Network Structure of the Digital Advertising Marketplace

Venkatesh Subramanian, Chun-Yen Pan, Mckayla Sharp, Sarah Stukalin, Xiaoyan Zhang

Abstract

Motivated by the urgent need to regulate digital platforms, our research depicts a descriptive analysis of the structure of the online advertising (Adtech) marketplace which acts as a financial backbone of the digital ecosystem. The Adtech marketplace's structure was studied through scraping the ads.txt of a sample of online news sources. The data from the sample were cleaned and analyzed, and the resulting data was used to create a graph database on Neo4j. We provide a flavor of the analysis that can be done with this data using the mathematics of the 'Method of Reflections' and set the stage for scaling up our analysis to big data methods.

Introduction

Advertising on digital platforms began in the 1990's, creating an online advertising technology (Adtech) marketplace. This marketplace has existed generally unregulated, but now several bipartisan bills have been introduced in DC proposing to regulate Adtech via Antitrust regulations. The purpose of this research is to analyze the structure of the Adtech marketplace to contribute to the literature in terms of structural analysis. Through analyzing the structure of the marketplace, this data can be used to support or reject the price analysis that policy makers are currently using to determine if Antitrust policy needs to be enacted.

In order to analyze the structure of the Adtech marketplace, we chose to examine digital advertising on newspaper websites through concatenating multiple online newspaper directories: Jasmine directory, W3 Newspapers directory, and United States Newspaper Listing (USNPL) directory. These directories create a list of newspaper websites from which we scraped the data via the ads.txt extension in URLs. The ads.txt extension provides data on the advertising exchange, specifically the supply side through which websites sell their advertising opportunities to markets who want to buy advertising space. Through the advertising exchange, the websites sell their slots directly or indirectly (as a reseller) to marketers who want to buy the slots to display their advertisements. From the list of newspaper websites, we deduplicated the data and removed newspaper websites that did not have the ads.txt extension. Next, we wrote code to scrape the newspaper directories to collect the ads.txt data.

To view the data for structural analysis, we created a database on Neo4j. This database allows the relationships between any publisher and platform in the dataset to be examined. Further, we explored the quantity of unique relationships for both direct and reseller relationships in the advertising exchange and used descriptive analysis to compare the two. Because there are distinct differences between direct and reseller relationships, the data collected may elucidate if a specific type of relationship dominates over the other. We also setup the mathematics of the 'Method of Reflections' to analyze the bipartite network database we created on Neo4j.

Motivation

Several bipartisan bills were recently introduced in the House and the Senate to regulate digital platforms, such as advertising technology marketplaces (*Competition and Transparency in Digital Advertising Act* 2022; *American Innovation and Choice Online Act* 2022; *Platform Competition and Opportunity Act* of 2021, 2021; *ACCESS Act* of 2021, 2021). These electronic trading marketplaces facilitate marketers and publishers to trade digital advertisement spaces.

The impetus behind the bills is the perceived market failure due to market concentration. In the case of digital advertising, just three firms have held ~64% market share for three years in a row and this trend is projected to continue (Lebow 2021). In addition, the market "exhibits characteristics that would trigger concerns in other electronic trading markets: market growth is distorted, trading costs - between 30% to 50% of the trade - are high and non-transparent, and conflicts of interests abound." (Srinivasan 2020).

The internet is playing an increasing role in the US economy. As per the latest industry estimates, in 2020, the internet economy contributed to 12% of the US GDP while in 2016 it contributed 6% (Deighton and Kornfeld 2021). Within the internet economy, the Adtech marketplaces comprise a significant revenue share: 81% of Alphabet's revenue (Alphabet Inc 2021) and 97% of Meta's revenue (Meta Platforms Inc 2021) came from digital ads in 2021. Thus, it is important to address any market failures in the Adtech marketplace, which is the intent of the aforementioned bills.

But many of these bills create structural interventions via broad antitrust regulations without addressing the unintended consequences (Chin 2022). These consequences may include barriers to entry in the marketplace and the stifling of innovation. Anticipating these unintended consequences is complicated by the precarious state of current antitrust scholarship which is actively debating Structuralist and Consumer welfare theories (Khan 2017).

The precarious state of Antitrust scholarship

As explained by Khan (2017), Antitrust scholarship is split across two competing theories - Structuralist theory and Price theory. Structuralists (also called "Neo-Brandiesians") believe that the structure of the market determines good market outcomes. They believe that this structure should be free of concentration to determine a fair price for the consumer.

On the other hand, Price theorists believe that as long as price for the consumer (often referred to as 'Consumer Welfare') is favorable, structure of the market shall remain as is, even if it means that the market is concentrated at a given point in time. They believe that market outcomes (price) must determine structure and not the other way round as the Neo-Brandiesians would have it.

Khan (2017) claims that the Sherman Act was originally introduced with Structuralist intentions. She adds that the ‘Consumer Welfare’ school of thought, championed by the Chicago school of price theorists, ended up diverting the original structural paradigm and made the Price paradigm the dominant one. Detractors disagree (Dorsey, 2020) with this. This debate has led to precariousness in the current antitrust scholarship making antitrust regulations a questionable affair.

Numbers for the Neo-Brandiesians

In order to provide data to clarify Antitrust scholarship, one needs to add empirical evidence to both the Structuralist and Price theories. This project’s specific scope is to add evidence to the Structuralist theory championed by the Neo-Brandiesians. We collected data on the structure of the Adtech marketplace, without regard to the price. We then proceed to model the data using a network structure and provide a flavor for the kind of analysis that is possible with this dataset.

Aim

The aim of our research is to map the network structure of the digital banner advertising market. This is done by web scraping across websites that display digital banner ads. Since there are no APIs for this web scraping, we developed our own code to facilitate the scraping, taking inspiration from the template code provided by IAB (Interactive Advertising Bureau, 2022). Given that this is a fairly novel undertaking, we aimed to reduce the scale of the project to only focus on newspaper websites. However, we set up the code and infrastructure to scale up our research later to map the network structure of the rest of the Adtech marketplace. In addition, we also set up the mathematical model of the ‘Method of reflections’ (Hausmann et al., 2014) that enables the analysis of the network structure. This setup should give a glimpse of the analytical possibilities to analyze the network structure of the Adtech marketplace.

