The Inventory is Dark and Full of Misinformation: Understanding Ad Inventory Pooling in the Ad-Tech Supply Chain

Yash Vekaria
University of California, Davis

Rishab Nithyanand
University of Iowa

Zubair Shafiq
University of California, Davis

***Abstract*—Ad-tech enables publishers to programmatically sell their ad inventory to millions of demand partners through a complex supply chain. The complexity and opacity of the ad-tech supply chain can be exploited by low-quality publishers (e.g., misinformation websites) to deceptively monetize their ad inventory. To combat such deception, the ad-tech industry has developed transparency standards and brand safety products. In this paper, we show that these developments still fall short of preventing deceptive monetization. Specifically, we focus on how publishers can exploit the ad-tech supply chain, subvert ad-tech transparency standards, and undermine brand safety protections by pooling their ad inventory with unrelated sites. This type of deception is referred to as "dark pooling." Our study shows that dark pooling is commonly employed by misinformation publishers on various major ad exchanges, and allows misinformation publishers to deceptively sell their ad inventory to reputable brands. Our work suggests the need for improved vetting of ad exchange supply partners, the adoption of new ad-tech transparency standards that enable end-to-end validation of the ad-tech supply chain, and the widespread deployment of independent audits like ours.**

1. Introduction

The complexity of online advertising lends itself to fraud. A key to the success of online advertising is the ability of advertisers and publishers to programmatically buy and sell ad inventory across hundreds of millions of websites in real-time [1]. Notably, Real-Time Bidding (RTB) allows publishers to list their ad inventory for auction at an ad exchange [2]. The ad exchange then asks its demand partners to bid on the ad inventory listed by its supply partners, based on the associated contextual and behavioral information. The ad-tech supply chain is complex because it relies on hundreds of specialized entities to effectively buy and sell the ad inventory in real-time and at scale [3]. Adding to this complexity, each ad impression often gets sold and resold through multiple parallel or waterfall auctions [4]. Such scale and complexity, combined with the opaque nature of the ad-tech supply chain, makes it a ripe target for fraud and abuse [5]–[13]. One of the most common types of ad fraud involves creating low-quality websites and monetizing their ad inventory. Fraudsters attempt to drive large volumes of traffic to their website through various illicit means such as bots, underground marketplaces, traffic exchanges, or even driving legitimate traffic through click-bait and viral propaganda [14]–[16]. A notable example that motivated our work is that of the "Macedonian fake news complex" [17]–[19]. In this scheme, fraudsters created misinformation websites with misleading and clickbait headlines, aiming to go viral on social media, which led to tens of millions of monetized ad impressions.

Advertisers are invested in preventing fraud. Ad-tech has safeguards to protect against this type of ad fraud by blocking the ad inventory of low-quality websites even when the ad impressions might be from legitimate users. Specifically, brand safety features supported by demand-side platforms aim to allow advertisers to block ad inventory of web pages that contain hardcore violence, hate speech, pornography, or other types of potentially objectionable content [20]. All the effort of fraudsters would be wasted if they are unable to monetize their ad inventory through programmatic advertising due to these brand safety features. Fraudsters are known to exploit the opaque nature of the complex ad-tech supply chain to undermine brand safety protections by misrepresenting their ad inventory [21]. For example, in domain spoofing [22], low-quality publishers mimic the URLs of reputable publishers in their ad inventory, thus deceiving reputable brands into purchasing their ad space even when their original domain is blocked due to brand safety concerns [23]–[25]. To combat ad fraud resulting from misrepresented ad inventory, the Interactive Advertising Bureau (IAB) introduced two transparency standards. `ads.txt` [26] requires publishers to disclose all authorized sellers of their ad inventory. `sellers.json` [27] requires ad exchanges to disclose all publishers and intermediate sellers involved in selling the ad inventory. Together, when correctly implemented, these standards can reduce ad fraud by enabling buyers to verify the sources of the inventory they are purchasing.

Transparency mechanisms to prevent fraud are falling short. There is increasing concern that the `ads.txt` and `sellers.json` standards are either not widely adopted, implemented in ways that do not facilitate effective supply-chain validation, or intentionally subverted by malicious actors in a variety of ways. In this paper, we empirically investigate these concerns. We find that the `ads.txt` and `sellers.json` disclosures are plagued by a large number of compliance issues and misrepresentations. Most notably, we find extensive evidence of "pooling" of ad inventory

from unrelated websites — a practice known in the industry as "dark pooling." This makes it impossible for a buyer to reliably identify the sources of the ad inventory (i.e., where their ad will ultimately be placed). Dark pooling effectively enables low-quality publishers to "launder" their ad inventory, making it indistinguishable from that of well-reputed publishers. To gain insight into how low-quality publishers might circumvent the transparency required by the `ads.txt` and `sellers.json` standards, we selected a set of well-known misinformation websites as a case study. This choice is motivated by the known instances where ads from reputable brands have inadvertently ended up on such websites in the past [28]–[33]. Focusing on these misinformation websites, we confirm: (1) their widespread failure to comply with the `ads.txt` and `sellers.json` standards; and (2) widespread prevalence of ad inventory pooling. We also find instances of reputable brands buying ad impressions on these misinformation websites, perhaps unintentionally. Taken together, we make three key contributions.

