Attachment 3

Before the
FEDERAL COMMUNICATIONS COMMISSION
Washington, D.C. 20554

In the Matter of	
Expanding Consumers' Video Navigation Choices	MB Docket No. 16-42
Commercial Availability of Navigation Devices	CS Docket No. 97-80

DECLARATION OF MICHAEL KEARNS

April 22, 2016

TABLE OF CONTENTS

EXECUTIVE SUMMARY 1

1. Google's Data Empire.....4

2. The Importance of Data Volume: Heavy Tails.....9

3. The Importance of Data Diversity: Triangulation and Correlation 13

4. The Centrality of Machine Learning.....15

5. Video Navigation Devices and the NPRM: A Case Study in Google's Data Acquisition Strategy.....17

6. Regulatory Risks: Missing the Data Forest 20

7. Conclusions.....22

Google's Data Dominance: Harms to Competition, Consumers and Innovation from the FCC's Proposal to Disintermediate Video Distributors

Prof. Michael Kearns[1]
University of Pennsylvania

1. **Abstract:** The FCC has proposed requiring video distributors to unbundle and commoditize their content, program information and data. Google is the primary supporter and beneficiary of this extraordinary regulatory intervention. In this paper, we describe how Google's efforts in this proceeding are part of a longstanding and systematic strategy of data dominance, and the crucial role that a Google navigation device would play in expanding that dominance. We then explain the serious harms to competition, consumers and privacy that may result if the FCC adopts the regulation Google seeks.

EXECUTIVE SUMMARY

2. During its relatively brief history, Google has amassed an extraordinary amount and diversity of data on both its own users and the population at large. Initially confined to the world's largest corpus of desktop search phrases, Google's data dominance has been aggressively and systematically expanded at an explosive pace through acquisitions of other data-centric businesses, entry into new data-rich markets, development of new devices, and control of infrastructure such as operating systems and wifi networks. Collectively these massive and varied datasets provide Google with an unprecedented, high-resolution view of the thoughts, interests, concerns, plans, locations and movements of not only the general population, but of hundreds of millions of specific individuals. It is not an exaggeration to view Google as a company that specializes in surveillance and its monetization,[2] and one whose powers in this regard are unrivaled and growing.

3. In addition to its longstanding and massive data on its users' search behavior (via Google's first product, its desktop search engine), Google now routinely gathers data on its users' web surfing (via Chrome), physical location and movements (via Android, Google Maps, Waze and other services), device usage (via Android), future travel plans (Google Flights), reading and musical tastes (Google Books and Play), video consumption (YouTube), schedules (Google Calendar), personal documents and photographs (Google Docs and Drive), social activity (Google Plus), activity within the home (Nest Labs), personal health (Google Fit), and virtually any other proprietary or personal information that users "volunteer" to provide via Gmail or Google Voice. Google does not just passively collect these powerful data sources, but uses search, Android, and Google Play to drive customers to Google services and third-party apps,

---

[1] Work conducted on behalf of AT&T. All analyses, exposition and opinions are exclusively those of the author.

[2] See privacy and security expert Bruce Schneier's book "Data and Goliath" (W.W. Norton and Company, 2014) for extensive discussion of Google's surveillance practices and interests.

and exploits its massive ad network and analytics services to collect additional data from those apps and websites.

4. One of the few remaining "final frontiers" of consumer data for Google is information about what is happening inside of private residences. But they have been aggressively addressing this in recent years, making a concerted push with their acquisition of Nest Labs, which specializes in smart home monitoring; Dropcam, which provides cloud video for home surveillance; and similar acquisitions and efforts. As we shall discuss, Google's recent lobbying efforts regarding video navigation devices, including set-top boxes (STBs), are a logical next step in their campaign to get inside the home, and a case study in their systematic efforts to acquire ever more and ever new types of data, integrate it with their extant sources, and tilt the competitive landscape in their favor in yet another technological domain – this time the video navigation device, video distribution and television industries. This would cause grave harm to the entire video ecosystem, ultimately to the substantial detriment of consumers.

5. In many consumer data domains, sheer volume is not only helpful but an important competitive advantage for Google. This is partly because of the frequently "heavy-tailed" nature of many types of consumer data, in which there are always innumerably many events that are individually rare but collectively important. This is well-known to occur with any kind of language data, such as web search phrases, but also holds for consumer preferences in arenas where the number of choices is essentially limitless (such as music, movies, and books), as well as the distribution of physical locations of people and many other phenomena. As we shall detail, an important consequence of such heavy-tailed domains is that there is no such thing as "enough" data, and amassing ever-larger amounts provides Google with competitive leverage that is amplifying with time.

6. Despite the centrality of data volume to Google's strategy, equally important is data diversity – the gathering of as many different types of data on users as it possibly can. Technology industry observers sometimes express mystification at Google's forays into areas that appear unrelated to its "core" business (which we claim is actually consumer surveillance and its monetization), but the mystery vanishes if ones simply views each such excursion – whether internal projects like self-driving cars or external acquisitions – as a deliberate effort towards data diversification.

7. Data diversification is important for both simple and subtle reasons. The simple derives from the fact that combining multiple disparate consumer data sources allows Google to "triangulate" more accurately than its competitors – if you know someone's musical preferences, and their plan to travel to Rome next month, you can suggest concerts there in a way not possible from either data source alone. The subtle derives from the remarkable correlational power of apparently "orthogonal" consumer data sets – the fact that at the individual level, our habits, preferences and activities in one arena (such as the decision to use a Mac or Windows laptop) are often strikingly predictive of our behavior in others (such as political party affiliation).

8. Realizing the complementary powers of triangulation and correlation afforded by data diversity are clearly behind Google's "privacy" policies permitting them to freely combine data from any and all of their services. And again, in the same way that Google has achieved competitive domination in data volume, it has also done so in data diversity. No other company knows nearly as much about citizens, and in so many different aspects of their lives.

9. Data volume and data diversity each amplify the power of the other. When a new type of service emerges – for instance, cloud-based calendar management – it is appealing to users to choose a provider that already knows a great deal about them from other services (such as cloud-based email or contact lists), and can integrate that data into the new service. And the more information and data users have committed to Google, the easier it is for Google to enter new markets, and the harder it is for small companies or even incumbents, who do not have the same range of services or data, to compete. Since Google also dominates the massive ad delivery and analytics business, they are able to effectively both subsidize and monetize their data empire, which in turn increases their competitive advantages in both advertising and data.

10. Finally, as ambitious as it is, the mere acquisition of such phenomenal amounts of heterogeneous data is not enough on its own to achieve the powerful position Google has attained. As we shall detail, competently and fully realizing the power of this data requires carefully building complex, high-dimensional predictive models from the data. Towards this end, Google has also become the most dominant recruiter of the top talent in fields like machine learning, statistics, artificial intelligence and computer science. Despite the need for such expertise at virtually every company using data and technology, few are able to compete with the commanding recruiting position Google has managed to obtain. As with so many technology-driven trends, there is an amplifying or "rich-get-richer" phenomenon at work here: experts in fields like machine learning thrive on working with massive data sets and elite colleagues, so as Google has acquired more of both than its competitors, it makes further cornering the talent market even easier.