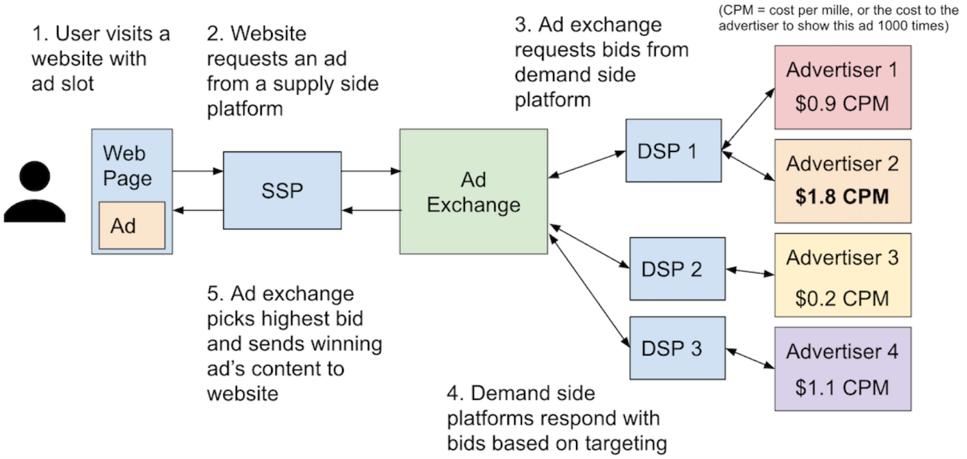
Theory

In this section, we set up the theoretical backdrop necessary for designing our methods and analysis in the rest of this report.

The Adtech Marketplace

Advertising technology (AdTech) refers to the software and tools that companies and publishers use to publish and deliver online advertisements to consumers. This system, which has existed since the 1990’s, is referred to as the AdTech marketplace, or digital advertising marketplace. However, the marketplace today is changing rapidly (Gordon et al., 2021). The online digital marketplace is defined by four parts: advertisers, Demand Side Platforms (DSPs), Supply Side Platforms (SSPs), and publishers and functions as shown in figure 1.

Figure 1: *An illustration of the stakeholders involved in the AdTech marketplace*


1. User visits a website with ad slot
2. Website requests an ad from a supply side platform
3. Ad exchange requests bids from demand side platform
(CPM = cost per mille, or the cost to the advertiser to show this ad 1000 times)
Web Page
Ad
SSP
Ad Exchange
DSP 1
DSP 2
DSP 3
Advertiser 1
$0.9 CPM
Advertiser 2
$1.8 CPM
Advertiser 3
$0.2 CPM
Advertiser 4
$1.1 CPM
5. Ad exchange picks highest bid and sends winning ad’s content to website
4. Demand side platforms respond with bids based on targeting

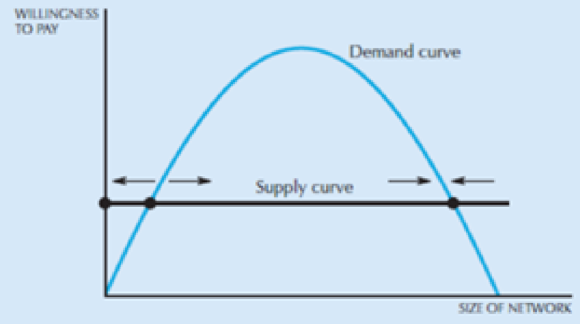
***Note.* From *Why “bad” ads appear on “good” websites – a computer scientist explains,* by Eric Zeng, 2022, The Conversation.**

Markets with Network Externalities

In a naive model of demand and supply, the number of buyers will decrease when the price increases, while the number of buyers will increase when the price decreases. The demand curve is a standard downward-sloping curve. In this case, people's willingness to buy a good is not only affected by price but also by whether the others are willing to buy the good. But when network externalities become a concern as it does in the Adtech marketplace, this argument breaks down. In the Adtech marketplace, it is natural for marketers to flock to the platform that has a good network of publishers. The greater the number of publishers, the higher the demand on the platform.

In general, in such marketplaces, when the demand increases, the demand curve slope will increase, as seen in Figure 2. After a period of time, a large amount or majority of the population will have the goods, and the network externality gets estimated – the capacity for demand has been satiated. In other words, after a platform acquires a critical mass of publishers all marketers will naturally flock to that platform hence creating a natural monopoly.

Figure 2: *An illustration of the demand curve based on the size of the network on the x-axis and willingness to pay on the y-axis.*



Network externalities. The demand is given by the curved hump, the supply by the horizontal line. Note that there are three intersections where demand equals supply.

***Note.* Varian, H. R. (2014). *Intermediate microeconomics: a modern approach: ninth international student edition.* WW Norton & Company.**

Two-Sided Markets and Networks

A two-sided market is the platform that allows users to trade, enabling bilateral or multilateral markets to remain on the platform (Easley & Kleinberg, 2010). For each transaction on such a platform, it charges a fee from both buyers and sellers. If the buyers and sellers have different price elasticities, the market is two-sided. Price elasticity refers to the effect of a small change in price on the buyers and sellers. If price change has no influence on one side, it can be said that that side has low price elasticities.

This theory can be applied to the online advertising marketplace. For example, if advertisers have low price elasticities, the platform will charge money from advertisers. Conversely, consumers have high price elasticity because they are sensitive to price change. Platforms usually give consumers free access or provide a large discount. Consider Google ads, which are free for consumers. When there are many people searching for ads that are published on this platform, it will attract advertisers to post their ads. In the section on 'Descriptive statistics' we show some preliminary evidence hinting how the traders in the middle play an oversized role in the Adtech marketplace.

