to increasing bidding delegation to specialized intermediaries and how this rise of buyers' power affected Google's revenues. These intermediaries, known as "digital marketing agencies" (DMAs or, simply, agencies) are responsible for bidding in about 75% of Google's sponsored auctions in our data. DMAs contributed to major transformations in the market, making the bidding process faster and better targeted through the use of more sophisticated technologies and better data. The key innovation on which we focus is the creation of agency trading desks (ATDs): most DMAs belong to an agency network and, for each network, the ATD is the centralized unit that conducts all bidding activities for the advertisers affiliated to the network's agencies. In our data, just seven ATDs bid for about 50% of all the 6,000 advertisers, or for about 40% of the the 36 million keywords.

Agency networks can influence the auction outcomes in a number of distinct ways: from choosing the set of keywords and optimizing the individual bids to coordinating the advertisers' actions with those of rival advertisers whenever they are handled by the same DMA. The theoretical relationship between delegated bidding to DMAs and auctions' outcomes is complex as it might involve a multiplicity of actions: choice of keywords, bids, targeting, ad content, etc.. What seems clear, however, is that the increasing concentration in DMA and ATD has potentially important effects on the types of strategies available to agencies.

Earlier theoretical work by [Decarolis, Goldmanis and Penta, 2017] considers the situation in which agencies are jointly optimizing the bidding strategies of advertisers that have already chosen to bid on common keywords. However, a broader set of strategies is available, as the agency might sustain bid rotation or market split schemes, such that its clients are never directly competing in the auctions. Furthermore, beside any collusive intent, agencies can select keywords different from those proposed by the advertisers, enhancing efficiency and lowering costs. Indeed, earlier theoretical work has identified at least three problems of individual bidding that joint delegation to a common intermediary might help to alleviate: externalities between ad [Jeziorski and Segal, 2015], limited information leading to winners' curse [McAfee, 2011] and budget constrains leading to inefficiencies [Balseiro et al., 2017]. For all these aspects, the increase in intermediaries' concentration might create efficiencies that could also benefit the search engine. Thus, as typical in horizontal merger analyses, it is ex ante ambiguous how changes in market concentration for the ad networks would impact Google's revenues.

Our approach to answer the question of how intermediaries' concentration affects Google's revenues is based on three ingredients. First, a novel dataset obtained by combining multiple data sources and covering the years 2014-2017. We have obtained from Redbooks - the most comprehensive database on marketing agencies - a list of advertisers representing nearly the universe of major US firms active in online marketing. For each of these advertisers, the Redbook data gives us the full list of marketing agencies (both DMAs and traditional marketing agencies) affiliated with them, as well as the link of each individual agency to the ad network to which it belongs. We have combined these data on agencies with data on the Google's sponsored search auctions from SEMrush. For all Redbook advertisers, we know which keywords, if any, they bid on via Google. For each keyword and year, we know the position of the domain in the search outcome page, the volume of searches associated with the keyword (i.e., the average number of search queries for the given keyword in the last 12 months); the visible URL of the ad; the content of the ad; and, most importantly, the keyword-specific average CPC (i.e., the price advertisers pay for a user's click on an ad triggered by the given keyword).

The second element is a definition of relevant markets that exploits the richness of our data to cluster together keywords that represent markets, albeit not in a strict antitrust sense. We move from the advertiser- to the keyword-level in the definition of markets by grouping the keywords in thematic clusters. Such an approach is helpful in textual analysis in order to generate measures of distance among *documents* (i.e., groups of words) within the set of all documents. We first proceed to vectorize the K documents - that is, the universe of keywords bid by the universe of Redbooks advertisers 2014-2017 - through state of the art natural language processing techniques (the GloVe algorithm described below). We then cluster keywords together to form the equivalent of markets. We see this part of the analysis as a valuable contribution that addresses some of the concerns emerged in the current debate on increasing concentration.

The third element is an instrumental variable strategy that accounts for the fact that changes in network concentration might be driven by unobserved factors that also drive changes in the search auctions revenues. Therefore, after having specified our measures of both ad auctions revenues and ad network concentration, we construct for the latter an instrument exploiting the fact that the DMA market is highly dynamic and characterized by a relatively high number of mergers and acquisitions during the study period. We use such M&A events as exogenous variations to build an instrument for the changes in concentration (measured in terms of the Herfindahl-Hirschman Index, HHI). In particular, for each market-time combination we compute the "simulated change in HHI" as the counterfactual HHI change in market m at time t *induced* by the merger, absent any other changes.

The (preliminary) findings of our analysis reveal that ad network concentration induces lower growth of the search engine's revenues. This effect appears to be robust to a number of sensitivity assessment involving the way in which the clustering analysis to define the markets is conducted and the set of controls included in the model specification. The effect seems to be driven by the set of chosen keywords, which shifts toward less expensive (in terms of average cost per click, CPC) keywords. Therefore, despite the potential efficiencies created by delegation to the DMAs, the evidence seems to indicate that the increased buyer power of DMAs results in a worsening of the search engine's revenue prospects. Our ongoing work will aim at further disentangling the forces behind our main finding.

Related Literature. – TBA

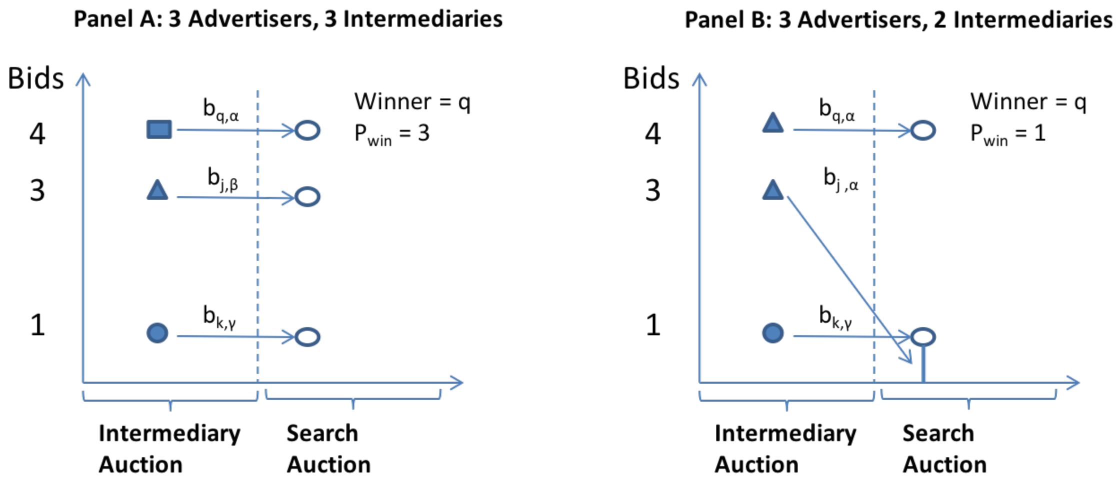
II Basic Framework

Suppose there is a monopolist search engine selling ad slots on its results page. Suppose also that there are three advertisers (q, j and k) interested in showing their ad to consumers searching some keyword w. Allocations and payments will clearly depend on how many ad slots the search engine places on its web page and on the selling mechanism adopted. For instance, if there is only one slot available and a second price auction is used, the winner will be the advertiser with the highest bid and his payment will equal the second highest bid.

Now suppose that advertisers do not bid directly on the search auction. They submit their bid to an intermediary who runs internally a second price auction among its clients (we shall refer to this as the *intermediary auction*) and then bids on their behalf in the search auction. To see why this can affect the functioning of the search auction, consider the two cases illustrated in Figure 1. In panel A, each advertiser bids through a different intermediary, which we indicate as α, β and γ. In this case, intermediaries have no incentive to distort bids in the search auction. Hence, if for instance the bids placed in the intermediary auction are $b_q = 4$, $b_j = 3$ and $b_k = 1$, the same bids will enter the search auction: $b_{q,\alpha} = 4$, $b_{j,\beta} = 3$ and $b_{k,\gamma} = 1$, as indicated by the straight arrows in Figure 1. Advertiser q obtains the slot and pays 3 to the search engine, while the others do not pay anything. In panel B, the situation differs because we have only 2 intermediaries: both q and j employ α. This intermediary can now alter the search auction outcomes by retaining or emending the bids it places on behalf of its two clients. It might report just the highest bid among the two, $b_{q,\alpha} = 4$, or both bids, but setting $b_{j,\alpha} \in [0, 1]$, as indicated in panel B by the interval to which the bid $b_{j,\alpha}$ leads. In all cases, q wins the slot, but paying only 1 instead of 3.