11. Google's recent lobbying efforts regarding navigation devices for video programming distribution are a logical next step in their campaign to acquire data inside of private residences, and an unfolding case study in their strategic and competitive practices. If these efforts are successful, Google (as well as other parties) will acquire a rich trove of behavioral data that is currently one of the few tools outside of their arsenal – data on people's television viewing habits and preferences, as well as data from a demographic segment that uses the Internet less than Google's current user population. We shall argue that should the provisions of the current NPRM be adopted, it will swiftly and irreparably damage the entire video ecosystem, in part by tilting the competitive landscape for video navigation devices and video programming distribution in Google's favor, since no current MVPD (or any other party) has amassed the data and surveillance power of Google. Because Google does not currently receive viewing data and history from "over-the-top" video services such as Netflix and Amazon, they are turning to the regulatory process in an attempt to force the unpaid exposure of the valuable and proprietary data of MVPDs.

12. Google can be expected to design a search-based navigation device – one that freely mixes television programming with free Internet video, advertising from both traditional television and media sources with display, contextual and search-based advertising, and paid placements and promotions for Google products and services. Naturally the data would flow in both directions – all of Google's extant data sources will be brought to bear on the design, interface and operation of its navigation device, and the rich new stream of television viewing data will inform Google's extant products. We can expect Google to push navigation device design towards its successful ad-supported search model, and to use its data and AdWords dominance to subsidize navigation device operating systems, as they have done with Android in mobile. While perhaps initially convenient and appealing for consumers, we shall discuss how the resulting commoditization of television content may easily result in diminishment of quality, much as has occurred with print journalism since the advent of the consumer Internet.

13. The NPRM under consideration would require MVPDs to completely unbundle their proprietary data streams – content, programming, scheduling, and permissions – and permit Google to design their own video navigation devices and user interfaces around these streams. They would be permitted to do so without attribution or compensation, and be able to display their own highly targeted advertising against the proprietary content of MVPDs, and fully integrate the resulting behavioral data with the rest of their data empire.

14. Google can be relied on to promptly bring everything it knows about consumers – their locations, interests, plans, purchases, concerns – to bear on their television viewing habits. Google will have the ability to use their knowledge of users' personal interests, such as shopping, search and travel habits, in order to select which programming to highlight in their devices' user interface, as well as what advertising inventory to place alongside that programming. No current MVPD will have the capability to match such integration, since no current MVPD has anything approaching the fine-grained resolution of Google's knowledge of individual consumers, which of course derives from the aforementioned and longstanding diversification of Google's data empire. And in turn, the detailed behavioral data a Google video navigation device could collect – users' program preferences, the times at which they watch television and the duration, which ads they click away from and which they don't – can all be used to further improve the overall accuracy of user profiles, which benefits all of Google's businesses horizontally.

1. Google's Data Empire

15. Since its founding in 1998, Google has systematically created a data empire whose scale and scope are unprecedented. In its earliest years, Google was a highly focused and

specialized company,[3] offering just their core search engine service. But as was quickly discovered at the dawn of the consumer Internet, search is the most data-intensive of businesses. Early efforts to "catalog" the content of the Web using hand-coded human expertise or knowledge (such as Yahoo!'s original hierarchical index) were quickly overwhelmed by the exponentially growing content. Since click-throughs provided an objective measure of search result quality, the race for data supremacy was on – the more users a search engine had, the more feedback or "training data" it received to improve its search algorithms via machine learning, which would in turn bring more users and data.[4] This cycle was accelerated by the introduction of pay-per-click advertising against keyphrases as the dominant monetization model for search. It seems fair to say that Google effectively won this first data arms race sometime around 2003, and has comfortably dominated the search market ever since.

16. Beginning in the early 2000s, Google has gradually expanded the scope of its services and therefore the data it collects on consumers. The list of distinct Google products and services easily numbers in the hundreds,[5] and was amassed through a remarkable combination of external corporate acquisitions[6] and internal development efforts. While many of these efforts, especially the earlier ones, were related to the core search business, most of them are not. Rather, these services and products span a staggering variety of human activity, and collectively provide Google with an unprecedented view of its users (and even non-users). It seems implausible that the scale and scope of this expansion[7] can be explained by occasional experimentation in new markets; rather, it appears to be a core part of corporate strategy.

17. An only partial taxonomy of the various types of consumer and other data Google has acquired includes the following broad categories:

18. **Search:** In this oldest, original category, for its entire history Google has collected the queries entered by users of its search engine. It is now well-understood and well-documented that users tend to display uninhibited search behavior – they routinely search for information about diseases or conditions they suffer from, their worries or problems in life, their professions and hobbies, literally anything that "comes to mind". It is also well-known, and unsurprising, that the search history of a user can act as a near-

---

[3] Google's first external acquisition appears not to have occurred until 2001, and only a handful of their almost 200 documented acquisitions occurred during their first five years. *See* https://en.wikipedia.org/wiki/List_of_mergers_and_acquisitions_by_Google.

[4] In a later section we shall detail why this cycle essentially never terminates – i.e., there is no such things as "too much" data – due to the statistical properties of language.

[5] https://en.wikipedia.org/wiki/List_of_Google_products.

[6] https://en.wikipedia.org/wiki/List_of_mergers_and_acquisitions_by_Google.

[7] In a recent two-year period, Google spent $17 billion on acquisitions, outspending Apple, Microsoft, Amazon, Facebook and Yahoo! combined. *See* http://www.businessinsider.com/google-acquisition-spending-2014-1.

fingerprint for their identity.[8] Google can associate search histories with individuals either directly (if they log into a Google account), or indirectly (via information like IP address, device type and operating system, GPS coordinates, and techniques such as browser fingerprinting). Search is also special since the Web largely consists of free-form, natural language text, which must be crawled and indexed. As we shall discuss, the particular challenges of search data call for massive machine learning efforts, another area in which Google is dominant.

19. **Browsing History:** Similar comments hold for general web-surfing data (i.e., the sequence of URLs visited by a user, as opposed to the search terms they type into the Google search engine), which Google collects for any user of its Chrome browser (which is by most estimates[9] now the dominant browser, and is the default installation on all Android devices). Chrome is also used to power third-party browsers such as Opera, further increasing Google's browsing market share and data.[10]

20. **Media Preferences:** Over time Google has created a stable of services whose usage gives them a detailed profile of a user's tastes and preferences in media consumption of many varieties. These include online video consumption (via YouTube), music (again via YouTube, and other services including Google Play Music), books (via Google Books), News (Google News), Photos (Picasa and Google Image Search), and many other categories. Of course, information about media consumption from these more services is amplified by user searches in these same categories of items, as well as shopping interests from services such as Google Shopping.