Network Theory

The network model is a flexible way in which the database model is conceived to represent objects and their relationships. It describes the method of structuring and manipulating data in the database. Network model is unique in that it is not limited to hierarchies but is viewed as a graph of object-type nodes and relationship-type arcs. In the network model, some components may be very well connected in a single network, while others may not. The model has greater power to represent connections among system components, nodes and edges, more realistically. Given these properties, we believe a network theoretic paradigm might be well suited to add evidence to Structural theory which is the aim of this project. In this section, we set up the technical terminology of Network Theory required in the other sections of this report.

The term *Networks* and *Graphs* are used synonymously in this paper. The network model consists of three different elements: the node representing the constituent elements of the system, the edges that reflect the relationship between the constituent elements, and the traffic flowing in the network. Additionally, it reflects both the quantitative relationship between elements and determines the goal and direction of network model optimization. The degree of a node in a network is the number of connections or edges between that node and other nodes. If a network is directed, meaning that edges point from one node to another in one direction, then nodes have two distinct degrees. A network consists of some nodes and edges that connect them. The number of all edges connected by each node is the degree of this node. Degree distribution refers to the overall description of the node degree in a network.

A graph whose nodes can be divided into two (or *n*) groups so that no edge connects nodes within each group, is called a *bipartite* (or n-partite) graph (Sayama, 2015). Our data collection methods here result in a bipartite graph of two nodes, namely publishers and platforms.

Edge list is an array, listing, or serial, that records the edge between all points. We construct an edge list in the creation of our Neo4j database.

Centrality is a common concept in social network analysis. As soon as the social actor gets a higher centrality, it means they get closer to the center of network, that higher power, influence, convenience from the network they may acquire (Hochberg et al., 2007), which presents the importance of the centrality in the network. In the analysis section, we establish the method of reflections which essentially creates centrality measures of higher order nuances.

Literature Review

In this section, we review some of the previous work done on a similar dataset.

Compliance

Previous studies have observed the effects of ads.txt on helping ad buyers avoid ad fraud from spammers who sell counterfeit ads and claim to be from high-value publishers (Bashir et al. 2019). The research utilizes ads.txt compliance as a measure for the potential for combating fraud. Bashir et al. (2019) found that from the sample of Alexa Top-100K, 60% of websites adopted ads.txt who participate in displaying ads via real time bidding. The researchers conclude that this is an impressive adoption since ads.txt was founded two years prior to the study, in 2017. The researchers conclude that the high compliance rate suggests that “SSPs and ad exchanges are honoring the standard by not attempting to sell unauthorized inventory” (Bashir et al. 2019). However, this study is limited by not being able to test fraud levels directly.

Further, ads.txt has been examined from a transparency perspective to test whether the data can be used for research and privacy advocates. The study highlights that previously, the only way to research the online ad ecosystem was through heuristics or crowdsourcing data (Bashir et al. 2019). With ads.txt, the datasets are useful by being able to aggregate ads.txt files, from the seller-side, and even more useful if coupled with inclusion data on the buyer-side platforms.

Finally, (Bashir et al. 2019) highlight limits to using ads.txt data. They found that there are many errors in that data. Additionally, the study found that the seller domains listed on ads.txt do not include all seller domains and that additional manual work is required to map to all seller domains (Bashir et al. 2019).

Fake News

Ads.txt data has also been used to study who is responsible for monetizing news websites via ads, specifically in regards to fake news. Using ads.txt, Papadogiannakis et al.(2022) mapped the relationship between the top 10 most popular digital ad sellers for the DIRECT relationship, which ads.txt defines as where the publisher is the owner of the specified account. The study found that "fake news websites form direct business relationships with 27 ad systems, while surprisingly, real news websites do so with 41 systems" (Papadogiannakis et al. 2022). Evidently, through ads.txt, Papadogiannakis et al. (2022) is able to map the network of business relationships between news websites (both real and fake news sites) to the ad networks.
Ads.txt is based on data given from the websites themselves. Sellers.json is complementary, however, the data is provided from the sellers themselves. Papadogiannakis et al. (2022) utilized the sellers.json files to verify the business relationships found in ads.txt. However, it is important to note that there were found to be some discrepancies between the data (Papadogiannakis et al.

2022).

Operationalization of Marketplace Structure

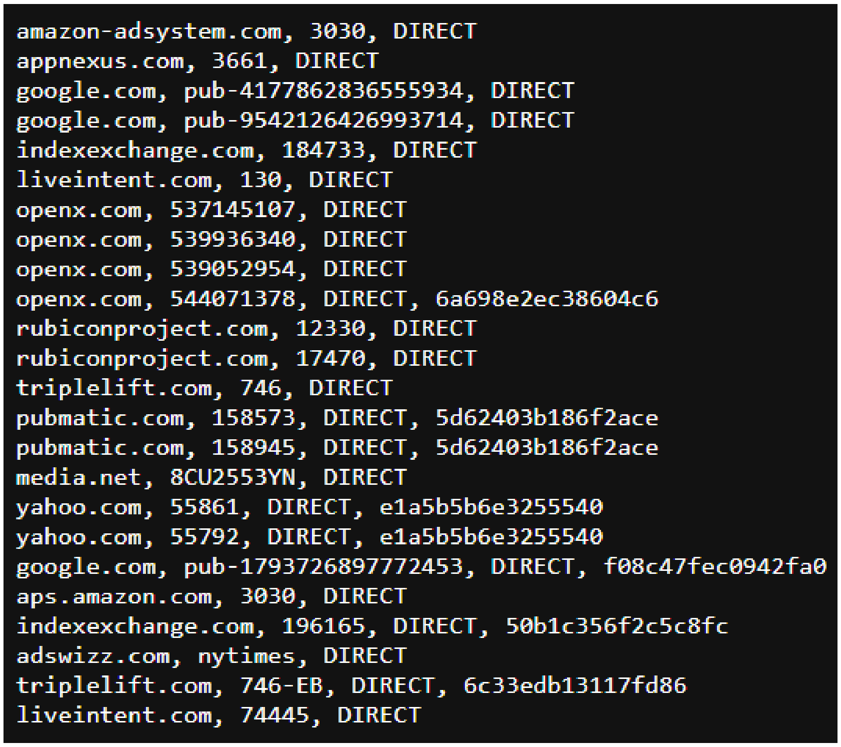
While previous literature has used ads.txt to study compliance rates and fake news, we use it to study market structure. So from our perspective, the ads.txt dataset is an operationalization of the market structure. Below we explain the specifications of this dataset.