Measuring compliance with the transparency standards of `ads.txt` and `sellers.json`. We study a set of control and well-known misinformation websites to compare their compliance with `ads.txt` and `sellers.json`. We find that although compliance issues are widespread even in the control set of websites, they are significantly more prevalent on misinformation websites.

Measuring the prevalence of (dark) pooling. We measure the high prevalence of ad inventory pooling by our control and misinformation websites. By analyzing the `ads.txt` and `sellers.json` files, we identified nearly 80 thousand instances of pooling. We find that the misinformation pools are significantly more than twice as likely to pool ad inventory from unrelated websites than those that do not contain a misinformation website. Upon further analysis of ad-related metadata in network traffic, we confirmed the use of 297 pools across 38 ad exchanges by misinformation websites.

Measuring the (in)effectiveness of brand safety tools. We find ads from 55 reputable brands, including Forbes, GoDaddy, Harvard, Intel, Microsoft, Nike, Samsung, Tumblr, Yahoo!, Verizon, and Wayfair, on misinformation websites. We investigate the correlation between the prevalence of pooling and ads from reputable brands on misinformation websites. We find that misinformation websites that are part of at least one dark pool are nearly 20% more likely to attract ads from reputable brands than those that are not part of a dark pool. The responses to our disclosures indicate that reputable brands are generally unaware of their ads appearing on misinformation websites despite several using a brand safety service.

While there is some anecdotal evidence of a general lack of compliance with the ad-tech transparency standards and dark pooling [34], [35], it does not systematically study these issues at scale. To the best of our knowledge, our work is the first to systematically study compliance with ad transparency standards and (dark) pooling at scale.

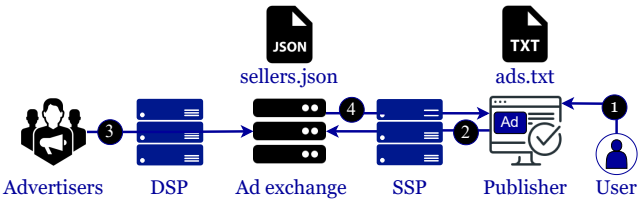


Figure 1: Programmatic advertising ecosystem: When a user visits a publisher website (Step ❶), the publisher puts its ad-inventory for sale on ad exchanges via SSPs in real-time (Step ❷). Advertisers bid for these slots via DSPs (Step ❸). Advertisement of the winning bid is displayed to the user on the publisher website (Step ❹). To mitigate fraud, advertisers use `sellers.json` of ad exchanges and `ads.txt` of publishers to verify who is and who is not an authorized seller of a given inventory.

2. Background

In this section, we provide a high-level overview of the mechanisms behind the supply of programmatic ads (§2.1) and the vulnerabilities in the ad supply chain (§2.2).

2.1. Programmatic advertising

Although there are a variety of mechanisms for programmatic advertising (e.g., real-time bidding, header bidding, exchange bidding) and the participating organizations might differ, the types of entities involved in the supply chain remain the same for each mechanism.

The programmatic advertising supply chain. Programmatic advertising is made possible by the following entities illustrated in Figure 1: *supply-side platforms* (SSPs) for publishers to list their ad inventory in real-time, *ad exchanges* (AdX) which aggregate the inventory of multiple SSPs and facilitate bidding on individual ad slots, and *demand-side platforms* (DSPs) which allow advertisers and brands to identify targets for their ad creatives by suitably bidding on the inventory listed at ad exchanges. These entities work together to create a supply chain for ads as follows: When a user visits a publisher, the ad inventory associated with that visit is put up for auction at an AdX by the SSP. DSPs, operating on behalf of advertisers and brands, then make bids on the ad inventory available at the AdX. These bids are informed by what is known (to the DSP) about the user and the publisher. The winner of the auction is then notified by the AdX and the associated ad creative is used to fill the ad slot on the publisher's website.