The logic of the previous example can be easily generalized in several ways to show why increasing intermediaries' concentration can lower the search engine's revenues. Suppose, for instance, that there is a single slot for sale, but now there are N advertisers and K intermediaries. Advertisers are endowed with arbitrary bids in the intermediary auction ($b_1 \geq ... \geq b_N$), so that b_2 is what the search engine would earn were each advertiser to hire a different intermediary. But, if advertisers are independently and uniformly assigned

Figure 1: An Example of Bidding through Intermediaries



Notes: There are three advertisers (q, j and k) submitting arbitrary bids ($b_q = 4$, $b_j = 3$ and $b_k = 1$) to a second price auction held by the intermediary to which they are affiliated. In panel A, each advertiser has a different intermediary (α, β and γ). In panel B, q and j share intermediary α. The arrows indicate how the intermediary translate the bids in its auction into the bids placed on the search auction. In panel A, bids are transmitted without distortions; in panel B, j's bid is reduced. q wins in both cases, paying the second highest bid which is either 3 (panel A) or 1 (panel B).

at random to intermediaries, the search engine's expected revenues is lower and equal to:

$$E(\pi) = \sum_{n=2}^{N} b_n(\frac{K-1}{K^{n-1}}). \tag{1}$$

$E(\pi)$ declines as K gets smaller, being minimized at zero when $K = 1$.[3] With the three bids of the previous example, $E(\pi)$ is 2.22 with 3 intermediaries and 1.75 with 2 intermediaries.

This problem with intermediaries' concentration had been noticed by the theoretical literature in computer science and economics. Mansour, Muthukrishnan and Nisan [2012] was the first study to point this out in the context of the ad exchanges where one slot per time is sold. Decarolis, Goldmanis and Penta [2017] further extended the analysis to the situation of multiple, heterogenous slots that is typical in sponsored search auctions. They show that when the multiple slots are sold via the Generalized Second Price (GSP) auction, as done by Google, both search engine's revenues and allocative efficiency are damaged by

[3]Equation (1) is assumes intermediaries behaving as in in Figure 1, panel B. Thus, $\pi = b_2$ whenever advertiser 2 is not with intermediary 1, which happens with probability $(K-1)/K$; $\pi = b_3$ if advertiser 2 is with intermediary 1, but advertiser 3 is not, which happens with probability $(K-1)/K^2$; etc..

intermediaries' concentration. Their most surprising result is that these effects are even more pronounced than under a benchmark system (the VCG auction, used, for instance, by Facebook) that is known to perform poorly when bidders play coordinated strategies.

In the typical situation where advertisers bid over a multitude of keyword auctions, the scope that an intermediary might have to coordinate bids is even greater. For instance, the market can be split across keywords or users' demographics. In general, the feasibility these strategies and the intermediary's incentive to implement them depend on technological, contractual and strategic considerations. For instance, an intermediary must both set up an internal system to select among its clients and contractually specify how the outcomes of this system will determine actions in the search auctions. The details can be crucial. For instance, if in the example in Figure 1 panel B part of q's surplus from having $b_{j,\alpha} \leq 1$ in the search auction is rebated back to advertiser j in proportion to j's bid in the intermediary's auction, then j might have an incentive to overstate his bid in the intermediary's auction. This would clearly reduce the extent to which intermediation can reduce prices in the search auctions. Relatedly, the losses will be reduced if advertisers that are close competitors choose not to share the intermediary.

When presenting the institutional details in the next section, we will indeed emphasise at least three motives why intermediaries' concentration might benefit the search engine's revenues. These forces are all linked to the superior technological capabilities of intermediaries to bid faster, on more keywords and with better data for targeting bids to users. Therefore, while we expect the features described in this section to imply a negative association between intermediaries' concentration and auctions' prices, a more nuanced relationship exists with the search engine's revenues. It is thus an empirical question what forces will dominate.

III Industry Background

Internet advertising is mostly subdivided into sponsored search and display advertising. Our study focuses on the former, whose basic functioning is described by the classic works of

Edelman, Ostrovsky and Schwarz [2007] and Varian [2007].[4] A basic version entails ad slots sold via auctions with advertisers that: *i)* open an account with the platform through which the search engine auctions off ad space (for instance, *Google Ads*, formerly *AdWords*) and *ii)* enter a bid, a budget and a brief ad for the keywords of interest. Each time a user queries the search engine for a keyword, an auction is run to allocate the available slots (typically up to eight) among the advertisers bidding on that keyword, if any, and to determine payments.

In recent years, however, the market has evolved, becoming more complex. Whether an internet user is querying a keyword on a search engine like Google or Bing, making online purchases on Amazon or eBay, using social media like Facebook or Twitter, or simply reading a newspaper or a blog, ads are likely to appear. "Ad exchanges" (ADX) have emerged as marketplaces connecting the demand of ad space with the supply by many, differentiated publishers.[5] Therefore, even though the search auctions remain firmly dominated by Google, advertisers often reach these auctions thorough ad exchanges. But unlike in the early days of the search auctions when advertisers directly bid, the ad exchanges – like in the typical financial exchanges – can be accessed only by qualified bidders and specialized intermediaries.

This evolution of the marketplace has gone hand in hand with technological innovations on the demand side. The typical case in our data involves at least three steps: *i)* an advertiser contracts with a digital marketing agency (DMA) the management of an internet marketing campaign, *ii)* this agency transmits to an intermediary specialized in online bidding the budget and the campaign's content decided with the advertiser; *iii)* the intermediary optimizes the bidding campaign on the ad exchange (or directly on the search engine's auction platform). The DMAs in steps *i* and *ii* are just the modern version of the traditional "Madison Avenue" agencies. The specialized intermediaries in step *iii* are, instead, more interesting in light of what discussed in the previous section. There are a few kinds of such intermediaries, but the most relevant for our study are the so called "agency trading desks" (ATDs).

[4] Display advertising entails the sale of ad space on web pages, videos and apps. The owners of these spaces (*publishers*) connect to advertisers through a *display network*. Google's *AdSense* is an example of such network. In it, advertisers select the contextual environments in which they are interested and bid to show their ad there. Auctions are used to select the winning ad (typically only one) and determine the payments. See Choi et al. [2018] for a recent, detailed review of the literature on display advertising.

[5] *Google Marketing Platform*, formerly *DoubleClick*, and Microsoft AD-ECN are examples of ad exchanges.

To understand what ATDs are, consider that there are a few thousand DMAs active in the US advertising market. Most of them, however, are not fully independent entities, but part of broader agency networks. The ATD is the centralized entity within an agency network that is responsible for bidding in the online ad auctions for all the hundreds of agencies and thousands of advertisers within the network. There are seven main networks (IPG, WPP, Publicis Groupe, Omnicom Group, Aegis-Dentsu, Havas and MDC) and each of them developed its own ATD in the last 10 years.[6] They represent the demand-side response to the incentive to improve bidding performance through better data and faster algorithms.