Explanation of Data: ads.txt

The ads.txt extension of the base url contains three or four data points. Sequentially, these data points are the name of the seller, the publisher ID, and whether the seller is direct or a reseller (IAB Tech Lab, *ads.txt*, 2019). The direct or reseller relationship indicates “whether the publisher is the contractual owner of the advertising account in [column] 2 (former) or that the publisher has contracted with a third-party to manage the account” (Bashir et al. 2019). Some file formats may also include an optional fourth column. This field is referred to as the TAG or the Certification Authority ID. The Certification Authority ID is “an ID that uniquely corresponds to the company in [column] one” (Bashir et al. 2019). Figure 3 shows an example acquired from nytimes.com/ads.txt.

While duplicates are not prohibited within an ads.txt file, it may provide inaccurate information. The single ad exchange/SSP is unlikely to actually have that many working relationships with the publisher under each of the IDs. Thus, duplicates would be considered errors. For our research, these errors were resolved during the data cleaning process by de-duplication (Bashir et al., 2019).

Figure 3: *An example of the nytimes.com/ads.txt extension from which the data was collected.*



```
amazon-adsystem.com, 3030, DIRECT
appnexus.com, 3661, DIRECT
google.com, pub-4177862836555934, DIRECT
google.com, pub-9542126426993714, DIRECT
indexexchange.com, 184733, DIRECT
liveintent.com, 130, DIRECT
openx.com, 537145107, DIRECT
openx.com, 539936340, DIRECT
openx.com, 539052954, DIRECT
openx.com, 544071378, DIRECT, 6a698e2ec38604c6
rubiconproject.com, 12330, DIRECT
rubiconproject.com, 17470, DIRECT
triplelift.com, 746, DIRECT
pubmatic.com, 158573, DIRECT, 5d62403b186f2ace
pubmatic.com, 158945, DIRECT, 5d62403b186f2ace
media.net, 8CU2553YN, DIRECT
yahoo.com, 55861, DIRECT, e1a5b5b6e3255540
yahoo.com, 55792, DIRECT, e1a5b5b6e3255540
google.com, pub-1793726897772453, DIRECT, f08c47fec0942fa0
aps.amazon.com, 3030, DIRECT
indexexchange.com, 196165, DIRECT, 50b1c356f2c5c8fc
adswizz.com, nytimes, DIRECT
triplelift.com, 746-EB, DIRECT, 6c33edb13117fd86
liveintent.com, 74445, DIRECT
```

Note. https://www.nytimes.com/ads.txt

An example of ads.txt is seen in Figure 3, from nytimes.com/ads.txt. Line one of Figure 3, “amazon-adsystem.com, 3030, DIRECT”, is data that informs that the seller, amazon-adsystem.com (as seen in column one of line one) is an authorized ad network (digital seller), or SSP, with the publisher, nytimes.com. This means that the publisher, nytimes.com, authorizes amazon-adsystem.com to sell their ad space inventory. The publisher ID, located in column two of line one and reads 3030, is a “string that uniquely identifies the publisher’s account within the ad system hosted by the company” (Bashir et al. 2019). In column three of line one, “DIRECT” informs the viewer that amazon-adsystem.com is a direct seller to the publisher, nytimes.com. The fourth item, if included, indicates the certification authority ID (TAG). Although line 1 “amazon-adsystem.com” does not include a TAG, you can reference line

10 “openx.com, 544071378, DIRECT, 6a698e2ec38604c6” for an example of the TAG. In this case, the TAG is 6a698e2ec38604c6.

Methods

Data Collection

Data was collected via a list of URLs of news sources to act as a sampling frame to facilitate the scraping of ads.txt files from each of these URLs. To collate this list of URLs we first web scraped three URL directories: Jasmine directory (*Jasmine directory*, 2022), W3newspapers (*W3Newspapers - World Newspapers, News Sites, and Magazines.*, 2022), and USNPL (*USNPL*, 2022). To define the sampling frame, we collated URLs of newspapers from all the 50 states of the United States, as well as the District of Columbia (DC). This ensures that the sample is representative of both local and national news sources in the USA.

This data collection process has two main limitations. Due to a lack of information regarding documentation, it is unclear how the aforementioned directories were created. Specifically W3 newspapers and USNPL, wherein if the makers of these directories were biased to collect a certain type of news source our data might suffer from the same bias. The second limitation is the partial compliance with the ads.txt standard (Bashir et al., 2019), as explained in the following section. We found that many news sources lacked an ads.txt file. In addition, many of the complying domains did not follow the formatting standards outlined for ads.txt. It is also possible that publishers do not update their ads.txt files regularly, resulting in misleading data.

Data Cleaning

Once we obtained the initial list of URLs, they needed to be cleaned before scraping for the ads.txt files. As per the standards set by IAB (*ads.txt Version 1.1*., 2022), the ads.txt file has to be nested under the base URL (also called root domain) only. First, we *extracted the base URLs* from the initial list. For instance, if the initial list had “https://example.com/somepage” our base URL extraction process led to “example.com”. After base URL extraction, we went through a process of *enriching the dataset* with details about the states and topics of the news sources, wherever such information was available from the directories. This enrichment procedure was followed by *deduplication of URLs* especially when it got repeated across directories.