21. **Location, Movement and Travel:** From its inception Google has possessed information regarding the physical location of its users via IP addresses and their close association with geographic coordinates.[11] But such mappings can be imprecise compared to actual GPS data, and also are less applicable to mobile devices. So over time Google has developed and acquired a collection of services that provide much broader and more precise geolocation data on its users. Much of this data is collected by products that are infrastructure, rather than consumer-facing applications. A notable effort in this regard is the Android operating system that runs all Android devices, and which gathers the GPS coordinates of those devices on behalf of all location-based services. Location data is also gathered by the growing wireless network Google has been building out (Google Fi;[12] partnerships with Starbucks and New York City;[13] and many similar projects). In recent

---

[8] http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=0.

[9] https://en.wikipedia.org/wiki/Usage_share_of_web_browsers.

[10] http://www.cnet.com/news/a-nail-in-the-coffin-for-firefox-mozilla-struggles-to-redefine-browser/.

[11] As a demonstration, simply type "what is my IP address" into Google, and visit a site like http://ip-api.com/ to see how much this address conveys.

[12] https://fi.google.com/about/.

[13] http://www.cnet.com/news/at-starbucks-at-t-is-out-and-google-is-in-for-wi-fi/.

years, Google has made significant investment in services that not only determine where users *are*, but where they *will* be or *plan* to be in the future. The navigation services on Google Maps and the popular mobile app Waze (acquired for almost $1BN in 2013) clearly invite users to share short-term driving or travel plans, and via GPS can track actual progress on trips. On a longer temporal and geographic scale, Google Flights (acquired as ITA Software for $676M in 2010) provides views into users' commercial airline travel plans and purchases.

22. **Social Activity:** In addition to an explicit social networking application (Google+), there are myriad sources from which Google can extract information about the friends and other social, business and family relationships of users. These include Google Contacts, a cloud-based address and contact list management service; Google Hangouts, an Internet videoconferencing service; Google Voice, (call forwarding and voice messaging); and many other products with a social or sharing component, such as YouTube.

23. **Communications and Documents:** Some Google services permit a more unrestricted view into the lives of users – effectively anything they "volunteer" to share with Google via the use of products like Gmail. It is well-known that Google algorithms inspect the content of Gmail messages for ad personalization and other purposes. Similarly, Google Docs provides another free-form data channel, and Google Calendar allows users to document their entire schedules on company servers.

24. **Device Usage:** An incredibly broad, rich and detailed source of behavioral data that Google enjoys derives from the dominance of the Android mobile operating system, which provides data on virtually everything a user does on their device. In addition to massive volumes of data in the categories above, this includes detailed data on app usage.

25. **Residential Activity:** One of Google's more recent and systematic forays seems to be products and services that provide the opportunity to collect data inside the home. Most notable in this regard was their acquisition of Nest Labs for $3.2BN in 2014; Nest provides "smart" devices for home monitoring and control, such as thermostats, smoke alarms and cameras (via Dropcam, which was acquired shortly after Nest in 2014 for $555BN). There is also an extensive list of "Works with Nest" partners that provide a wide selection of home control, monitoring and surveillance products.[14] It is unclear which of these products might share the residential data they collect with Google via the Nest Partnership, but there certainly seems nothing explicit to prevent such sharing. Google Fiber, Chromecast and Google Voice are other services that provide data on residential activity.

26. One major type of data that falls in this last category is data about the television viewing habits of consumers. Despite significant recent competition for the television industry from the Internet and other media channels, there remain well over 100M television-viewing households in the U.S. alone, presumably representing a far greater number of individual viewers. Furthermore, there are many aspects of television viewing that remain

[14] https://workswith.nest.com/category/all.

unique to the medium, and that therefore capture information not easily obtained through other data sources. These include the fact that many shared cultural and entertainment events remain exclusively or primarily delivered via linear (i.e., live) television, such as the Super Bowl (which this year had over 110M viewers, and whose ads have even become an event in themselves), important or final episodes of television series (witness the outbreak of parties that accompanied the conclusion of *Breaking Bad*),[15] presidential debates and speeches, and live newscasts of important events that still require the traditional reporting and operational staffs of live television. In addition, there are known ways in which the demographics of television viewers differ from populations primarily consuming more recently introduced forms of media consumption. For instance, elderly people tend to watch more television on average,[16] meaning that television is a mechanism both to reach that group and gather data on their interests and habits.

27. These facts certainly cannot be lost on an organization as sophisticated about and experienced with data as Google. Google's goal is to effectively become the chief intermediary for all electronic content. Their current efforts to disintermediate MVPDs (via the forced unbundling of their programming and other content), if successful, would allow them to access this rich source of data missing from their current flows. As we shall argue below, this will enable them to bring all of those other flows to bear on the content and user interface of their video navigation device, and to bring video navigation device data to bear on all their other services. We shall detail how this data amplification cycle will irreversibly alter the competitive landscape in video distribution and navigation devices in Google's favor, since no current MVPDs have anything approaching Google's data empire in terms of either volume or diversity.

28. The Google data taxonomy we provide above mentions only a fraction of the hundreds of products Google offers and companies it has acquired. It also omits a number of notable past efforts that Google may significantly invest in again (such as health monitoring and data collection, in its ambitious but discontinued Google Health platform; related efforts include the active Google Fit Android exercise and health app platform), and ongoing efforts that open up entirely new categories of surveillance data for Google (such as the Google Glass "ubiquitous computer" and self-driving cars).

29. Even though a typical user might use just a few Google products on a regular basis, it is worth contemplating the comprehensive view of an individual provided to Google by the data collected. Regular use of Google Search will already provide a tremendously detailed and private background profile of a user's personal, professional, social, medical and commercial interests. Use of Gmail will provide regular, free-form natural language data on their personal and professional activities. Navigating using Google Maps or Waze will track the user's movements, and use of any of the above on a mobile device, or the Android operating system more generally, will tell Google accurate geographic

---

[15] http://variety.com/2013/tv/news/breaking-bad-viewing-parties-1200671090/.

[16] *See, e.g.*, https://www.washingtonpost.com/news/business/wp/2014/09/05/tv-is-increasinglyfor-old-people/.

coordinates for the user at frequent intervals. By integrating the data from just these few of the more common products, Google knows who you are, what you're doing, and where you are in real time throughout a typical day. As privacy expert Bruce Schneier puts it, "Google knows more about what I'm thinking than I do, because Google remembers all of it perfectly and forever."[17]

30. While the extent to which Google is currently integrating and exploiting data across most of its hundreds of products is unknown, their intention to do so was clearly signaled by a sweeping change to its privacy policies in 2012 that explicitly allowed them to combine data across all their services.[18] And as we shall detail, there is no doubt that Google has the necessary scale, technology, and scientific expertise to extract all the power that such unification makes possible – not merely in the sense of creating integrated views of disparate data sources (which indeed is itself already a considerable technical challenge), but in the sense of building highly accurate predictive models for every aspect of their users.