Indeed, bidding through ATDs differs from traditional direct bidding by advertisers in several ways that are important for this study. The key technological advancement upon which all these differences are based is the use of automated systems for bidding. This allows faster bidding, to the point that a large portion of internet ad are currently traded in real time. It also allows to more effectively use large amount of consumers' s profiling data, both internal ATD data and data purchased from third parties (*data exchanges*), possibly in real time. The need for speed and data explain why the backbone of online ad bidding takes the form of the simplified two-auction structure of the example in the previous section.[7] First, an (automated) auction within the ATD serves to select which advertisers and their relative bids to transmit to the ad exchange. Second, the pooling under one ATD of advertisers active in the same market allows the ATD to access bigger, relevant data and this can be profitable both to save on the costs charged by the *data exchanges* and to improve speed.[8]

There are at least four ways through which the more sophisticated bidding by intermediaries can affect the revenues of ad auctions. First, ATDs can select more and different keywords relative to what individual advertisers would do. Automated bidding can allow to bid on a large number of keywords and, through machine learning techniques, to select different, more refined keywords known as *long tail keywords*.[9] Second, bids and ad content

[6]The corresponding ATDs are Cadreon, Xaxis, Vivaki, Accuen, Amnet, Affiperf and Varick Media. For some networks, ATDs' names and structures evolved over time. See the web appendix for further details.

[7]A detailed discussion of this complex environment is clearly beyond the scope of this paper. See McAfee [2011] for further details.

[8]To grasp the value of higher speed, consider for instance that as page load time goes from 1 second to 3 seconds, the probability of bounce increases 32%, see `https://www.thinkwithgoogle.com/data-gallery`.

[9]These are longer, more specific keyword variations that contain at least 3-4 keywords. They tend to be

for the chosen keyword can be better optimized thanks to the superior data and the ability to experiment faster with different bids and ad text messages.

The other two aspects are both related to joint profit maximization of competing advertisers under the same ATD. On the one hand, bids can be strategically retained or, at least, reduced as the previous section explained. Relatedly, an ample set of (tacit) collusive strategies can also be implemented: bid rotation or market split (by keyword, time, geography or other user's demographics). On the other hand, there are several problems with bid optimization by individual advertiser that joint bidding can address. The literature on sponsored search has focused on externalities: for a given keyword, advertiser and slot, the number of clicks that this advertiser will receive under different configurations of the set of rivals displayed might be very different [Jeziorski and Segal, 2015]. In the context of the ad exchanges, the literature has identified problems related to limited information driving to winners' curse [McAfee, 2011] and budget constrains leading to inefficiencies [Balseiro et al., 2017]. For all these aspects, the presence of an intermediary can, under certain conditions, allow to internalize the externalities, reduce the risk of winner's curse and improve market efficiency in the presence of budget-constrained bidders.

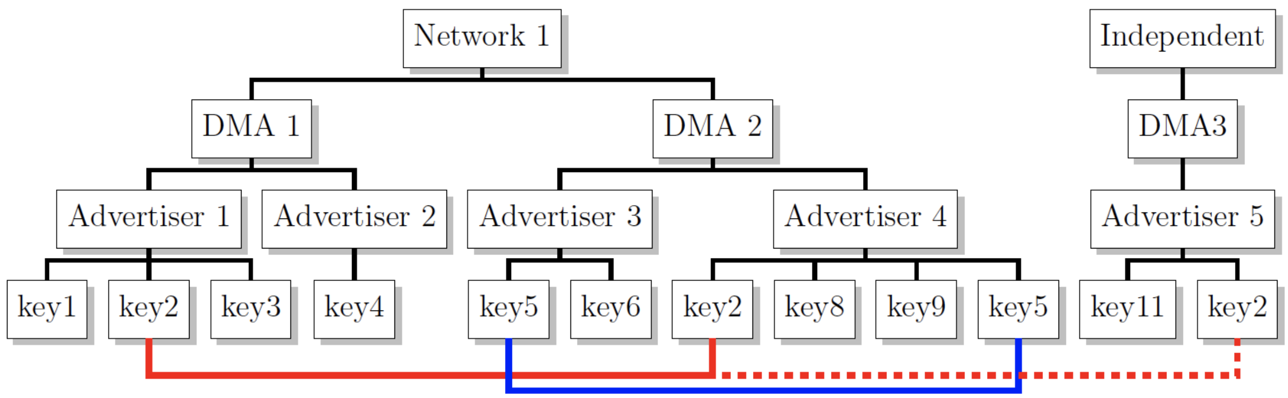
The changes in the degree of ad networks' concentration that we discuss in the next section can imply changes along all the dimensions described above. Either positive or negative effects could therefore result for the search engine. An expansion of the set of keywords on which bidding takes place might bolster revenues, but not necessarily so if the ad budget is diverted toward less competitive long tail keywords. An enlargement of the set of rival advertisers handled by the same DMA could enhance efficiencies through reduced negative externalities, but not if collusion motives cause lower bids. Determining ex ante what type of effect would prevail is clearly impossible. We therefore resort to the detailed data described below to quantify the effects of changes in ad network's concentration.

valuable in part because by being more specific they are exposed to less competition and, in part, because they are likely searched for when the user is closer to the bottom of the funnel. For instance, while an advertiser might bid for "charity donations," an ATD might bid on thousands of more specific variants, one of which could be "charity donations forniture pickup." This latter keyword is associated with a donation that is rather likely to happen as the query involves "forniture pickup." Long tail keywords have become more common over time and now account for roughly 70% of all online searches.

IV Data

The minimal data requirements to test the effects of bidders' concentration on the search engine's revenues are information on: *i)* the advertisers' affiliation to intermediaries, *ii)* the set of keywords on which they bid and *iii)* the associated average CPC of these keywords. Our analysis is based on a new dataset that contains all this information, and more.

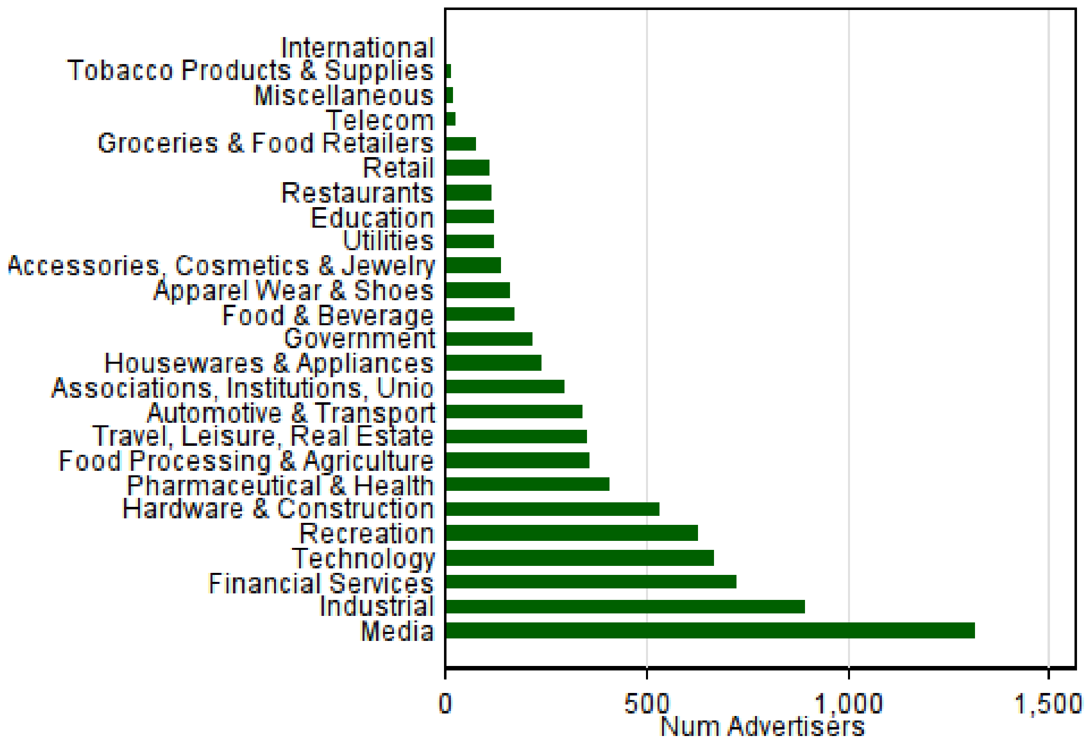
Figure 2: Redbooks-SEMrush Data Structure



Notes: Hierarchical structure of the data: keywords (SEMrush), advertisers (Redbooks/SEMrush), agencies and networks (Redbooks). Solid lines represent examples of coalitions: within DMA (blue) and network (red).