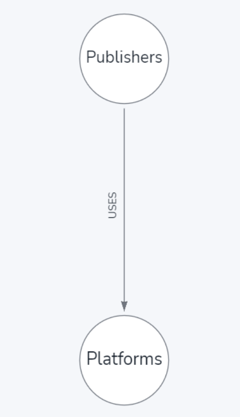
URLs on the internet are often dynamic– new URLs get created and old URLs go offline frequently. Because of the dynamics of URLs, we determined it was pertinent to ensure that all the URLs in the dataset were valid. Our *URL validation process* involved creating a HTTPS GET request to, for instance, ‘https://example.com/ads.txt’. If the response from this request had a 200 status code then we considered the URL to be valid and retained it for further scraping, else removed it. Further, we also had issues from the URLs even if they were considered valid, such as the problem of redirection. A valid URL that gives a 200 code could still be redirecting to a different URL in which case the ads.txt should be nested under the base URL of the redirected URL. We accounted for this by getting the redirected URL of all URLs in the valid URL list, repeating the URL validation process for the redirected URLs and creating a mapping between the original URL we had collected from the directory to the valid redirected URL if it existed. For the rest of the process, we used these validated redirected URLs.

After cleaning, verifying, and solving the problem of redirection we extracted the ads.txt data from the URLs. However, as per IAB standards (*ads.txt Version 1.1*., 2022), some of the websites had placeholder ads.txt files. We removed all placeholder files. In addition, wherever records in the ads.txt file didn’t conform to the standards of IAB, we removed these records. After removing placeholder files and files that didn’t conform to IAB standards, we combined the extracted ads.txt data all into a final edge list (explained in section on ‘Network Model’) to create our network structure.

Neo4j Database creation

Following the data collection, through scraping of the ads.txt files, each publisher was mapped to a unique key. The unique key is called the publisher key, and each platform was connected to a unique key called the platform key. From this, three files were created – ‘publishers.csv’, containing a publisher’s base URL and publisher key; ‘platforms.csv’, containing a platform’s base URL and platform key; and ‘uses.csv’, containing the aforementioned edge list to indicate which publisher ‘uses’ which platform. From these files, we created a Neo4j database. Neo4j is “a native graph database” (*Neo4j,* 2022*)* which “stores and manages data in its more natural, connected state, maintaining data relationships that deliver lightning-fast queries, deeper context for analytics, and a pain-free modifiable data model” (*Neo4j,* 2022*)*. This database creation will help in scaling up this database in the future, as necessary. Figure 4 presents the graph data model.

Figure 4: *Neo4j graph model showing how Publishers is mapped to Platforms with the edge defined as Uses.*



Analysis & Results

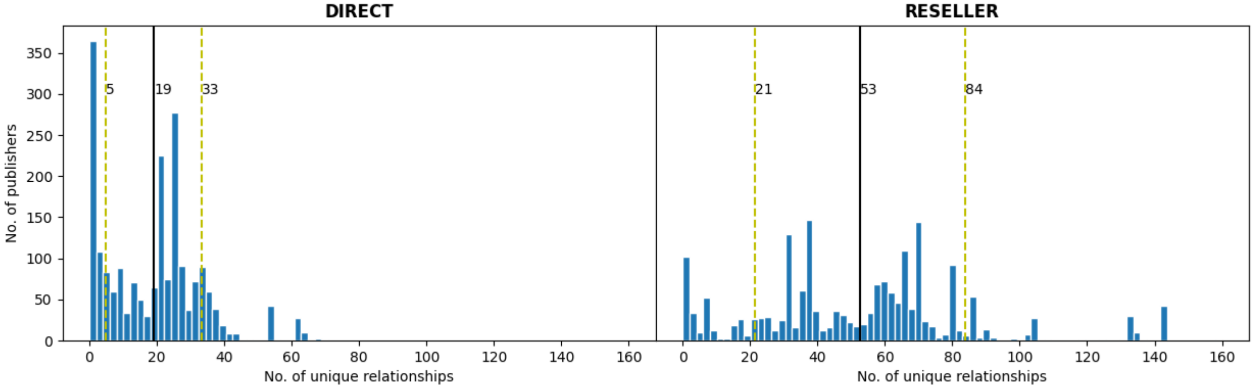
In this section we provide some descriptive statistics and set up the mathematics for ‘The Method of Reflections’. While we don’t provide substantial results, we hope this section sets up the flavor of analysis for future research.

Descriptive Statistics

The network dataset contains 2050 publisher nodes which use 628 platform nodes. In terms of edges, there are 137,038 unique edges and 1,064,527 non-unique edges. The unique edge count is based on the number of platforms a publisher uses. The non unique count is based on the number of accounts each publisher has on a platform. On an average, each publisher has 8 different accounts on a platform. We are unsure why this is the case. We speculate that each platform provides different types of ad related services and each of these services requires a separate account from the publisher.

In Figure 5, we plot a histogram of the number of unique direct and reseller relationships between publishers and platforms. The dark line indicates the mean and the yellow dashed lines indicate one standard deviation below and above the mean. As can be seen, the mean number of direct relationships is much lesser than the mean number of reseller relationships. This tells us that many publishers go through intermediaries to sign up with the platforms. This is an indication that the platforms in the middle matter. This is further validation of the theory of multi-sided markets.

Figure 5: *Direct relationship histogram on the left and Reseller relationship histogram on the right. Each graph displays the number of unique relationships with publishers.*



Method of Reflections

Having described the network and provided some validation of the multi-sided market theory, we will now analyze the bipartiteness of the network. In the literature on Economic Complexity (Hausmann et al., 2014), a general mathematical framework for studying bipartite networks has been introduced. We apply that framework for the bipartite network of publishers and platforms in the following sections. We highly recommend the interested reader to read Haussmann et al. (2014) for a deeper understanding of mathematics, since our treatment here follows their treatment of the subject very closely.

Adjacency matrix for a bipartite network

Let the publisher nodes be denoted by p and platform nodes be denoted by c. Then the adjacency matrix describing this bipartite network is denoted by M_{cp} where $M_{cp} = 1$ if publisher p uses platform c else $M_{cp} = 0$. This creates a matrix of dimensions $n_p \times n_c$ where n_p is the number of publishers and n_c is the number of platforms. Note that a general adjacency matrix is a square matrix whose dimensions are given by the total number of nodes. But M_{cp} is an adjacency matrix that is defined specifically to bring out the bipartiteness of our network to the forefront of the analysis. In the passages that follow, adjacency matrix means this special M_{cp} matrix (unless specified otherwise).