31. The Google description of these policy changes, as usual, emphasizes all the product improvements that can result by the commingling of cross-service data sources. Left unmentioned are all of the potential privacy, surveillance and competitive harms that result from allowing a single organization a virtually unlimited, centralized repository of human behavior at the individual level. In discussion of its privacy policies, Google often emphasizes – and is also often lauded by observers for – its refusals to sell its data to third parties. Such remarks miss the point that Google's dominance derives precisely from its unprecedented proprietary data empire. Selling their data would undermine this crucial competitive leverage. Their business model is not to monetize the sale of data to others, but to exploit it internally for the predictive perfection of their own products. Thus far it has succeeded brilliantly.

32. In the coming sections, we will detail more fully the important ways in which both the volume and diversity of consumer data Google has obtained give it competitive advantages that are amplified with time and each new market Google enters. These advantages largely derive from technical aspects of the distribution of consumer behaviors and data, and the use of machine learning to extract value from them. We shall discuss these topics in layperson's terms, and along the way examine the implications, both technical and competitive, of allowing Google to disintermediate MVPDs.

2. The Importance of Data Volume: Heavy Tails

33. Given the discussion of Google's consumer data empire above, it would be reasonable to wonder whether so much data is necessary, and exactly how it provides a competitive advantage. After all, at some point doesn't one have "enough" data, and therefore any party that crosses this threshold amount can compete in principle? We shall argue that

[17] "Data and Goliath," Bruce Schneier (W.W. Norton and Company, 2014).

[18] *See, e.g.*, http://www.cnet.com/news/google-wants-ability-to-combine-your-user-data/ and *also* https://googleblog.blogspot.com/2012/01/updating-our-privacy-policies-and-terms.html.

this answer to this question is no, due to the frequently "heavy-tailed" nature of many types of consumer data, including but not limited to any type of language-based data (as with search), expressions of preferences across very large numbers of choices (as with choices of media like books or movies), and physical locations (as with local search queries, physical location of users, and mapping and navigation applications).

34. It is scientifically well-documented that many types of data generated by human and social processes – such as language, where people live and travel, and their preferences in areas where the number of choices is immense (such as music, movies or books) – obey what is known as a heavy-tailed distribution. As we shall describe, an important implication of this for Google and its competitors is that events rarely or never observed before continue to appear regardless of how much data has already been gathered.

35. Suppose we are trying to build a search engine from scratch, and that we have already managed to index the contents of the Web (itself no small feat) and attract some users. In order to decide or optimize what links or ads to show in response to a user query, we need to get feedback: by showing users potentially relevant links, and seeing what they click on, a search engine can *learn* what is more and less relevant, and promote or demote it accordingly.[19] At the highest level, this explorative learning process is at the heart of building a search engine, or any other kind of consumer recommendation system.

36. If there were only 1,000 or 10,000 words in the English language, we might hope to eventually get enough queries that we would have seen each word many times, enough to learn the optimal links to show with each. Of course, more important than the actual number of words is their relative frequency – even if there were an unlimited number of words, but only the 1,000 most frequent "mattered", it might be relatively easy to build a competitive search engine.

37. But it turns out that language is not like this: while the most frequent words of course account for most occurrences, the infrequent words are still important. For example, Zipf's Law,[20] which has been observed empirically since the 19th century, refers to the fact that in most languages, the kth most common word occurs about 1/k as frequently as the most common word. Such a distribution is called heavy-tailed ("tail" referring to the low-frequency end of the distribution) because it means that *words that are individually rare are collectively common*, and thus cannot be ignored in building a comprehensive search engine. The more queries we have seen, the more words we have rarely or even never seen before will arise,[21] which in turn creates the need and opportunity to gather

---

[19] We shall have a good deal more to say later about the use of machine learning across all of Google's application and data domains.

[20] https://en.wikipedia.org/wiki/Zipf%27s_law.

[21] For instance, it is easy to work out that for any quantity obeying Zipf's Law, the number of items whose frequency is roughly 1/n is a constant fraction of the distribution no matter how big n is. This means that no matter how big our sample size is, we are always observing new items – they are just being shifted further into the tail as the sample size grows, but never diminishing in

more data on them in the learning process. This phenomenon is especially exacerbated for search, where queries are not individual words, but often compound phrases consisting of many relatively infrequent words and proper nouns (like "Philadelphia florist Fairmount neighborhood open Sundays"), or even words, names, terminology and slang that have only been in existence a short while (like "splash brothers dubnation").

38. Let us illustrate this phenomenon with another data domain in which Google, after a late entry, has managed to quickly become the dominant player: navigation and maps – that is, data about where people are and where they are going.[22] Like search, mapping and navigation is an area where the queries are heavy-tailed, simply because where people live and travel is also heavy-tailed. If Zipf's Law holds here, we would expect the kth largest U.S. city to have about 1/k the population of the largest city (New York City). Thus if we plot the logarithm of k on the x-axis, and the logarithm of the population size on the y-axis, then since log(1/k) = -log(k), we would expect a straight line with negative slope.