From Redbooks, a comprehensive database on marketing agencies, we obtained a list of advertisers representing nearly the universe of major US firms active in online marketing. For each of these advertisers, the Redbook data gives us the full list of marketing agencies (both DMAs and traditional marketing agencies) affiliated with them. The data is yearly for the period 2012-2017 and covers around 6,000 advertisers (i.e., web domains) per year active in all sectors of the economy. For the years 2014 and 2017 only, we also have access to a linkage variable that relates each individual agency to its agency network, if any. Given the ATDs' role, this letter variable is crucial: we will consider the networks as the intermediaries and only for the DMAs outside the networks we take them as the intermediary. This leaves us with seven agency networks and about a thousand independent agencies. For the remaining advertisers with no agency affiliation, we consider them as bidding autonomously. We define all advertisers linked to networks as *network advertisers*. The different cases of agencies with and without network are illustrated by the top two rows of Figure 2. Figure 3, instead, indicates the main industry for each advertiser: 25 different industries are represented, with the three largest ones being media, industrial and financial services.[10]

Figure 3: Number of Advertisers per Industry: Redbooks data



Notes: Number of unique advertisers per industry according to Redbooks data. In the above panel we report the original data, in below panel we report the original data (green bars) and the imputed industries (red bars). After running the matching algorithm, 1,800 advertisers still have a missing industry.

We combine the intermediaries data with sponsored search data from SEMrush. For each keyword on Google, we know which advertisers appeared in the sponsored ad slots and when. We also know the position of the ad in the search outcome page, the volume of searches associated with the keyword (i.e., the average number of search queries for the given keyword in the last 12 months); the visible URL of the ad; the content of the ad; and, most importantly, the keyword-specific average CPC (i.e., the price advertisers pay for a user's click on an ad triggered by the given keyword).[11] Therefore, as shown in the bottom rows of

[10]Since for a third of the advertisers Redbooks does not contain the information on industry affiliation, we impute the industry by using the keyword data from SEMrush. An advertiser will be assigned to the industry with which it shares most keywords.

[11]Since the Redbook data are a snapshot of the agency affiliation during the first half of January each year, er also take the January data for SEMrush.

Figure 2, through the advertisers' identity we can link the agency network data to the data on the individual keywords.

Table 1: Summary Statistics: Keywords, Networks and Markets

Panel A. Statistics by Keywords

	Network Advertisers				Non-network Advertisers			
	Mean	SD	Median	Obs	Mean	SD	Median	Obs
Cost-per-click	2.34	0.90	5.80	15,353,369	2.39	0.89	6.11	21,565,995
Volume	497	40	34,949	15,353,369	362	40	99,950	21,565,995
Traffic	0.01	0.00	0.52	15,353,369	0.06	0.00	1.27	21,565,993
Competition	0.58	0.69	0.39	15,353,369	0.59	0.73	0.39	21,565,995
Num of Advertisers	1.30	1.00	0.68	15,353,369	1.21	1.00	0.52	21,565,995
Coalition	0.15	0.00	0.36	15,353,369	0.00	0.00	0.00	21,565,995
Coalition Size	2.38	2.00	0.69	327,809	-	-	-	-

Panel B. Statistics by Network

	Search Volume Share				Keyword Share			
	2014	2015	2016	2017	2014	2015	2016	2017
IPG	0.18	0.16	0.18	0.16	0.09	0.12	0.12	0.14
WPP	0.14	0.13	0.15	0.13	0.15	0.14	0.13	0.12
Omnicom	0.14	0.16	0.14	0.16	0.10	0.10	0.12	0.13
Publicis	0.12	0.10	0.11	0.13	0.10	0.11	0.10	0.09
MDC	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.07
Havas	0.04	0.06	0.05	0.04	0.04	0.05	0.04	0.02
Dentsu-Aegis	0.04	0.07	0.09	0.06	0.05	0.06	0.07	0.07
Indep Age	0.11	0.08	0.07	0.07	0.16	0.14	0.13	0.13
Indep Adv	0.16	0.17	0.15	0.18	0.25	0.24	0.22	0.23

Table 1 presents summary statistics, by keyword (panel A) and by network (panel B). In the left columns of panel A, we report the statistics for the network advertisers. We observe them across 15 million keywords during the sample period. In the right columns of panel A, we report analogous statistics for the non-network advertisers. For both groups, we see a similar CPC; although the mean and median CPC is lower for the network bidders. In terms of volume, for both groups the substantially lower value of the mean relative to the median indicates a tendency to bid on many keywords that are infrequently searched. Traffic indicates what share of the advertiser's traffic can be associated with the specific keyword. Thus the lower value of 1% observed for the network advertisers relative to the 6% of the non network advertisers is compatible with the former placing ad over more keywords.

Competition is a measure, ranging from 0 to 1, assigned by the data provider to score how easy it is to raise to the top slot for the given keyword. Its value is rather similar for the two groups and this despite the next variable, number of bidders, indicates that often there is a slightly higher number of participants in the keywords on which the networks advertisers bid. Coalition measures the number of keywords where more than one of the ad shown belongs to different advertisers represented by the same agency network. Within this subset of cases, the following variable shows that the average coalition size is 2.38 advertisers.[12]

Panel B, cuts the data in a different dimension. It aims to show the relative size of each one of the seven networks, both in terms of the volume of the advertisers covered and in terms of the number of keywords. If we consider just the four largest, the big four as they are often referred to (WPP, Omnicom, Publicis Groupe, and Interpublic Group of Companies), their combined market share denotes a very high degree of concentration. This is the case regardless of whether we measure it in terms of volume of traffic generated by the keywords (left columns) or straight keywords (right columns). To conclude the data description, we note that a key element of our analysis will be the changes in network concentration. These will mostly be driven by the networks' activity of DMA acquisitions. As an example of this phenomenon, consider the case of what used to be up until June 2016 the largest non-network affiliated DMA, Merkle. The only activity of Merkle was digital marketing. In July 2016, Aegis-Dentsu acquired Merkle for $1.5 billion dollars, and, at that time, many of Merkle's clients were bidding on the same keywords as some of Aegis-Dentsu's advertisers. For instance, in the electronics sector, *Dell* and *Samsung* were in Merkle's portfolio, placing bids on keywords also targeted by Aegis-Dentsu's clients *Apple*, *HP*, *IBM/Lenovo* and *Intel*. Other examples include: in the financial sector, Merkle's *Lending Tree* and *Metlife* were bidding in auctions alongside Aegis-Dentsu's *Capitalone, Discover, Fidelity, Equifax, JP Morgan-Chase*; for car manufacturers, Merkle's *FIAT-Chrysler* and *Mercedes-Benz USA* bid

[12]Although the vast majority of cases involves coalitions of size 2, there are a few examples of larger coalitions. For instance, for the keyword "online banking" there are four advertisers (Bank of America, Travelers, Geico and State Farm) all affiliated with a single DMA (The Martin Agency, a major marketing agency that is also part of the Interpublic Group of Companies). Furthermore, not reported in the table is another interesting statistic that shows that there is typically a single multi-advertiser DMA per keyword. This is consistent with the agencies' specialization and helps explaining why the phenomenon of common DMA appears to be expanding over time.

alongside Aegis-Dentsu's *Toyota, Volkswagen, Subaru*; in phone services, Merkle's *Vonage* bid alongside Aegis-Dentsu's *T-Mobile*.[13] This acquisition therefore increased the potential for coordinated bidding.