Diversity of publishers and Ubiquity of platforms

There are two essential concerns to keep in mind to study the structure of this marketplace. First, are publishers stuck with one platform or do they have options? Secondly, are platforms stuck with one publisher or can they do business with several publishers effectively? The first concern is captured by a metric called the diversity of publishers. The diversity of publishers counts the number of unique connections a publisher has with different platforms, also referred to as the degree of publishers in the network:

$$k_{p,0} = \sum_{c} M_{cp} \qquad (1)$$

The second concern is captured by a metric called the ubiquity of platforms, which counts the number of unique connections a platform has with different publishers. In other words it is the degree of platforms in the network:

$$k_{c,0} = \sum_p M_{cp} \qquad (2)$$

Average Diversity across publishers and Average Ubiquity across platforms

Now, even if a publisher is connected with many platforms, if all the ads come from one platform (due to market concentration), then losing the connection with just that platform would be fatal for the publisher's business. Similarly even if a platform is connected with many publishers, if one publisher is consistently buying most of the ads from the platform then losing that one publisher would be fatal for the platform's business. So the Diversity and Ubiquity metrics alone can't capture the intricacies of market concentration. There is a need for further characterization. That is, there is a need to characterize the publisher by the average ubiquity of the platforms it is connected to as shown below:

$$k_{p,1} = \frac{1}{k_{p,0}} \sum_c M_{cp} k_{c,0} \qquad (3)$$

In addition, there is a need to characterize the platform by the average diversity of publishers it is connected to as shown below:

$$k_{c,1} = \frac{1}{k_{c,0}} \sum_p M_{cp} k_{p,0} \qquad (4)$$

There is no reason to stop the analysis here. We can create a recurrence relationship as shown below and use average of averages and so on:

$$Generalizing\ (3),\ k_{p,N} = \frac{1}{k_{p,0}} \sum_c M_{cp} k_{c,N-1} \ (5)$$

$$Generalizing\ (4),\ k_{c,N} = \frac{1}{k_{c,0}} \sum_p M_{cp} k_{p,N-1} \qquad (6)$$

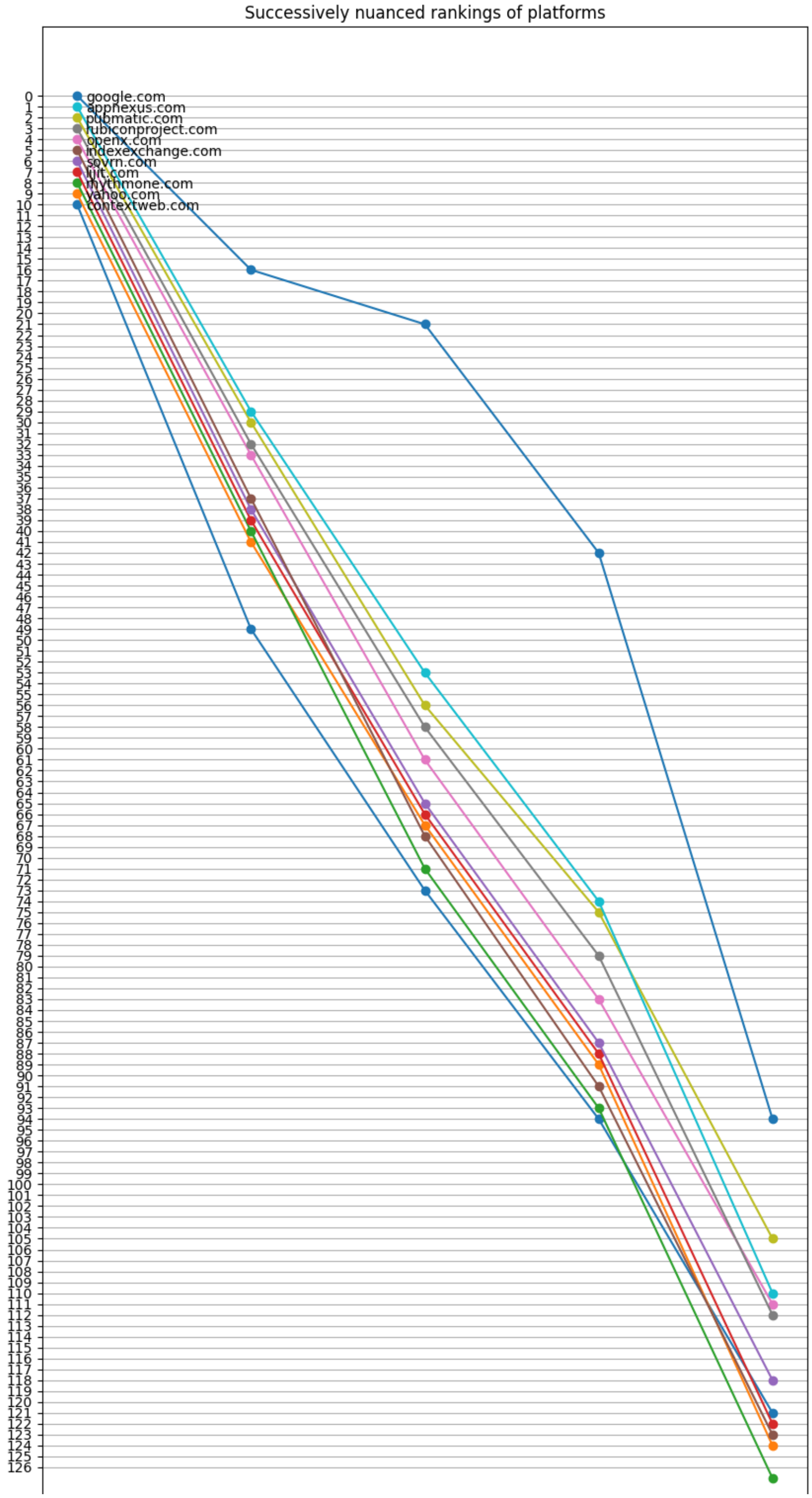
It is interesting to note that these generalizations are essentially centrality measures (described earlier in the ‘’ section). This generalization leads to the creation of $\vec{k}_p = \left(k_{p,0}, k_{p,1}, \ldots, k_{p,N}\right)$ and $\vec{k}_c = \left(k_{c,0}, k_{c,1}, \ldots, k_{c,N}\right)$. Note in $\vec{k}_c$, elements with even subscripts characterize the platforms while those with the odd subscripts characterize the publisher. A similar explanation holds for $\vec{k}_p$. This mathematical analysis can continue further and this problem can be converted into solving an engine problem. But given the scope of this project, we stop the analysis here.