Transparency in the supply chain. Crucial to the operation of the ad supply chain is that the participating organizations can trust that publishers and AdXs are not misrepresenting their inventories or their relationships with other entities. For example, DSPs need to confirm that the ad inventory that they are bidding on is actually associated with a particular publisher. Similarly, DSPs also need to confirm that the AdXs that they are purchasing ad inventory from are actually authorized to (re)sell that inventory. The absence of trust in

this supply chain can lead to situations where DSPs place premium bids for ad slots that are actually associated with non-premium publishers — ultimately leading to a brand's ad creative appearing on websites that they may not want to be associated with. To foster trust and enable DSPs (Demand-Side Platforms) to perform basic verification of the ad inventory, the Interactive Advertising Bureau (IAB) introduced two standards: `ads.txt` and `sellers.json`.

The `ads.txt` standard. The `ads.txt`[1] standard (introduced in 2017) aims to address ad inventory fraud by requiring each publisher domain to maintain an `ads.txt` file at the root level directory (e.g., `publisher.example/ads.txt`). The `ads.txt` file is supposed to contain entries for all AdXs that are authorized to sell or resell the ad inventory of the publisher. Each entry in the `ads.txt` file contains the following fields:

- the authorized AdX,
- the publisher ID assigned to the publisher domain within the AdX network, and
- the authorized relationship between the publisher and authorized AdX — i.e., whether the AdX is authorized as a `DIRECT` seller or `RESELLER` of inventory for the domain.

How `ads.txt` helps prevent fraud. When an ad request is sent by a publisher to an AdX (which issues bid requests to DSPs), the request contains the publisher ID and the domain associated with the inventory being listed. Importantly, because publisher IDs are typically associated with an organization and not a domain, it is possible for multiple domains to share the same publisher ID. `ads.txt` enables verification that a website is not spoofing the domain in their ad requests. More specifically, `ads.txt` allows:

- AdXs to verify that the publisher ID in the ad request matches the publisher ID associated with the domain in the ad request and
- DSPs to verify that the AdX claiming to (re)sell the inventory of a domain is authorized by the domain to do so.

Before the `ads.txt` standard, there were no mechanisms to facilitate such checks and the sale of fraudulent inventory was widespread [21].

The `sellers.json` standard. Similar to the `ads.txt` standard, `sellers.json` aims to mitigate ad inventory fraud and misrepresentation. The `sellers.json` standard[2] requires each AdX and SSP to maintain a `sellers.json` file at the root level directory (e.g., `adx.example/sellers.json`).[3] This `sellers.json` file *must* contain an entry for each entity that may be paid for inventory purchased through the AdX — i.e., one entry for each partner that is an inventory source for the AdX. Each entry in the `sellers.json` file contains the following fields:

- the seller type which indicates whether the entry is associated with a `PUBLISHER`, an `INTERMEDIARY` (i.e., inventory reseller AdX), or `BOTH` (i.e., this entity has their own inventory and also resells other inventory);
- the seller ID associated with the inventory source (same as the publisher ID in `ads.txt` if this entry is associated with a publisher. From this point onwards we will refer to seller ID or publisher ID as seller ID); and
- the name and domain associated with the seller ID (these fields may be marked as "confidential" by AdXs to protect the privacy of publishers).

How `sellers.json` helps prevent fraud. When a bid request is received by a DSP from an AdX that is compliant with the `sellers.json` standard, it must contain information about the provenance of the inventory in a Supply Chain Object (SCO).[4] At a high level, the `sellers.json` file provides a mechanism for DSPs to identify and verify all the entities listed in this SCO. This is done as follows:

- When a bid request is received by the DSP, it should use the AdX's `sellers.json` file to verify that the final AdX has an authorized relationship with the prior holder (an SSP or another AdX) of the inventory.
- The previous step is applied recursively (on all intermediate neighbors in the SCO) to verify the end-to-end authenticity of the inventory.
- The DSP then uses the `sellers.json` files of all intermediaries and the `ads.txt` file of the publisher to verify that the publisher is legitimate and (re)sellers who handle the publisher's inventory are authorized to do so.

This capability for end-to-end validation of the SCO (Supply Chain Object) allows DSPs to identify instances where the ad inventory originates from low-quality publishers using fraudulent `ads.txt` files or is being sold by malicious intermediaries.

2.2. Supply chain vulnerabilities

Despite the introduction of the `ads.txt` and `sellers.json` standards, there remain various vulnerabilities in the ad inventory supply chain. Our investigation focuses on the vulnerabilities that enable low-quality publishers to monetize their ad inventory by misrepresenting or obscuring its source. Some of these vulnerabilities arise from misrepresentations in the `ads.txt` and `sellers.json` files, while others arise from pooling their low-quality inventory with the inventory of unrelated high-quality publishers. We refer to the former as *inventory misrepresentation* and the latter as *dark pooling*.