The above argument is, however, still imprecise. While it is intuitively true that *Dell* and *Lenovo* are competing firms, understanding the effects of increased network concentration requires a more precise definition of the markets. Ideally, we would like to apply the same type of approach to market definition that is used in antitrust analysis. Nevertheless, we both lack the type of data that such an analysis would require and we also have such a large number of firms that this type of detailed analysis would be unfeasible. We therefore resort to an alternative approach based on machine learning methods that we detail below.

V Market Definition Via Thematic Clustering

Advertisers' industries are too aggregated to be treated as the relevant markets, but the individual keywords are too disaggregated to serve this role.[14] Our solution to find a useful middle-ground is to apply state of the art natural language processing methods to form keyword clusters interpretable as markets. There are two main steps in our approach described below: first, an (unsupervised learning) algorithm is used to represent keywords as numerical vectors; second, these vectorialized keywords are then aggregated into clusters.

A key element for the first step is the availability of a *corpus* (i.e., body of text) on which the algorithm learns the association between words. Given the goal of identifying relevant markets within the online advertisement industry, the ideal corpus should be informative on how consumers find products and services online. With such corpus, the approach described below can mimic what is sometimes done in antitrust cases by surveying consumers about what products they see as living in the same product space. Without aiming for the same accuracy required for competition cases, we nevertheless see this approach as a valuable

[13] Source: Redbook data fro January 2016.

[14] The excessive aggregation of industry data for competition analysis has been clearly discussed in the recent work of Werden and Froeb [2018].

contribution.[15] We first discuss its details and then discuss some of its limitations.

Step 1 - Keyword vectorization For each keyword appearing in our SEMrush data, we need a vector representation. The motive is simple: "red car," "blue car" and "automobile" are three keywords that we would like to see grouped together. But using keywords directly, only "red car" and "blue car" will be pooled together. The vector representation systems developed in natural language processing solve this type of problem. We use an unsupervised learning algorithm (GloVe, by Pennington, Socher and Manning [2014]) to obtain vector representations for each word within the keywords. The GloVe model builds on the classical matrix of word co-occurrences in a corpus - i.e., a sparse matrix with one row per document in the corpus, and one column per word, populated with the number of occurrences. In particular, the novelty of the GloVe model with respect to previous approaches is that it combines the benefits of matrix factorization approaches - i.e., reduce the dimensionality of co-occurrence matrices - with the good performances of skip-gram models (like Google's *word2vec*) in word analogy tasks.

We use a GloVe dataset pre-trained on 840B documents, corresponding to ≈ 2.2M unique terms, from Common Crawl in English, featuring 300 dimensions. Using as corpus such an extensive body of text which originates from mimicking the web crawling behavior of typical internet users is what makes the resulting vectorization analogous to surveying people about proximity between keywords.[16] Hence, when applied to the sponsored search keywords in our data, the vectorization shall identify which products and services are related to each other.

[15] The definition of the relevant market is one of the main challenges in the applied Industrial Organization literature and competition policy. The size and characteristics of the markets have dramatic effects on estimated parameters, and usually there is no ex-ante market definition available and commonly accepted. Ideally, we would like our market definition to be suitable for an anti-trust analysis: in competition law, in fact, the relevant market is the market in which a particular product or service is sold. In turn, it is the intersection of a *relevant product market* and a *relevant geographic market*. The European Commission states that the former "comprises all those products and/or services which are regarded as interchangeable or substitutable by the consumer by reason of the products' characteristics, their prices and their intended use", whereas the latter "comprises the area in which the firms concerned are involved in the supply of products or services and in which the conditions of competition are sufficiently homogeneous".

[16] The dataset, and GloVe code, are open source and freely available at https://nlp.stanford.edu/projects/glove/. There is a number of other datasets available (e.g., trained on Wikipedia, on Twitter, with 25, 50, 100 or 200 dimensions, etc.): we plan to re-run all our analyses using these datasets. The Common Crawl project gathers high-quality crawled data and make them freely available for analyses and research purposes at http://commoncrawl.org/big-picture/what-we-do/.

To use the GloVe vectors, we split every keyword in its constituent *terms* and merge every word with GloVe vectors.[17] Finally, we obtain the vector representation of each keyword by summing together the vectors relative to all its underlying *terms*.

Step 2 - Text clustering at the keyword level We perform this step within each one of the 24 markets in which the advertisers are partitioned in the Redbooks data.[18] In particular, take the vector representation of all the keywords belonging to all the advertisers within an industry and apply a clustering algorithm to aggregate them into what we will treat as markets. More in details, consider the vector representation of K kewywords:

$$\begin{aligned}\overrightarrow{\boldsymbol{d_1}} &= (w_{1,1}, w_{2,1}, ..., w_{T,1}),\\ &\vdots\\ \overrightarrow{\boldsymbol{d_k}} &= (w_{1,k}, w_{2,k}, ..., w_{T,k}),\\ &\vdots\\ \overrightarrow{\boldsymbol{d_K}} &= (w_{1,K}, w_{2,K}, ..., w_{T,K}),\end{aligned}$$

where T is the dimension of the vector space used by GloVe, or 300. Using the deviation of the angles between each document vector, it is possible to measure their *distance*: in figure 4 we plot an example two-dimensional vector space (given by two terms, $t1$ and $t2$) with two vectors $\overrightarrow{\boldsymbol{d_1}}$ and $\overrightarrow{\boldsymbol{d_2}}$, and their angular distance θ.

In practice, to compute the distance it is easier to use the *cosine* of the angle instead of the angle itself. In particular, in the example:

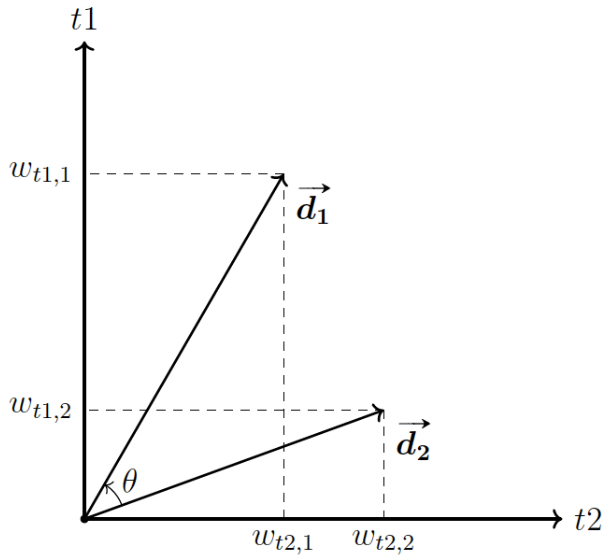
$$\cos\theta = \frac{\overrightarrow{\boldsymbol{d_1}} \cdot \overrightarrow{\boldsymbol{d_2}}}{||\overrightarrow{\boldsymbol{d_1}}||\;||\overrightarrow{\boldsymbol{d_2}}||}, \tag{2}$$

where $\overrightarrow{\boldsymbol{d_1}} \cdot \overrightarrow{\boldsymbol{d_2}}$ is the intersection between the vectorized documents and $||\cdot||$ is the norm operator, and reads $||\overrightarrow{\boldsymbol{d_1}}|| = \sqrt{\sum_{i=1}^{n} d_{1,i}^2}$. Since all vector components are nonnegative by

[17]In the total sample, the share of unique words merged with GloVe data is $\approx 80\%$.

[18]There are 25 industries, but we ignore the "Miscellaneous" one.