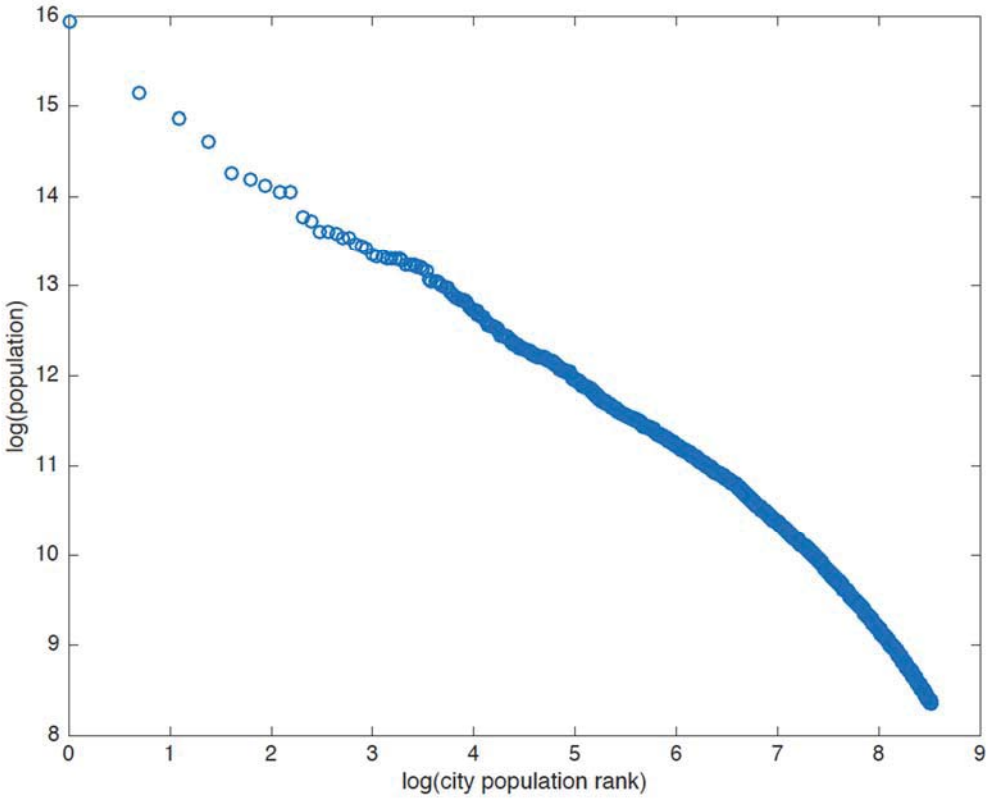
39. This is exactly what we have plotted in the figure below, for the 5,000 largest U.S. cities,[23] and we can see that Zipf's Law holds quite well here. In practical terms this means that as we see more place queries, we keep observing new places for the first time, or getting crucial additional observations of places we have seen mentioned only a handful of times. For example, the 1,000 *smallest* cities among the 5,000 largest have a combined population of more than 5 million, which is larger than Los Angeles; yet the largest of them is only 6,000. The 100 smallest have a combined population of 440,000, which is larger than Omaha; yet the largest has only 4,400. This is what we mean by individually rare but collectively frequent: if we "ignore" the locations with a small population or number of queries, in the aggregate we end up ignoring the population of a sizable city.

---

frequency. This property will not hold for other common types of distributions, such as the "bell-shaped curve" of the Gaussian or Normal distribution.

[22] Google Maps was introduced in 2005 (via acquisition of Where 2 Technologies), long after early Internet navigation services such as MapQuest (1996), Yahoo! Maps (2002) and others; Waze was acquired by Google for $1.3BN in 2013. While Google does not break out separate market share estimates for its mapping services, since Google Maps is the default installation on all Android devices (roughly 83% of the smartphone market worldwide in 2015), and also has significant adoption on iOS devices, it seems safe to say they are the dominant player.

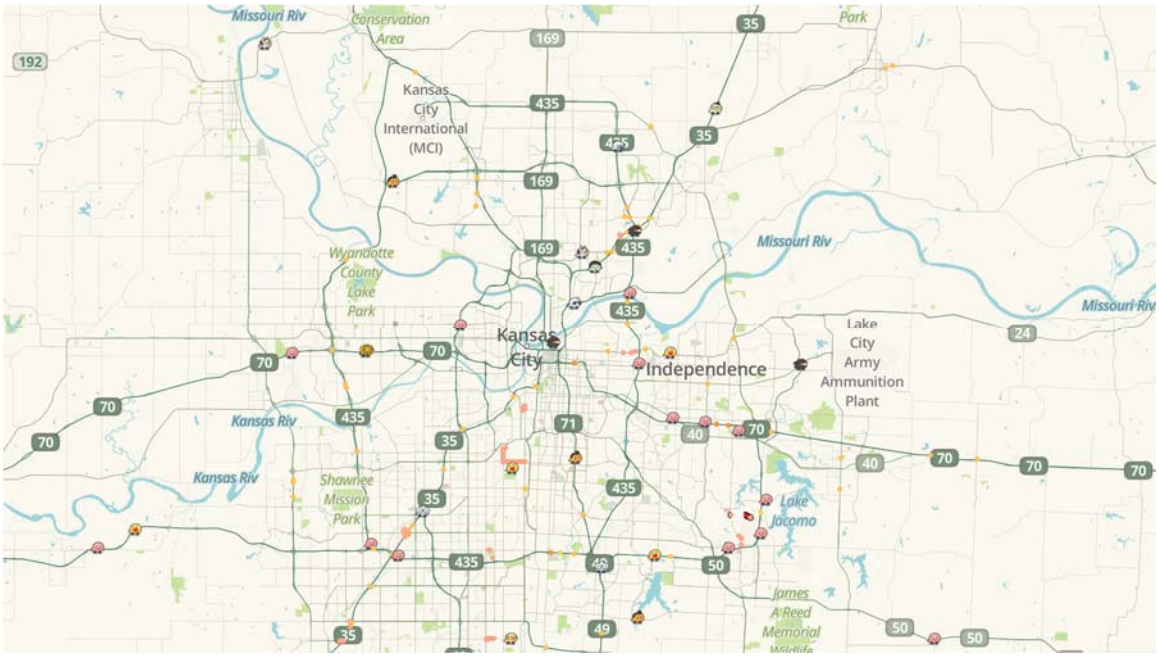
[23] Raw data from http://ezlocal.com/blog/post/Top-5000-US-Cities-by-Population.aspx.



Plot of the logarithm of a city's rank by population size vs. the logarithm of its population size itself, demonstrating the heavy-tailed property – small cities collectively account for large populations.

40. And this is just at the rather coarse level of cities, towns and villages. Within each of these, no matter how small, there lies another heavy-tailed distribution of traffic patterns, road closures and construction, gas stations and restaurants, etc. that drivers or locals might want to know about. The image below is a recent screen shot from the Live Map service of the crowdsourced navigation application Waze[24] for the Kansas City area. Each tiny pink, orange or yellow icon denotes some type of traffic-related information or event. While there are a lot of them, it is hard to imagine there are not many more that simply have not been reported or observed yet. Again, there is no such thing as too much data in such domains. As with search, the party that manages to keep getting more place and navigation queries can optimize and monetize them better, by learning more about local traffic patterns and road construction and closures, learning more about local businesses, etc.

[24] https://www.waze.com/livemap.



Snapshot of Waze's live map of traffic events in the Kansas City area, showing the fine granularity at which such local data is generated.

41. Practically all of the data markets Google has chosen to enter over time have this heavy-tailed property – not only search and location/navigation, but also advertising for search, contextual and mobile, where advertisers can target incredibly specialized long-tail niche audiences. They now seek to enter the heavy-tailed market for television and video content search and advertising as well, which is akin to music, books, and other forms of consumer media consumption. As we shall discuss, the centrality of machine learning in such markets makes them extremely well-suited for Google to extend its dominant position via AdWords and its massive advertiser demand.

3. The Importance of Data Diversity: Triangulation and Correlation

42. In the discussion so far, we have emphasized the importance of having massive amounts of data of a particular type – for instance, search queries – as a prerequisite for dealing with heavy-tailed phenomena. But equally important (and equally central to Google's long-term strategy) is data *diversity*: the gathering of as many different *kinds* of consumer data as possible. We might characterize data volume in a particular domain as being analogous to "depth", whereas data diversity across domains is more concerned with "breadth". A review of our opening section, which gave an extensive (yet still only partial) catalog of the many types of data Google now collects, should make it clear that they have systemically and aggressively grown the diversity of their data empire over the last 15 years.