Inventory misrepresentation. Inventory misrepresentation arises from misrepresentations of ad inventory by publishers. It can be identified by discrepancies in the publisher's

1. "ads" in `ads.txt` stands for Authorized Digital Sellers. Full specification of the `ads.txt` standard is available at: https://iabtechlab.com/wp-content/uploads/2021/03/ads.txt-1.0.3.pdf

2. Full specification of the `sellers.json` standard is available at: https://iabtechlab.com/wp-content/uploads/2019/07/Sellers.json_Final.pdf

3. We observed that several AdXs, including Google, use non-standard paths — e.g., Google's `sellers.json` is located at https://storage.googleapis.com/adx-rtb-dictionaries/sellers.json

4. Supply Chain Object (SCO) contains an ordered list of all the entities involved in the ad transaction (e.g., publisher → SSP → reseller → AdX).

ads.txt file and is possible when DSPs and AdXs do not follow the ads.txt and sellers.json standards. Some examples of these misrepresentations include:

- a publisher's ads.txt file might incorrectly use seller IDs of other publishers to suggest an authorized relationship with an AdX to boost the perception of its inventory. (Misrepresentations #1 and #2)
- a publisher's ads.txt file might incorrectly indicate that a popular AdX is an authorized (re)seller of its inventory to boost its reputation with other AdXs. (Misrepresentation #3)
- a publisher's ads.txt file might have more than one entry of the same seller type for an AdX or sellers.json files might associate a seller ID with multiple publishers or sellers making ads.txt and sellers.json verification unreliable. (Misrepresentations #4 and #9)
- a publisher's ads.txt file might list authorized relationships with (re)sellers that do not have sellers.json files, making end-to-end verification impossible. (Misrepresentation #8)

Dark pooling. *Pooling* is a common strategy to share resources in online advertising. Consider, for example, the case where two or more publishers are owned by the same parent organization. In such scenarios, the ability to share advertising infrastructure and AdX accounts allows for more efficient operation and management. One way to identify the occurrence of pooling is by noting a single AdX-issued 'seller ID' shared by multiple publisher websites. *Dark pools* are pools in which seller IDs are shared by organizationally-unrelated publishers (possibly of differing reputation). Note that "dark pooling" is a term of art that is commonly used in industry. While pooling is not itself a "dark" practice, pooling seller IDs of unrelated publishers is considered a "dark" practice because it deceives potential buyers about the actual source of the ad inventory [34], [35].

The seller ID defined in ads.txt and sellers.json standards is also defined in the RTB protocol [36], [37]. Note that the payment after successful completion of an RTB auction is made to the publisher (i.e., the seller) associated with the seller ID [38]. Hence, it should be noted that simply using another domain's seller ID in ad requests from a website will result in any ad-related payments being made to the owner of the seller ID. Therefore, for revenue sharing, the creation of these pools needs to be facilitated either through intermediaries (e.g., SSPs) or by collaboration between publishers.

End-to-end validation of pooled supply chains. Pooling leads to a break down of any brand or DSP's ability to perform end-to-end verification of the ad inventory supply chain. Specifically, the final step of verification highlighted in §2.1 cannot be meaningfully completed unless *all* domains associated with a publisher's account are publicly known (and unfortunately, this is not the case). This is because the end-to-end verification of the ad inventory supply chain, as specified by the IAB, implicitly relies on trust that seller IDs are actually associated with specific organizations and that these associations are verified by AdXs. We illustrate this with an example.

- Consider a publisher website sportsnews.example which has a legitimate subsidiary: nbanews.example. The publisher registers for an account with a popular AdX (adx) and is issued the seller ID sellerid after being vetted by adx. It is expected that this website can now share this seller ID with its subsidiaries. Both websites will now list adx as a DIRECT seller through the sellerid account in their ads.txt files.
- The publisher now decides to share adx-issued seller ID with fakesportsnews.example, another sports news website but of low quality, for a cut of the revenue generated from ads shown on fakesportsnews.example. In its ads.txt file, fakesportsnews.example now adds adx as a DIRECT seller and also lists sellerid as its seller ID. Note that fakesportsnews.example would otherwise be unable to get directly listed on adx and monetize its ad inventory due to its low quality.
- When an ad request for some inventory is sent from fakesportsnews.example, all basic supply chain validation checks are successful because the seller ID sellerid is in fact registered by adx in their sellers.json file. Any bidding DSP will therefore operate under the assumption that the website receiving their ads has been vetted by adx and is associated with sportsnews.example.
- Complications only arise if the verifier notices that sellerid was only registered to the owner of sportsnews.example and the bid request actually originated at fakesportsnews.example. However, invalidating the bid request simply because of this inconsistency will mean that even legitimate subsidiaries such as nbanews.example cannot pool their inventory. Instead, additional checks are required to ascertain whether fakesportsnews.example and sportsnews.example are related or whether adx vetted fakesportsnews.example as well. This issue remains unaddressed by current validation mechanisms.