Figure 4: Distance



Notes: Angular distance θ between vectors $\overrightarrow{\boldsymbol{d_1}}$ and $\overrightarrow{\boldsymbol{d_2}}$ in a two-dimensional vector space.

definition, $cos\theta = 0$ if $\overrightarrow{\boldsymbol{d_1}} \perp \overrightarrow{\boldsymbol{d_2}}$. To sum up, for each document pair d_i, d_j we compute the cosine similarity measure:

$$sim(d_i, d_j) = \frac{\overrightarrow{\boldsymbol{d_i}} \cdot \overrightarrow{\boldsymbol{d_j}}}{||\overrightarrow{\boldsymbol{d_i}}||\ ||\overrightarrow{\boldsymbol{d_j}}||} = \frac{\sum_{k=1}^{n} w_{ki} w_{kj}}{\sqrt{\sum_{k=1}^{n} w_{ki}^2}\sqrt{\sum_{k=1}^{n} w_{kj}^2}}. \tag{3}$$

Finally, we run a spherical k-means clustering algorithm (see Dhillon and Modha [2001]) on the cosine distance matrix with 10,000 centroids on different samples - i.e., using the first 100 up to the first 600 keywords per advertiser. In Table 2, for each sample (col 1), we report the average number of keywords per cluster (col 3), the average number of words per keyword (col 4) and the size and average number of keywords relative to the biggest cluster (col 5 and 6, respectively). Due to the nature of the keyword searched (i.e., very short or single-word documents), the spherical k-means algorithm yields a single, large cluster (the *cluster 0*) containing keywords that are all "singletons" or orthogonal to one another. As expected, the absolute size of cluster 0 increases with N, while its keywords are shorter.

Testing the quality of the clusters obtained requires a reference sample where keywords and markets are correctly associated. Lacking this type of sample, we resorted to random

Table 2: Descriptives on k-means results

First Keywords	# Clusters	Avg # keywords	Avg # Words	Cluster 0 # keywords	Cluster 0 # Words	N
100	10,000	48.51	3.73	30,472	2.12	409,322
200	10,000	68.32	3.79	54,555	2.22	630,921
300	10,000	80.15	3.82	69,020	2.21	742,162
400	10,000	95.95	3.80	78,393	2.29	815,341
500	10,000	110.16	3.81	98,271	2.31	900,127
600	10,000	121.67	3.81	101,032	2.95	990,361

inspection of the clusters. We find overall very satisfactory results with our initial motivating concern of related but different keywords (like car and automobile) systematically pooled together. We plan to sharpen the assessment of the clusters along two dimensions: i) using a corpus constructed from past EU and US competition cases to partition keywords according to proper antitrust assessments of the relevant markets; ii) using survey systems (like Amazon's Mechanical Turk) to test the clusters as done in the natural language processing literature, for instance by counting the frequency with which a random word added to the cluster is correctly identified as the foreign object by a human inspecting the keywords in the cluster.

In doing this testing, we also plan to assess the degree to which our approach suffers from two problems that are intrinsic to our method. The first is that both substitute and complement products/services are likely to be pooled together. To the extent that, despite the complementarity between products, the advertisers are in competition for the limited ad space, our analysis would not be distorted. But the existence of joint marketing efforts by advertisers of complementary products requires further attention. Similarly, our keyword-based method is able to identify different geographical markets only to the extent that the geographical aspect is particularly salient in the keywords (and in the training corpus). Visual inspection of the clusters reveals that this is only sometimes the case (like "car rental Boston" and "car rental New York" being sometimes pooled together). Depending on the outcome of the cluster testing, we will consider how to enhance the relevance of the geographical terms in the keywords so to better partition geographical markets.

VI Empirical Strategy

Having defined markets, we now study whether increases of intermediaries' concentration is associated with changes in Google's revenues. We begin by describing the main variables used and then present the empirical strategy.

A. Outcome Variables - Suppose that the clustering procedure has identified M markets, $m = 1, ..., M$. Denote as K_m the set of k keywords in market m. We can use our keyword-level data to construct a measure of Google's search revenues in market m at period t by aggregating revenues over keywords:

$$R_{mt} = \sum_{k \in K_m} CPC_{kmt} * Volume_{kmt} * CTR_{kmt} \tag{4}$$

where CPC_{kmt} is the average Cost-per-Click of keyword k of market m at time t, $Volume_{kmt}$ is its overall number of searches of k over an year and CTR_{kmt} is the cumulative Click-through-Rate of all the sponsored ad slots shown for keyword k.[19]

There is substantial heterogeneity in the levels of revenues across markets, mostly driven by heterogeneity in volume and CPC. To perform a meaningful analysis of the association of the revenue's level and the level of concentration, we should successfully standardize markets. Rather than attempting this route, we focus on the growth rate of revenues for the same keywords in a specific market over time and analyze how this relates to the intermediaries' concentration in the market. By focusing on growth, we also avoid concerns related to time-invariant unobservable differences in the characteristics of the keywords that may be correlated with revenue levels. We will also control for the influence of time-varying features at the level of keywords by using information for the organic results of the query.

Other outcomes that will be interesting to explore in addition to the revenue measure R are the average CPC and Volume of keywords, as well as the total number of keywords,

[19] More specifically, for each k, the overall CTR_k is the cumulative sum of all the ad slots j appearing on the search outcomes page of keyword k: $CTR_k = \sum_{j \in k} CTR_j$. We use the click-through rate typical in the industry for each advertisers' position, but we are in the process of obtaining more accurate CTR data.

possibly separating regular vs. long tail and branded vs. non-branded. These additional outcomes allow us to explore the channels through which concentration affects the search engine revenue and they can offer direct evidence of the different types of bidding strategies discussed earlier.

B. Concentration Measure - We measure intermediaries' concentration using an analogue of the well known Herfindahl-Hirschman Index (HHI) for the sponsored search. Suppose that we have a market m composed by the set K_m of keywords. For each keyword $k \in K_m$, there are sponsored ad slots j, each occupied by an advertiser a. Each of these slots brings a certain number of clicks, which are ultimately the advertisers' object of interest. We therefore measure the "market size" (S_{mt}) as the sum of all the clicks of all the ad slots allocated in all the keywords in the market. That is: $S_{mt} = \sum_{k \in K_m} Volume_{kmt} * CTR_{kmt}$.[20] The intermediaries' concentration is measured accordingly by summing together all the clicks of all the market keywords associated with the slots occupied by each of the advertisers that the intermediary represents. That is, for intermediary i, representing the set of advertisers A_i, the market share in market m at time t is:

$$s^i_{mt} = \frac{1}{S_{mt}} \sum_{a \in A_i} \sum_{k \in m} \sum_{j \in k} CTR_{jkmt} * Volume_{kmt} * 1\{a \text{ occupies } j \in k\}. \tag{5}$$

Thus, our concentration measure for market m at time t is the squared sum of each intermediary's market share, or: $HHI_{mt} = \sum_{i=1}^{I} ({s^i}_{mt})^2$.[21] As discussed earlier, we consider as intermediary the agency network, or, if not present, the DMA, or, if also this is not present, the individual advertiser.

[20] An alternative definition could involve using an arbitrary fixed number of slots (say eight) and then consider as the market size all the clicks potentially associated with all the eight slots in all keywords. Aside from the fact that the total number of slots admissible across Google's auctions is neither fixed nor observable to us, this alternative definition could lead to find little concentration in markets where many keywords have only one ad appearing. But this does not seem idea as for these keyword the concentration of clicks in the hands of the advertisers that are shown is clearly very high.

[21] Despite several theoretical and practical shortcomings of this measure, it is commonly used in both academia and competition policy to proxy for concentration. See, among others Hastings and Gilbert [2005], Dafny, Duggan and Ramanarayanan [2012] and the US Horizontal Merger Guidelines.

C. Empirical strategy In an ideal setting, the following OLS regression of revenue growth on the level of concentration would reveal the causal effects of concentration. In practice, it is not possible to assign a causal interpretation to this conditional correlation. For instance, a keyword might have become suddenly fashionable for some exogenous reasons; advertisers that were previously little interested in this keyword now hire an intermediary to bid for this keyword; they all hire the same intermediary as it is the one specialized in the market to which the keyword belongs. This situation would likely induce observing a positive association between intermediaries' concentration and the search engine's revenues, but this does not imply the existence of a causal relationship between the two phenomena. We, nevertheless, report below this OLS regression which will represent our starting point in the analysis:

$$\Delta ln(R)_{mt} = \beta HHI_{mt} + \phi X_{mt} + \tau_t + \gamma_z + \epsilon_{mt}. \tag{6}$$

The dependent variable is the change in revenues in market m (of industry z) between t and $t-1$. This regression controls for year (τ_t) and industy (γ_z) fixed effects, as well as for characteristics of the market-time included in X_{mt} (we use the number of organic links, as well as experimenting with other variables constructed from the information on the organic links associated with the market-time keywords).