Is the perceived market power just a mirage? – Successive rankings of platforms from the ‘Method of reflections’

The figure below uses 6 to characterize the platforms in our network. This figure specifically characterizes the platforms that have the top 10 uniquity score. That is, the top 10 platforms that are connected to the most number of publishers. Traditional literature on market concentration would only rank the platforms based on ubiquity score (i.e) the element with subscript 0, to characterize the market and claim that the top platforms hold too much market power. But the figure goes a step further and shows how the rank of these platforms evolve as higher order elements are accounted for, specifically those elements with 2nd, 4th, 6th and 8th subscripts in kc. The figure shows that their market leadership has a steady downward trend as higher order ranks are considered. For instance ‘google.com’ which is ranked 1 based on ubiquity score alone, successively drops down to rank 16, rank 21, rank 42 and so on. This figure raises questions on how we should go about characterizing platforms in a multi-sided market. Is the method of reflections a viable characterization? If it is, then looking at the figure, is the market power in this market, just a mirage?

We believe it would be too premature to draw a substantial conclusion from the figure. As was mentioned earlier the scope of this project is to give a flavor for the type of analysis possible. The aim of the figure is to achieve exactly that and nothing more. It might be necessary to observe how the rank evolves further (maybe it goes up at a much higher order?). It would also be necessary to address the criticisms levied against the operationalization provided by the method of reflections that currently exists in the currently nascent literature on Economic Complexity (Hausmann et al., 2014).

Figure 6: *Successively nuanced rankings of platforms with ranks (with rank of 0 being the highest) on the y-axis and iterations of the method of reflections on the x-axis.*



Conclusion

Our project was motivated by the need to reduce the precariousness of current antitrust scholarship and the urgency to act on the problem, given the introduction of the Competition and Transparency in Digital Advertising Act (2022) on the floor of the Senate. Our research acted on this motivation by adding evidence to the Structuralist theory of antitrust. This was accomplished by collecting data on the advertising partners of news sources across the USA by web scraping their ads.txt files. This data was used to create a Neo4j database to analyze the network structure. The Neo4j database can also be utilized to scale the data source in the future and determine how changes between platforms and publishers would change the structure of the marketplace.

With the data collected, 2,050 news sources and 628 platforms connected by 137,038 unique edges, we provide an introduction to the type of analytical possibilities provided by this dataset. We used descriptive statistics to say that reseller relationships are more prevalent in the market than direct relationships indicating that middlemen play a huge role and hence adding credence to the notion of a multi-sided market at play here. Finally we set up the mathematics of the 'Method of reflections' and characterized the top 10 platforms in the marketplace by successive rankings provided by the method. By illustrating a sharp downward trend in successive rankings, we speculate if the market power that is being perceived here is a mirage while pointing out the need for further analysis.

For the sake of this study, we limited our population to online newspapers. This allowed us to apply our research to an appropriately representative population. This population also allowed us to collect a reasonable sample size relative to population, within our means.

An issue we ran into while scraping ads.txt data was a lack of adherence to IAB standards. This obstacle made the cleaning of the collected data more tedious than anticipated.

This is further indication that an increased level of management and cohesion within the marketplace would be beneficial. A second notable issue we ran into was creating a public facing Neo4j database. Although an AuraDB instance is available for free on the Neo4j cloud, a publicly hosted version of the database necessitates financial costs. Hence we failed in deployment although this is just a monetary constraint and not a technical constraint. In place of a public facing database, we have instead provided the .dump file for download in our Github repository and those interested should be able to download it and replicate the database with relative ease.

In the future, we hope to leverage the potential of Neo4j to scale up to big data methods so as to get closer to data on the whole population of publishers instead of relying on a sample of news sources as we did in this preliminary project. This would involve not just scrapping the ads.txt files at the publisher end, but also the sellers.json files at the platform end. Further it would be useful to understand the evolution of the structure here by collating a time series dataset (which is also supported by Neo4j), by leveraging the archival capability of the 'Wayback machine' (*Internet Archive: Wayback Machine*). By illustrating the surprising (although inconclusive) results of the method of reflections in characterizing the multi-sided market here, we have also shown the need for better operationalizations of market power that are needed in the future to make further headwinds on this very important issue that could determine the future of the internet ecosystem.

Synergy report

This project had 3 distinct phases. The first was ideation, the second was the project proposal and the third was the final presentation and writing of this report. The whole team gave it their best in all the 3 stages. All members attended almost every single meeting (which happened once a week on an average) we had and made sure to let everyone know in case they won't be available. Despite personal commitments (work, children, illnesses) everyone stepped up to the task even during Fall break! We worked through language barriers, inexperience in coding among other deficiencies each of us had and gave it our best shot. We all did the best we could with what we had. Some of us contributed significantly to the theoretical background, some of us to the coding parts, some of us to the analysis and some of us greatly enhanced the report and in bringing all the aforementioned pieces together to not lose sight of the big picture.

We wish we could have spent more time practicing our slides before the final presentation to communicate the excitement we had working on this project to the rest of our classmates. But personal commitments and scheduling conflicts made this difficult. But we hope this report is successful in communicating that excitement.