43. As was the case with data volume, there are particular technical reasons why data diversity provides significant competitive advantages. The simplest reason is what we might call *triangulation*: by having multiple consumer data sources, it becomes possible

to predict both collective and individual preferences and actions much more accurately than with any single source alone. For example, knowing only a user's search history might reveal quite a bit about their interests, profession, health, and so on. But also knowing their location and travel plans, as well as their upcoming schedule, permits personalization and targeting that combine properties of consumers (such as their tastes and preferences) with their activities (such as trips and important dates like birthdays and anniversaries). If I know about you, your spouse, your spouse's interests, and their birthday, I can effectively suggest presents you might purchase in a way not possible if I am missing even one of these data items.

44. Such direct triangulations, however, only scratch the surface. Each different consumer data source may also permit many indirect inferences about users, which in turn can be used to influence the interpretation and use of other data sets. For example, for many users, location data alone may give an accurate indication of approximate age, via apparent waking hours and extent of movement, speed (indicating walking, running, biking, driving, etc.), whether one goes to Florida for Spring Break – the inferences that can be drawn are limitless and powerful. Armed with this purely location-based "demographic" profile, one can then personalize search results accordingly. A user with a long history of exotic travel locations who enters "Java" is more likely to mean the island than a user whose GPS coordinates have not left Silicon Valley in a year, who is more likely to mean the programming language.

45. This is just one small example of the extraordinary, non-additive correlational power of combining multiple data sets. The simple fact is that despite our individuality, on average our activities and interests in one arena are highly predictive of our activities and interests in another – including in surprising, subtle and useful ways like your choice of a Mac or a PC being strongly correlated with your political affiliation (and many other things).[25] Furthermore, in social data sets (which includes not just overtly social data like that from Google+, but data from services such as Gmail and many other sources), even data generated not by a user but by their contacts may reveal much about the user, such as sexual orientation, ethnicity and religion.[26]

46. More than any other company or organization, Google realized the power of data diversity early in its history and began its systematic expansion into virtually any market that would provide it with a new source of consumer data, which could in turn be brought to bear on all its extant products. Some of these markets were pioneered by Google, such as search. Some of them were already mature, and Google was a late entrant, such as email and maps and navigation. Some of them were nascent or even speculative, such as

---

[25] *See* http://www.theatlantic.com/technology/archive/2011/05/infographic-the-ever-expanding-taste-graph-by-hunch/238737/, http://www.coolinfographics.com/blog/2011/4/28/self-described-mac-vs-pc-people-infographic.html and http://www.abctrick.net/2012/01/12-surveys-of-thay-teach-hunch-about.html for an extraordinary variety of examples of such correlations documented by extensive online surveys conducted several years ago by the startup Hunch.

[26] *See, e.g.*, http://www.pnas.org/content/110/15/5802.full.pdf.

Google Glass or self-driving cars. But all of them share the property of providing Google with a novel source of consumer data.

47. As we shall discuss in more detail below, the variety and aggressiveness of Google's expansion seems to have often confused and frustrated regulators, who are prone or forced to view each Google acquisition through a narrow lens, first attempting to "define" the market in question, and then asking questions such as whether the market is "nascent" or "mature" (such as that for mobile ad networks circa 2010),[27] whether there are yet "dominant" players, etc. This approach often misses the data forest for the trees, since the real competitive power Google derives from its acquisitions cannot be understood by looking at each in isolation, for the very reasons we have outlined above.

4. The Centrality of Machine Learning

48. As crucial as data volume and diversity are to Google's competitive position, there is an equally important third pillar, which is machine learning – the modern science underlying the construction of large-scale predictive models from massive data sets. Machine learning is a mixture of topics from areas as diverse as statistics, probability theory, pattern recognition, algorithms, artificial intelligence and most recently, distributed systems. While its origins lie in the 1980s, in recent years the data explosion enabled by the Internet has made machine learning one of the most important scientific fields, and one that has even entered the popular consciousness.

49. A detailed examination of the manifold ways in which machine learning is central to practically everything that Google does is beyond our scope, but we will try to provide some examples in the context of the remarks we have made thus far. The punch line here will be that the more diverse data you have, the more complex, subtle and better-performing predictive models you can build. This phenomenon has accelerated rapidly in very recent years through a combination of advances in computational power and scale, and machine learning research and engineering.

50. Let us first consider the role of machine learning in the context of our remarks on heavy-tailed data domains. Superficially it might seem that the solution to the problem of individually rare but collectively frequent observations would be sheer data volume, as we have discussed. But to maximize the advantages of raw scale, one must additionally discover, in an automated fashion, all of the correlations and duplicates that can occur in the heavy tail. These include language-based phenomena such as contextual synonyms – for instance, the fact that "bike" and "bicycle" might be interchangeable in the context of a query that also contained "mountain", but not in the context of a query that instead contained "Harley Davidson", where "motorcycle" is the better synonym. By recognizing such relationships, one effectively amplifies the number of observations of rare events, since they can be grouped together. But this recognition can't be done "manually" or based on human expertise – there are simply too many such relationships, and they appear and change too quickly. Instead Google uses advanced machine learning

[27] https://www.cornerstone.com/Publications/Articles/Merger-Review-Google-AdMob.

techniques to build extensive, large-scale language (and other) models that automatically discover synonyms, antonyms, multiple word senses (such as “fire” in the senses of burning, guns, and termination of employment), topic models (vocabulary associated with particular topics, which themselves must be inferred automatically from data), and many other things. The same expertise and set of techniques can also be profitably applied to non-language data like consumer preferences (where machine learning is widely used to build recommendation systems that suggest content or products a user might like based on their own history as well as collective modeling), location and navigation (where machine learning can be used to predict future traffic patterns and optimal routes based on past data), and many other domains in which Google has established a preeminent position.

51. Regarding data diversity, virtually all of the myriad cross-domain inferences we have already discussed are made possible by machine learning. While we might be able to articulate and measure particular correlations that we are curious about – such as whether there is indeed a correlation between one’s political views and one’s choice of a Mac or PC – these might not be the ones that turn out to be meaningful, predictive or profitable, and there are far too many for us to exhaustively enumerate and test them. Far better to have highly scalable, data-driven machine learning algorithms that can directly find the most important relationships without much or any human intervention. Common techniques here include classical methods from statistics such as Singular Value Decomposition, and more modern algorithms for low-rank matrix approximation. The basic idea behind both is to take many large (say) consumer data sets and use them to automatically find a small number of canonical “types” or clusters of users, and then represent individual users as combinations or mixtures of these basic types. In this way large population-scale data sets are used to first model collective behavior in a low-dimensional manner that avoids or limits overfitting, and then construct “personal” models for specific individuals.