To deal with the issue of causality, we use an IV strategy inspired by that of Dafny, Duggan and Ramanarayanan [2012] to study the effects of health insurers' concentration on insurance premiums. That is, we use changes in the market structure originating from mergers and acquisitions (M&A) between intermediaries as a source of exogenous shock to the local (i.e., market) concentration. The idea is that M&A operations between intermediaries, especially the larger ones are unlikely to be driven by the expectation of how the CPC would evolve in specific markets as a consequence of the merger. Since two merging intermediaries might have clients in a plethora of markets with possibly quite different starting levels of concentration, then the M&A operation generates useful variation in these market's HHI. More specifically, for each market-time we compute the "simulated change in HHI" ($sim\Delta HHI_{mt}$) as the counterfactual HHI change in market m at time t *induced* by the merger, absent any other changes. Hence, if in year t intermediary α merges with intermediary β, the merger-induced change in market m's HHI is:

$$sim\Delta HHI_m = \underbrace{(s^{\alpha}_{m,t-1} + s^{\beta}_{m,t-1})^2}_{\text{Share of merged firm } \alpha+\beta} - \underbrace{((s^{\alpha}_{m,t-1})^2 + (s^{\beta}_{m,t-1})^2)}_{\text{Sum of single firm } \alpha \text{ and } \beta \text{ shares}} = 2s^{\alpha}_{m,t-1}s^{\beta}_{m,t-1} \tag{7}$$

We use, for each market-year, the variable $sim\Delta HHI_{mt}$ as instrument for HHI_{mt}. Our first-stage, hence, takes the form of the following market-time level regression:

$$HHI_{mt} = \beta^{FS} sim\Delta HHI_{mt} + \phi X_{mt} + \tau_t + \gamma_z + \epsilon_{mt}, \tag{8}$$

where the set of covariates entering the regression along with $sim\Delta HHI_{mt}$ is the same of those in equation (6). What sign to expect on the estimate of β^{FS} is an interesting question. To the extent that there is persistency in the market shares, we would expect a positive sign and this is indeed the finding of Dafny, Duggan and Ramanarayanan [2012] for the US health insurance market. Nevertheless, it would not be unreasonable to see a negative sign in case a merger between intermediaries leads some clients to leave the new entity to avoid sharing a marketing resource with rivals (i.e., avoiding "sleeping with the enemy" Villas-Boas [1994]).

As discussed below, we estimate a positive sign for β^{FS}. This is in line with the companion work of Decarolis, Goldmanis and Penta [2018] where a discrete choice framework is used to model the advertisers' choice of intermediary and the findings reveal that an advertiser is more likely to select an intermediary the more of its rivals use the same intermediary.[22]

The next step of the IV strategy is the reduced-form regression, which is:

$$\Delta log(R)_{mt} = \beta^{RF} sim\Delta HHI_{mt} + \phi X_{mt} + \tau_t + \gamma_z + \epsilon_{mt}. \tag{9}$$

The proposed instrument captures large variations in the degree of concentration and hence should be relevant for $\Delta log(R)_{mt}$. This is what the test presented below will indeed confirm. Furthermore, while not directly testable the exclusion restriction requiring that all the effects of $sim\Delta HHI_{mt}$ on $\Delta log(R)_{mt}$ pass through HHI_{mt} is plausibly satisfied given the nature

[22]This is also reassuring because, in the presence of treatment effects that are heterogeneous across markets and with market's sorting into M&A based on the gain from merging, monotonicity is a necessary condition for the identification of the Local Average Treatment Effect (LATE), Angrist and Imbens [1995].

of the chosen instrument which is itself constructed as a function of the lagged HHI. Even more that in the case of the first-stage, the sign of β^{RF} is an open empirical question for the same reasons why it is ambiguous ex ante what effects on revenues the intermediaries concentration has.

VII Results

We begin the illustration of our results from the first-stage and reduced-form estimates in Table 3. In panel a, we report the estimates obtained from a sample in which, for each advertiser, we take at most the top 500 keyword (in terms of its traffic). In panel b, the threshold is set lower at 100 keywords.[23] The idea is that assessing through these two samples to what extent it might matter to include a broader, but likely less salient set of keywords. As revealed by the estimates in Table 3, there are indeed some differences between the two samples, but the broad sense of the estimates is the same: for both samples, there is a positive and significant effect estimated for the first-stage regression of HHI on $sim\Delta HHI$ and there is a negative and significant effect for the reduced-form regression of $\Delta log(R)$ on $sim\Delta HHI$. In essentially all cases the effects are significant and more so for the larger sample of 500 keywords where the clustering procedure allows to identify 14,212 market-year observations, relative to the 8,431 of the 100 keyword sample.

The positive sign of the $sim\Delta HHI$ estimate in the first stage regression is a confirmation from what was already indicated by the summary statistics in panel B of Table 1. The persistency in the shares of traffic volumes and keywords that we showed there explain this positive coefficient which is also in line with the discrete choice estimates of Decarolis, Goldmanis and Penta [2018] mentioned above. Regarding the reduced form, instead, we find consistent evidence of a negative relationship between the change in revenues and the simulated change in HHI. Taken together, these regression already inform us that the IV estimates will indicate a negative impact of intermediaries' concentration on the search engine's revenues' growth.

This is indeed what we observe in Table 4, where we report the OLS estimates (first three

[23]Both cases use 10,000 clusters for the k-means clustering algorithm.

Table 3: First-Stage andReduced-Form Estimates

Panel a) Top 500 Keyword per Advertiser

	(1) FS	(2) RF	(3) FS	(4) RF	(5) FS	(6) RF
$sim\Delta HHI$	0.978***	-0.271***	0.821***	-0.264***	0.902***	-0.231***
	(0.219)	(0.105)	(0.198)	(0.100)	(0.197)	(0.094)
Obs.	14,212	14,212	14,212	14,212	14,212	14,212
F_{IV}	39.72		40.11		38.31	

Panel b) Top 100 Keyword per Advertiser

	(1) FS	(2) RF	(3) FS	(4) RF	(5) FS	(6) RF
$sim\Delta HHI$	1.021***	-0.171	0.957***	-0.220**	0.903**	-0.214**
	(0.355)	(0.186)	(0.372)	(0.131)	(0.381)	(0.126)
Obs.	8,431	8,431	8,431	8,431	8,431	8,431
F_{IV}	24.18		32.28		26.12	
Industry FE			✓	✓	✓	✓
Year FE					✓	✓

Notes: Standard errors are clustered by industry-year. Clusters with less than 10 (or more than 10,000) keywords excluded.

columns) and IV estimates (last three columns). Both OLS and IV coefficients are systematically estimated to have negative signs, albeit for the 100 keyword sample the estimate is not significant in the model specification that does not include fixed effects for industry (and year). The magnitude of the IV estimates is smaller than that of the corresponding OLS ones. Including the industry fixed effects modifies the magnitude of the estimates, increasing their absolute value. This is indicative of the heterogeneity existing between industries. Indeed, in results not reported here, when we repeat the analysis industry-by-industry for all the 24 industries of the Redbook advertisers we find that the effect of intermediaries' concentration on revenues ranging from being zero to being large and negative (up to six times the baseline estimates in Table 4 for the Food-&-Beverage industry).

The findings above indicate that the effects of increased buyer power seem to dominate any efficiency gain from which the search engine might benefit. To better understand our findings, we explore next the association between intermediaries' concentration and CPC, volumes, positions and the number of keywords. Although there is quite substantial heterogeneity across industries, CPC and the number of keywords are two revenue components whose conditional correlation with concentration tends to be negative for most industries.