Caveat. The example described assumes collaboration between publishers — sportsnews.example and fakesportsnews.example. This might be inadvertent in some cases — e.g., if sportsnews.example and fakesportsnews.example are both assigned the same seller ID through a common intermediary (an SSP, for example as shown in Figure 2).

In sum, by pooling various unrelated websites under a single seller ID, low-quality publishers can "launder" their ad inventory, rendering it indistinguishable from the inventory of high-quality publishers. Moreover, this can occur when an AdX provides the seller ID to a trusted publisher (or an SSP), which then inadequately vets the low-quality publishers whose inventory it pools. Figure 2 illustrates this scenario of syndication-based pooling by some intermediary SSP. As we show later, such pooling is common. In fact, we

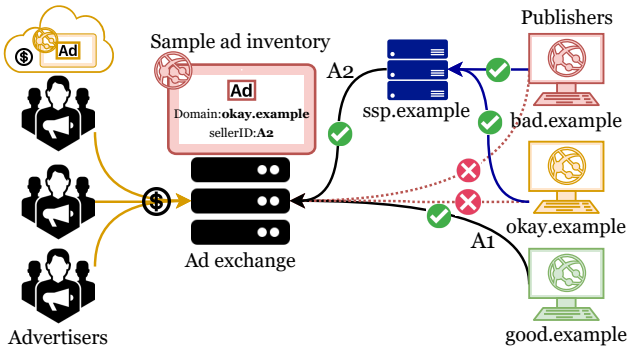


Figure 2: Illustration of pooling by an SSP — A premium publisher (`good.example`) or an AdX-trusted intermediary SSP (`ssp.example`) can list on the AdX to obtain seller IDs `A1` and `A2` respectively. Whereas, a low-quality publisher (`bad.example`) or legitimate but unrecognized publisher (`okay.example`) are unable to directly list on the AdX. A legitimate publisher may not get listed on the AdX because, for instance, traffic requirements are not met. However, `bad.example` and `okay.example` are able to list on SSP, which essentially pools multiple publishers together. Bid request may misrepresent the inventory on `bad.example` as that of `okay.example` using the seller ID of the SSP (i.e., `A2`). Reputable advertisers may bid on the inventory assuming that they are bidding on `okay.example`, when in fact their ad actually would end up on `bad.example`.

find some AdXs even providing services, via intermediaries, that facilitate pooling of unrelated entities.

3. Data

In this section, we describe the selection of publishers that we study (§3.1) and our methodology for collection of `ads.txt`, `sellers.json`, and ad-related metadata associated with these websites (§3.2).

3.1. Publisher website selection

Our goal is to identify practices that hinder the end-to-end validation of the ad inventory supply chain, both among high-quality and low-quality websites. We use misinformation websites as a case study for low-quality websites and use comparably ranked websites from the Tranco list [39] that have `ads.txt` as a stand-in for high-quality websites (referred to as a control).

Selection of misinformation websites. Since identifying misinformation websites is itself not the focus of our work, we leverage lists of misinformation websites curated in prior research by media scholars [40], [41] and computer scientists [42], [43].[5] To construct our list of misinformation websites, we began by aggregating all websites from these lists and removing duplicates. This left us with 1276 websites. Next, we discarded 434 websites that were no longer functional. Finally, we additionally classified each misinformation website using multiple independent sources including Politifact, Snopes, MBFC, OpenSources, PropOrNot, and FakeNewsCodex to ensure that each remaining websites contained content that was undeniably misinformation. We excluded the websites that were now parked domains, seemed to have been repurposed, or had conflicting labels across different sources. This left us with a set of 669 *misinformation websites* ($\mathbf{M}_{\text{full}}$). Of these 669 websites, we created a subset of all the 251 websites that had an `ads.txt` file and were also present in the Tranco top-million list [39] ($\mathbf{M}_{\text{ranked}}$). We use $\mathbf{M}_{\text{ranked}}$ to compare the prevalence of `ads.txt` and `sellers.json` discrepancies between misinformation and non-misinformation websites.