52. The algorithms of machine learning and the models they produce are largely automated once in operation. But the development of these algorithms, their improvement and evolution, their implementation in a distributed, cloud-based computing environment, and their specialization to the idiosyncrasies of new and ever-changing data sets remains a highly technical, research-intensive, and human-centric activity. Not surprisingly, this is again an area where Google has maintained competitive dominance for many years through aggressive and systematic recruiting practices. Google is without question one of the earliest and largest employers (probably the largest, though of course any such information is proprietary) of machine learning PhDs, including not only fresh graduates, but many senior, tenured academics they have lured from top universities. They are well-known in the academic machine learning research community; not only do their own scientists often publish in and attend the top journals and conferences, they are well-represented on the editorial committees of such venues, as well as reliable financial

sponsors at the most generous levels.[28] Google's acquisitions in recent years have had a strong focus on machine-learning and AI-based companies such as DeepMind (purchased for $625M, and whose recent success in the development of a learning-based, world-class Go-playing program made headlines when it defeated a champion player from Korea).[29]

53. As with almost everything Google does, there is a cyclical, rich-get-richer dynamic with respect to their machine learning recruiting efforts. Competition for PhD-level scientists and engineers in areas like machine learning and computer science is more intense than ever, in the technology industry and beyond. In addition to Google's established competitors like Facebook, Amazon and Microsoft, virtually every consumer-facing Silicon Valley startup is desperately seeking expertise in these areas. But especially the smaller companies find luring top candidates away from Google exceedingly difficult, because research and engineering remain social activities, and thus the best people want to be where the other best people already are, where the data is the most plentiful and most diverse, where the computing infrastructure has the largest scale, and so on. Thus Google's data empire reinforces its dominant recruiting position, and vice-versa.

5. Video Navigation Devices and the NPRM: A Case Study in Google's Data Acquisition Strategy

54. Google's recent efforts to lobby and influence the FCC (as evidenced by the current NPRM, which would allow Google and other technology companies to effectively disintermediate incumbent MVPDs) is an evolving case study in their aggressive expansion practices, and must be viewed in light of the technical remarks we have made so far, particularly with regards to data diversity. Simply put, the video navigation device data that Google would acquire under the NPRM would be perhaps the single most valuable data asset that they do not already possess. The detailed television viewing data – the programs and ads watched or switched away from, for how long, at what times of day and what days of the week, and all the demographic and behavioral inferences that can be made accordingly – for tens of millions of households and hundreds of millions of viewers, is a "final frontier" of consumer data that Google must certainly be salivating to acquire. And of course it is not simply the data that Google would acquire under the NPRM, but in fact the massive television advertising market as well.[30] Google does not currently receive viewing data and history from "over-the-top" video services like Netflix and Amazon, so they have turned to the regulatory process in attempt to force the unpaid exposure of MVPD data.

---

[28] Google had top billing among dozens of corporate sponsors at NIPS, the premier machine learning conference, in 2015 (https://nips.cc/Conferences/2015/Sponsors), as well as at many other prominent machine learning conferences and workshops.

[29] *See*, *e.g.*, http://www.nytimes.com/2016/03/10/world/asia/google-alphago-lee-se-dol.html for one example of the extensive coverage.

[30] *See*, *e.g.*, the "addressable" television advertising at http://adworks.att.com/.

55. Let us contemplate more specifically what a Google-designed navigation device might look like, and the issues that it would raise. It would be natural for Google to apply its core expertise in search and advertising to its device. Thus, a search-based user interface allowing consumers to do free-text queries for content would be expected, and the returned content would not be limited to television programming. Instead, it would likely mix television, online video and movies from many Internet sources, YouTube, paid video and music, perhaps even traditional web results. Like Google's desktop interface, we should expect the "organic" search results to be mixed together with advertising from search keyword auctions, direct paid placements from advertisers, and preferential placement of Google products and services. There would be no guarantees that traditional programming as currently offered by MVPDs would receive equitable treatment, nor would there be any restrictions on the nature of the advertising intermixed with the search results. Google advertisers may be encouraged or even compelled via packaging or bundling to have their ads shown across all of Google's properties – search, mobile, contextual, and now television advertising.

56. In short, there is no reason to expect that all of the competitive and technological practices Google has historically employed in its core business – which have drawn widespread criticism and scrutiny from regulators and competitors alike, around issues of fairness and transparency – would not be applied with equal force in a navigation device. While Google's navigation device may appeal to consumers initially, a serious concern is that the gradual demotion and commoditization of the original content currently offered by MVPDs will result in the loss of incentives to produce it in the first place – much as has happened with both print and television journalism since it was unbundled and commoditized by Google and other Internet access methods.

57. It might be less concerning if the impact of the NPRM were to simply "level the playing" field for Google and other technology companies in their efforts to enter the video navigation device and distribution markets. Unfortunately, the NPRM instead tilts the playing field wildly in favor of Google, and against incumbents. As per our remarks above, Google can be relied upon to quickly combine navigation device data with everything else it already gathers and knows about consumers, both individually and collectively, and use it to optimize the interface, content and advertising in its own video navigation device. Your search and browsing behavior, the contents of your Gmail, the physical locations you visit over time and where you travel, will all be brought to bear in determining the video programming and advertisements you are exposed to. And of course, because of the cyclical, rich-get -richer aspects of data volume and data diversity, the video navigation device data that Google acquires will in turn allow better monetization in its extant businesses: your television viewing habits will influence your search personalization, the interpretation of your Gmail, etc. In contrast, for regulatory and other reasons, incumbent MVPDs have largely only one valuable source of consumer data: television viewing data from video navigation devices, which the NPRM would now hand over to the already data-rich technology companies.[31]

[31] The uneven nature of the landscape is greatly exacerbated here by the positive feedback loop enabled by machine learning, which creates massive indirect network effects: the party with the

58. Consider the thought experiment in which Google is required to provide a direct, reliable and high-speed API to its core search technology to any party that wanted to build a competing search engine. The competitors could thus be "powered" by Google Search – shipping off user queries en masse to the API, receiving the Google results, but then reordering, rewrapping, mixing and rebranding them in any way a competitor liked, and showing the competitor's own advertising inventory against them – without any compensation or acknowledgement for Google. In order to remain competitive, Google would need to continue to maintain and improve its search technology, including by bringing all their other data sources to bear on search, but all such improvements would also immediately benefit competitors via the API.

59. Google might object to such a requirement, and legitimately so. After all, they built their search business from scratch over more than a decade, at great expense and risk. They have proprietary content and technology from their massive investment in this business. Why in the world should they be forced to give their core technology away to competitors?