There are a number of robustness exercises and further extensions that might be inter-

Table 4: OLS and IV Baseline Estimates

Panel a) Top 500 Keyword per Advertiser

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
HHI	-0.397***	-0.376***	-0.353***	-0.277***	-0.322***	-0.256***
	(0.091)	(0.087)	(0.089)	(0.117)	(0.128)	(0.113)
Obs.	14,212	14,212	14,212	14,212	14,212	14,212

Panel b) Top 100 Keyword per Advertiser

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
HHI	-0.388***	-0.365***	-0.351***	-0.167	-0.230**	-0.237**
	(0.075)	(0.072)	(0.072)	(0.179)	(0.131)	(0.136)
Obs.	8,431	8,431	8,431	8,431	8,431	8,431
Industry FE		✓	✓		✓	✓
Year FE			✓			✓

Notes: Standard errors are clustered by industry-year. Clusters with less than 10 (or more than 10,000) keywords excluded.

esting to explore in future versions of this study. We list here a few. Regarding robustness, aside from the obvious sensitivity checks on the clusters' size and the choice of top keywords to include, exploring features of the instrument seems the most relevant aspect. In particular, it would be interesting to separate the various M&A episodes to distinguish the ones operating across many markets from the more local ones where our key identification assumption is less likely to hold. Relatedly, a placebo test could be performed exploiting the failed attempt to merge between Publicis and Omnicom. This merger, announced on July 2013, but called off on May 2014, might serve to find out whether it is indeed only though consumed mergers that data and algorithms are pooled to create the effects that we observe.

In terms of extensions, there are two main avenues that could be relevant to explore with our data and approach. First, to disentangle better the market dynamics, it could be interesting to assess how intermediaries respond to their competitors' mergers. There are non-trivial theoretical implications of what competition between different intermediaries' coalitions of advertisers can imply for the functioning of Google's GSP auction (Decarolis, Goldmanis and Penta [2017]'s model predicts cycling patterns in bidding). But the extent to which the rise in buyers' power is effective in reducing the search engine's revenues crucially hinges on how the increased intermediaries' concentration eases competition among intermediaries, not just among advertisers within the intermediaries. Second, an important channel through which the intermediaries might curtail Google's economic power is by making it easier to divert ad budgets toward other platforms. Microsoft Bing is Google's main competitor in the search auctions and we have already collected data from its sponsored and organic search results. While information on the CPC is missing in these data, analyzing effects in terms of keyword number and search volumes might nevertheless be informative of how intermediaries's behavior affects competition between search engines.

VIII Conclusions

The (still preliminary) estimates indicate that concentration among the intermediaries bidding on behalf of advertisers in the search auctions negatively and significantly impacts the growth of the search engine's revenues. Despite the potential benefits for the search engine from the increased efficiency that intermediaries bring to the market along many dimensions, especially through enhanced speed and better data, the negative revenue result is indicative of the intermediaries capability to reduce the average prices. This is a novel insight on what is currently one of the largest and fastest growing advertising markets. In a period of increasing attention from the academic and policy worlds about industry concentration, our study reveals that technological innovation and countervailing power pose a limit to the economic power attainable by dominant firms, like Google is in the sponsored search.

Numerous questions are left open for future research. First, through the lens of a structural demand and supply model of advertising, it might be interesting to estimate to what extent Google might contrast the effects of intermediaries' concentration by revising its auction's reserve price policy. The optimal choice of the reserve price is a cornerstone of the mechanism design and market design approaches to auction markets. Rising the reserve price is the microeconmic's textbook answer to fighting bidders' collusion. It is therefore interesting that recently (May 2017) Google has opted to increase the reserve prices applied in its search auctions. A structural model might thus be helpful to assess the effectiveness of this policy and to evaluate how different choices of reserve prices might affect outcomes.

Second, it would be important to understand the consequences of intermediaries' concentration for consumers. On the one hand, a transfer of revenues from Google toward the more competitive industries where advertisers operate might induce a pass-through of some of the benefits to consumers. On the other hand, it is unclear to what extent a decrease in Google's revenues might worsen the quality of services that consumers attain on its search engine, or through its other services. More than thirty years after the breakup of the Bell System in 1982 it is still an open and key question how should an economist look at the dominance of companies like Google, or the Bell System.

References

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90(430): 431–442.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "Concentrating on the Fall of the Labor Share." *American Economic Review*, 107(5): 180–85.

Balseiro, Santiago, Max Lin, Vahab Mirrokni, Renato Leme, and IIIS Song Zuo. 2017. "Dynamic Revenue Sharing." In *Advances in Neural Information Processing Systems 30.* , ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, 2681–2689. Curran Associates, Inc.

CEA. 2016. "Benefits of Competition and Indicators of Market Power." Council of Economic Advisers 20160414.

Choi, Hana, Carl F. Mela, Santiago Balseiro, and Adam Leary. 2018. "Online Display Advertising Markets: A Literature Review and Future Directions." National Bureau of Economic Research Working Paper 18-1.

Dafny, Leemore, Mark Duggan, and Subramaniam Ramanarayanan. 2012. "Paying a Premium on Your Premium? Consolidation in the US Health Insurance Industry." *American Economic Review*, 102(2): 1161–85.

Decarolis, Francesco, Maris Goldmanis, and Antonio Penta. 2017. "Marketing Agencies and Collusive Bidding in Online Ad Auctions." National Bureau of Economic Research Working Paper 23962.

Decarolis, Francesco, Maris Goldmanis, and Antonio Penta. 2018. "Advertisers' Network Formation: Evidence from the Search Auctions."

De Loecker, Jan, and Jan Eeckhout. 2017. "The Rise of Market Power and the Macroeconomic Implications." National Bureau of Economic Research Working Paper 23687.

Dhillon, Inderjit S, and Dharmendra S Modha. 2001. "Concept decompositions for large sparse text data using clustering." *Machine learning*, 42(1-2): 143–175.

Economist. 2016*a*. "The rise of the superstars." 26 March.

Economist, The. 2016*b*. "Too much of a good thing." 17 September.

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwarz. 2007. "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords." *American Economic Review*, 97(1): 242–259.

eMarketer. 2018. "US Media and Entertainment Industries StatPack 2018: Digital Ad Spending Forecast and Trends."

Galbraith, John Kenneth. 1952. *American capitalism: the concept of countervailing power.* Houghton Mifflin, Massachusetts.

Gutierrez, German, and Thomas Philippon. 2017. "Declining Competition and Investment in the U.S." National Bureau of Economic Research Working Paper 23583.

Hastings, Justine S., and Richard J. Gilbert. 2005. "Market Power, Vertical Integration and the Wholesale Price of Gasoline." *The Journal of Industrial Economics*, 53(4): 469–492.

IAB. 2018. "Internet Advertising Revenue Report, Full Year 2017."

Jeziorski, Przemyslaw, and Ilya Segal. 2015. "What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising." *American Economic Journal: Microeconomics*, 7(3): 24–53.

Mansour, Yishay, S. Muthukrishnan, and Noam Nisan. 2012. "Doubleclick Ad Exchange Auction." *Computing Research Repository*, abs/1204.0535.

McAfee, R. Preston. 2011. "The Design of Advertising Exchanges." *Review of Industrial Organization*, 39(3): 169–185.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." 1532–1543.

Stiglitz, Joseph. 2016. "The new era of monopoly is here." The Guardian May, 13.

Varian, Hal. 2007. "Position auctions." *International Journal of Industrial Organization*, 25(6): 1163–1178.

Villas-Boas, J. Miguel. 1994. "Sleeping with the Enemy: Should Competitors Share the Same Advertising Agency?" *Marketing Science*, 13(2): 190–202.

Weche, John P., and Achim Wambach. 2018. "The Fall and Rise of Market Power in Europe." ZEW Paper Series, Discussion Paper 18-003.

Werden, Gregory J., and Luke Froeb. 2018. "Don't Panic: A Guide to Claims of Increasing Concentration." *Antitrust Magazine*, forthcoming.