Thanks to Xiaoyan Zhang for contributing to the theory section especially on the parts that needed the expertise of an economist. Thanks to Chun-Yen Pan for contributing to the theory section as well especially on the parts that needed Network theory. Thanks to Sarah Stukalin and Mckayla Sharp for bringing all the pieces together, editing this report with great patience and keeping track of the bigger picture of this project. As a coordinator, I am thankful to all of them for their patience in working with me on this complicated topic with a rich background in Economics, Political science, Public policy and Mathematics (Graph theory).

Code

Our codebase is available at:
VenkiPhy6 (2022). *admapper,* Github repository, https://github.com/VenkiPhy6/admapper

References

ACCESS Act of 2021, H.R.3849, 117th Congress (2021, June 24). https://www.congress.gov/bill/117th-congress/house-bill/3849

ads.txt Version 1.1. (2022). https://iabtechlab.com/wp-content/uploads/2022/04/Ads.txt-1.1.pdf

Alphabet Inc. (2021) *Form 10-K 2021*. Alphabet Inc. https://abc.xyz/investor/static/pdf/20220202_alphabet_10K.pdf?cache=fc81690

American Innovation and Choice Online Act, S.2992, 117th Congress (2022, March 2). https://www.congress.gov/bill/117th-congress/senate-bill/2992

Bashir, M. A., Arshad, S., Kirda, E., Robertson, W., & Wilson, C. (2019, October). A Longitudinal Analysis of the ads. txt Standard. In Proceedings of the Internet Measurement Conference (pp. 294-307) https://doi.org/10.1145/3355369.3355603

Chin, C. (2022). *Breaking Down the Arguments for and against U.S. Antitrust Legislation | Center for Strategic and International Studies*. https://www.csis.org/analysis/breaking-down-arguments-and-against-us-antitrust-legislation

Competition and Transparency in Digital Advertising Act, S.4258, 117th Congress (2022, May 19). https://www.congress.gov/bill/117th-congress/senate-bill/4258

Deighton, J., & Kornfeld, L. (2021). *The Economic Impact of the Market-Making Internet*. https://www.iab.com/insights/the-economic-impact-of-the-market-making-internet/

Dorsey, E. (2020). Antitrust in Retrograde: The Consumer Welfare Standard, Socio-Political Goals, and the Future of Enforcement. *The Global Antitrust Institute Report on the Digital Economy*, *4*. https://ssrn.com/abstract=3733666

Easley, D., & Kleinberg, J. (2010). Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge university press.

Gordon, B. R., Jerath, K., Katona, Z., Narayanan, S., Shin, J., & Wilbur, K. C. (2021). Inefficiencies in Digital Advertising Markets. Journal of Marketing, 85(1), 7–25. https://doi.org/10.1177/0022242920913236

Hausmann, R., Hidalgo, C. A., Busots, S., Coscia, M., & Simoes, A. (2014). *The atlas of economic complexity : mapping paths to prosperity*. The MIT Press.

Hochberg, Yael V., Ljungqvist A., & Yang, Lu. (2007). Whom You Know Matters: Venture Capital Networks and Investment Performance. 251–301. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2007.01207.x

Interactive Advertising Bureau (2022). adstxtcrawler, *GitHub repository*, https://github.com/InteractiveAdvertisingBureau/adstxtcrawler

Internet Archive: Wayback Machine. (n.d.). Internet Archive. Retrieved October 10, 2022, from https://archive.org/web/

Jasmine directory (2022). https://www.jasminedirectory.com/

Khan, L. M. (2017). Amazon's Antitrust Paradox. *The Yale Law Journal*, *126*, 710–805. https://papers.ssrn.com/abstract=2911742

Lebow, S. (2021, November 3). *Google, Facebook, and Amazon to account for 64% of US digital ad spending this year*. EMarketer. https://www.insiderintelligence.com/content/google-facebook-amazon-account-over-70-of-us-digital-ad-spending

Meta Platforms Inc. (2021) *Form 10-K 2021*. Meta Platforms Inc. https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/14039b47-2e2f-4054-9dc5-71bcc7cf01ce.pdf

Neo4j (2022). https://neo4j.com/

Papadogiannakis, E., Papadopoulos, P., Markatos, E. P., & Kourtellis, N. (2022). Who Funds Misinformation? A Systematic Analysis of the Ad-related Profit Routines of Fake News sites. arXiv preprint arXiv:2202.05079

Platform Competition and Opportunity Act of 2021, H.R.3826, 117th Congress (2021, June 24) https://www.congress.gov/bill/117th-congress/house-bill/3826

Sayama, H. (2015). *Introduction to the Modeling and Analysis of Complex Systems.* Basics of Networks, 295–404. https://lib.hpu.edu.vn/handle/123456789/21480

Sellers.json. (2019). https://iabtechlab.com/wp-content/uploads/2019/07/Sellers.json_Final.pdf

Srinivasan, D. (2020). Why Google Dominates Advertising Markets Competition Policy Should Lean on the Principles of Financial Market Regulation. *STANFORD TECHNOLOGY LAW REVIEW*, *24*(1), 55–175. https://law.stanford.edu/wp-content/uploads/2020/12/Srinivasan-FINAL-Why-Google-Dominates-Advertising-Markets.pdf

USNPL. (2022). United States Newspaper Listing . https://www.usnpl.com/

Varian, H. R. (2014). *Intermediate microeconomics: a modern approach: ninth international student edition.* WW Norton & Company.

W3Newspapers - World Newspapers, News Sites, and Magazines. (2022). W3Newspapers. https://www.w3newspapers.com/

Zeng, E. (2022, April 13). *Why "bad" ads appear on "good" websites – a computer scientist explains*. The Conversation. https://theconversation.com/why-bad-ads-appear-on-good-websites-a-computer-scientist-explains-178268

Appendix

An appendix providing further explanations, especially of the rich theory here can be found in the following Google document.