60. The NPRM instead imposes the analogous thought experiment on MVPDs. Video programming distribution is a business and technology built over many years, and relies significantly on agreements negotiated between programmers, advertisers and MVPDs based on production costs, consumer interest, security of content, ad placement within content, and many other moving parts. It is a complex ecosystem that has thrived in recent years, resulting in what many have called a new "golden age" of video and television. MVPDs are a central part of the industry that created and shaped this era. Not only is allowing Google to step between MVPDs and their users inequitable from a competitive standpoint, it threatens to disrupt the constellation of forces, agreements and investments that have resulted in the rise in the quality and quantity of video programming in recent decades. As just one example, disintermediation of MVPDs greatly undermines incentives to invest in the video ecosystem, since content producers and MVPDs will now have to share the benefits of those investments with Google, which is not paying them anything in exchange.

61. Google's advanced and unique power in contextual, targeted and personalized advertising, which has been honed and amplified over its 18-year history (in no small part due to its vast data trove) could trigger a cascade of negative consequences for video distribution and its customers. Instead of advertisers knowing when and in what content their spots would appear, they would now be at the mercy of Google's massive optimization engine, and also be competing with all manner of auxiliary advertising inserted by the navigation device interface. The reduced value of television advertising

---

most diverse data can create the best predictive models, which in turn attract more users and thus more data. In an Internet-centric world, this must all be done at no cost to consumers and thus must be cross-subsidized through advertising, which is the core of Google's competitive advantage.

would provide less incentive for programmers to produce quality content, in turn harming consumers.

62. We also note that video navigation devices and the NPRM seem to represent a departure from Google's typical expansion strategy. As aggressive as that strategy has been historically, in the many cases we have discussed so far Google has either "built" or at least "bought" – i.e., they either incubated a new business or product internally (as with search), purchased entry through an acquisition (as with video in YouTube), or some combination of the two. In contrast, their current efforts to enter the video navigation market instead consist largely of lobbying regulators to give them the keys.[32] Google's traditional search business relies on crawling and indexing the largely unstructured data of the Web. Proprietary sources of data like television require Google to negotiate or license structure data feeds, which does not fit with Google's extant business; it appears that in the present case, they are turning to the regulatory process instead.

63. Finally, Google has historically had very lax and opaque attitudes towards both consumer privacy and the protection of copyrighted content. With regard to privacy, like many technology companies Google often has complex policies that change frequently and without warning, such as the aforementioned policy alteration that permits them to freely commingle all sources of consumer data. Google's policies are generally designed to benefit Google and its primary customers: advertisers. As the Internet saying goes, if something is free, you are the product – or more accurately, your data is. This is not to say that Google needs to actually sell its data – as we have already pointed out, it is actually in their interests to instead keep their data empire proprietary, and instead offer advertisers opportunities to target audiences using Google tools. Regarding content protections, Google has benefitted enormously from the commoditization of many types of content, such as print journalism, and providing consumers "side-channels" to access previously paid content for free. They have also engaged in direct assaults on established protections, such as with "orphaned" copyrighted material in the Google Books case. Surely there is no reason to believe their attitudes towards video and television content will be any different should they be permitted to dictate navigation device design.

6. Regulatory Risks: Missing the Data Forest

64. We conclude with some remarks about the difficulties Google presents for regulators who have examined their competitive practices and acquisitions over the years, and the consequences for the debate over the current NPRM regarding navigation devices.

65. On many occasions when Google has tried to enter a new market, or to strengthen its position in one of its extant businesses, regulators have carefully considered whether threats to competition might be present, or whether Google's competitive practices needed curtailment or further scrutiny. Unfortunately the tendency in these investigations has always been to narrow the discussion as much as possible to the specific market in

[32] Google Fiber is a more direct effort, but is present in only a handful of U.S. cities and has only a small number of users (https://en.wikipedia.org/wiki/Google_Fiber).

question, whose very definition may be ambiguous. If regulators determine that the particular market is "nascent", and therefore that there are no clearly dominant incumbents, or that it is "mature" but competition is sufficiently present, they are reluctant to intervene.

66. Google's 2009 acquisition of AdMob, which was scrutinized but allowed by the FTC, is a good example of this dynamic. Rather than look at Google's online advertising business holistically, where it was already easily the dominant player domestically and globally, regulators instead chose to examine the acquisition through the narrow lens of contextual mobile advertising[33] – as if this were somehow a separate market from search advertising and everything else that Google does. The entire mobile industry was still in its relative infancy then – the first iPhone was introduced only two years before, in 2007 – and the landscape for mobile advertising was indeed far from settled, with the underlying technology and revenue models still under development and experimentation, and no single player the clear winner. Viewed in this narrow context, it might seem reasonable to be unconcerned about Google's entry.

67. We can be reasonably sure that Google leadership was not evaluating this acquisition wearing the same blinders, but instead thinking about all the ways it could integrate mobile advertising data with its already vast consumer data empire. Indeed, in light of the extensive remarks we have made here on the power of data volume and diversity, it seems quite likely that Google carefully considers how each acquisition or new business complements its existing data, and what synergies might be exploited. In this regard, the worldviews in D.C. and in Mountain View are polar opposites, and the narrowness of the former can and has been exploited repeatedly by Google. Indeed, a good acquisition strategy for Google is to enter markets when they are "nascent" and it is easier to clear regulatory hurdles, and then build a powerful position by integration with their overall data and resources.

68. This strategy seems to be at work once again in the current NPRM and in Google's related lobbying efforts at the FCC. While the television and video programming distribution industries are certainly not nascent, Google is once again benefitting from discussion and analysis that seems to be focused entirely on just those industries alone, and not on the myriad other data sources that Google would certainly bring to bear in its navigation device, nor on the competitive benefits the data from this device would bring to its other products and services. Google is clearly a company of unbounded ambition, and unlike regulators, they are not prone to view one market as separate from any other, but to instead adopt an aggressively holistic strategy… and in a technology-driven, data-centric world, there is every reason for them to do so.

---

[33] Or more specifically, the even narrower notion of "in-app" advertising. The FTC declined to challenge the acquisition only after Apple acquired the small AdMob competitor Quattro, deciding that Apple would become a formidable competitor to Google. This assessment turned out to be incorrect, as Apple did not have the advertising infrastructure or expertise to compete.

7. Conclusions

69. In this paper we have argued that Google has systematically and strategically grown its initial dominance in search and advertising to encompass a data empire that is unrivaled in its volume and diversity. We have outlined the technical reasons why this volume and diversity are essential to Google's vast and growing competitive advantages, and the central role that machine learning and related fields play. Throughout, we have shown how Google's lobbying efforts to disintermediate MVPDs are an attempt to obtain yet another valuable consumer data source, one of the few not already in their possession. Finally, we have argued that implementing these regulatory changes will unfairly and dramatically disrupt the current "Golden Age" of video and television content production, distribution and advertising, resulting in substantial harms to consumers.

I declare under the penalty of perjury under the laws of the United States that the foregoing is true and correct.

Michael Kearns
Professor of Computer and Information Science
University of Pennsylvania
April 20, 2016