5. For websites obtained from [41], we discard those labeled as 'state', 'political', 'credible', and 'unknown'.

Notation	Description	Size
$\mathbf{M}_{\text{full}}$	Complete set of misinformation domains studied	669
$\mathbf{M}_{\text{ranked}}$	Sites in $\mathbf{M}_{\text{full}}$ with `ads.txt` & part of Tranco-1M	251
$\mathbf{C}_{\text{ranked}}$	Similar-ranked NM with `ads.txt` for each $\mathbf{M}_{\text{ranked}}$	251
$\mathbf{C}_{\text{100K}}$	Tranco Top-100K domains with `ads.txt` presence	20K
$\mathbf{D}^{\text{static}}$	`ads.txt` and `sellers.json` crawled on 02/22	1.4K
$\mathbf{D}^{\text{crawls}}$	(PD, AdX, OD) tuples from dynamic crawl of $\mathbf{M}_{\text{full}}$	2.8K
$\mathbf{D}^{\text{brands}}$	(PD, Brand) pairs from dynamic crawl of $\mathbf{M}_{\text{full}}$	4.2K

TABLE 1: Description of dataset notations and sizes. NM represents non-misinformation websites. PD and OD represent publisher domain and owner domain respectively.

Selection of benign (control) websites. To facilitate comparisons of the prevalence of compliance issues between misinformation and benign websites, we created a control set of non-misinformation websites ($\mathbf{C}_{\text{ranked}}$). For each website in $\mathbf{M}_{\text{ranked}}$, we included the most similarly ranked non-misinformation website that also had an `ads.txt` file. We performed matching based on website domain ranks to avoid confounds related to website popularity. We also created a control set of the Tranco top-100K domains which contained an `ads.txt` file ($\mathbf{C}_{\text{100K}}$). This dataset was used to investigate the broad prevalence of pooling. These four sets of websites ($\mathbf{M}_{\text{full}}$, $\mathbf{M}_{\text{ranked}}$, $\mathbf{C}_{\text{ranked}}$, and $\mathbf{C}_{\text{100K}}$) are the subject of our study.

3.2. Data collection

Our analysis relies on three sources of data: (1) `ads.txt` and `sellers.json` files related to publishers, AdXs, and other intermediaries; (2) bid/ad requests and responses during visits to a publisher domain; and (3) brands placing advertisements on a publisher domain. An overview of our data collection is illustrated in Figure 3. Table 1 lists the notations for different datasets used throughout the paper.

`ads.txt` and `sellers.json` files. To build evidence for the occurrence of pooling and other misrepresentations, we need to analyze published `ads.txt` and `sellers.json` files associated with publishers and ad-tech entities.

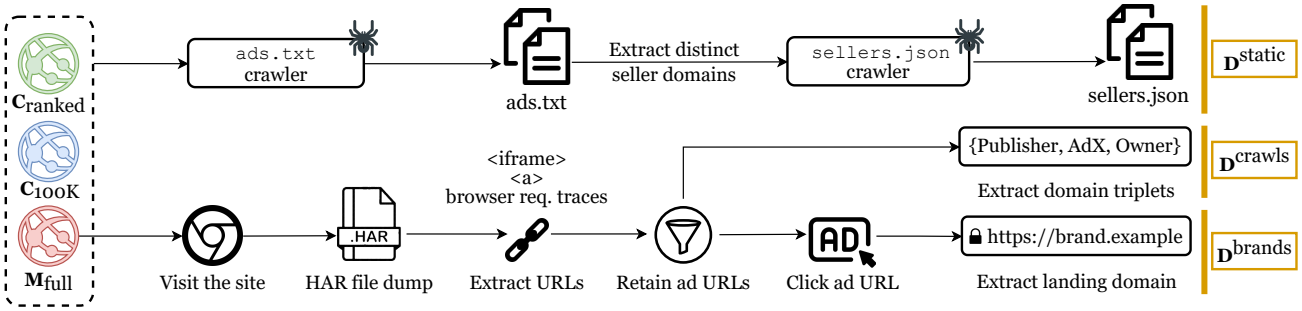


Figure 3: Overview of data collection methodology.

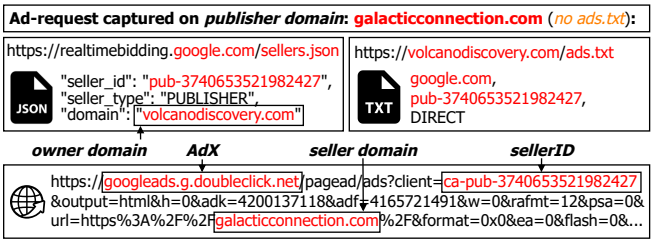
Processing `ads.txt` *files.* We searched for an `ads.txt` file at the root of each website in $\mathbf{M}_{\text{full}}$, $\mathbf{M}_{\text{ranked}}$, $\mathbf{C}_{\text{ranked}}$, and $\mathbf{C}_{100K}$. From these `ads.txt` files, we extracted the domains of all the entities that were listed as `DIRECT` sellers or `RESELLERS` of the publisher's inventory.

Processing `sellers.json` *files.* For each seller identified in our `ads.txt` files, we crawled the `sellers.json` file at the domain's root. When the `sellers.json` file was unavailable at this path, a best-effort attempt was made to manually identify any non-standard location of this file. We manually searched for the `sellers.json` for the top-1K ranked seller domains that were detected as `INTERMEDIARY` or `BOTH` and no `sellers.json` was extracted by the crawler for that domain. We performed a web search using "<domain> – sellers.json" query and looked for the JSON file on the official webpage of the seller. Two `sellers.json` were detected in this manner – google.com and pubmatic.com. We then parsed each `sellers.json` file to identify entities (and their domains) that were associated with `PUBLISHER`, `INTERMEDIARY`, or `BOTH` entries. Finally, until no new entities were discovered, we recursively fetched and parsed the `sellers.json` file associated with the entities labeled as either `INTERMEDIARY` or `BOTH`. This recursive fetching ensures that we have complete coverage of all the supply chain entities that may sell the inventory of all publishers in our datasets.

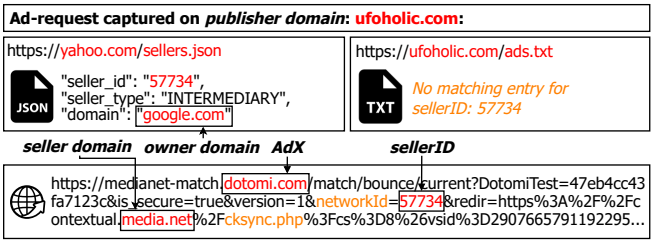
The $\mathbf{D}^{static}$ *datasets.* We crawled and processed `ads.txt` and `sellers.json` files in February 2022. We refer to the dataset as $\mathbf{D}^{\text{static}}$. In total, $\mathbf{D}^{\text{static}}$ included over 98K relationships from `ads.txt` files and 2.4M relationships from `sellers.json` files.

Limitations of this dataset. It should be noted that, by itself, this dataset *cannot present evidence that pooling is actually occurring*. This is because each publisher is responsible only for the content of their own `ads.txt` file, misrepresentation in other publishers' `ads.txt` files is not sufficient to imply pooling.

Obtaining real-time bidding metadata. To identify concrete evidence of pooling, we constructed a dataset of real-time bidding metadata. These include bid requests, responses, redirects, and payloads associated with ad requests and responses. Seller ID is communicated in requests and responses to different entities in the advertising ecosystem. Therefore, during a crawl of a given publisher's website,



(a) True positive case of ID matching an ad request



(b) False positive case of ID matching in an ad request

Figure 4: Illustration of seller ID matching in ad requests for (a) true positive on the misinformation website: galacticconnection.com and (b) false positive on the misinformation website: ufoholic.com.

observing an unrelated entity's seller ID in these metadata constitutes a more concrete evidence of pooling between them.

Crawling configuration. Following the best practices for crawling-based data collection [44], [45], we collected this dataset using a web crawler driven by Selenium (v4.1.0) and the Chrome browser (v91.0) with bot mitigation strategies (multiple randomly timed full page scrolls and randomized mouse movements), Xvfb from a non-cloud vantage point, and a 30-second waiting time after the completion of each page load. Prior work has shown that the bidders and content of ad slots are impacted by previous browsing history [46], [47]. Therefore, each page load was conducted with a new browser profile to avoid biases in our measurements of ad responses and content. With these settings, we loaded each website twice in $\mathbf{M}_{\text{full}}$ and saved the associated HTTP Archive (HAR) files and full-page screenshots.

Extracting real-time bidding metadata from ad-related requests and responses. From each HAR file, we first identified ad-related requests and responses by matching request URLs against well-known advertising filter lists used in

prior research [48]. We extracted the URLs, content, and HTTP POST-encoded data from each ad-related request and response. We then identified all (key, value) pairs using standard delimiters (e.g., & in query parameters). Finally, we matched the identified values with the seller IDs in from the $\mathbf{D}^{\text{static}}$ dataset. To mitigate false positives, we only matched ID strings with length greater than five characters.

Figure 4a shows a sample ad request for doubleclick.net that is matched for the highlighted seller ID. Since volcanodiscovery.com listed in Google's `sellers.json` and the misinformation website galacticconnection.com are unrelated, this represents a true positive instance of dark pooling. For each ad-related request and response, we identified the domain from which the request originated as the *publisher domain* (i.e., observed inventory source) and the AdX domain owning the detected seller ID as the *AdX* (i.e., inventory seller). We then used the `sellers.json` of the AdX/inventory seller to identify the domain that owned the seller ID found in the ad request. This domain is labeled as the *owner domain* (i.e., expected inventory source). The (*publisher domain*, *AdX*, *owner domain*) triples are used in later analysis. Figure 4b also shows a sample ad request on ufoholic.com, where one of the values matches with a Yahoo-issued seller ID that is owned by Google. However, Google-associated domains are absent from the ad request. The seller ID also does not exist in ufoholic.com's `ads.txt`. This match is deemed a false positive match and discarded from further analysis.

Methodology validation. We manually evaluated the accuracy of our method to extract metadata from ad-related requests. Specifically, we manually examined the requests and responses to verify that they did in fact include a key that suggested that the $value$ was associated with a seller ID. Our manual evaluation gave a false positive rate of 1.5%.

The $\mathbf{D}^{crawls}$ *dataset.* We label this dataset of (*publisher domain*, *AdX*, *owner domain*) triples as $\mathbf{D}^{\text{crawls}}$. In total, the $\mathbf{D}^{\text{crawls}}$ dataset consisted of 3.1K distinct triples observed across two crawls of 669 $\mathbf{M}_{\text{full}}$ websites. In §4, we use these triples to determine (dark) pooling on misinformation websites.

Limitations of this dataset. The programmatic advertising is auction-driven and participation from entities is non-deterministic. Therefore, any observations of entities and the IDs in requests and responses related to ads will vary from one crawl to the next, even when all other client-related factors are identical. Further, the browser provides a vantage point that typically only affords observations of the winners of real-time bidding auctions. Finally, it is possible that some communications regarding the involved seller ID are not visible to us due to hashing or other forms of obfuscation [49]. These limitations are unavoidable. It should be noted, however, that these limitations only impact the completeness of our findings and not the correctness. In other words, the prevalence of pooling and other discrepancies, as measured by our crawls, are only a lower-bound for their actual prevalence.

Identifying brands in advertisements. We also analyzed the brands whose ads appear on misinformation websites. To identify brands advertising on misinformation websites, we performed 10 separate crawls. This repetition was to account for the non-deterministic nature of programmatic advertising that results in a user receiving different ads on repeat visits to the same website. In each of the 10 crawls, after each page load was complete and the 30-second wait period ended, we clicked the DOM elements associated with each ad-related URL on the page. These clicks typically resulted in navigation to the brand's website. We used this website's domain to label the brand associated with the ad.

Methodology validation. To test the effectiveness of this methodology, we conducted a pilot test on one crawl where we compared the brand names identified through manual analysis and the automated approach. We found that in 30% of the displayed ads, the automated approach failed to identify the brand associated with an ad. In these cases, failure was largely because some ad-related request URLs were associated with"unclickable" elements of the ad. As a result, our automated approach could not trigger navigation to the brand's website. To mitigate this issue, we supplemented our automated approach by manually annotating the ads on all crawls that could not be associated with a brand. This process was relatively quick since most of the ads had been already automatically annotated with associated brands.

The $\mathbf{D}^{brands}$ *dataset.* We recorded all (*publisher, brand*) pairs identified with this methodology in $\mathbf{D}^{\text{brands}}$ dataset. In total, the $\mathbf{D}^{\text{brands}}$ dataset consisted of 4.2K distinct (publisher, brand) pairs and 2.1K unique brands.

Crawl success rate. Our crawling infrastructure for `ads.txt` and `sellers.json` had a 100% success rate (i.e., if a website had a file, we were able to crawl it without any failures). Dynamic web crawls did fail for a small percentage of websites ($< 5\%$) due to timeouts. However, we were able to crawl all websites at least once since we performed multiple crawls for each website.

Limitations. We performed all crawls from one IP address, which could impact our analysis of brands. In other words, we might have observed more or less brands had we performed crawling from multiple IP addresses.

Ethical considerations. We discuss the ethics of our web crawling along three dimensions: infrastructure costs, privacy risks, and advertising costs caused by this study. Overall, our study respects the principle of beneficence as outlined in the Menlo Report [50] and Belmont Report [51] by maximizing the possible benefits and minimizing the harms.

Infrastructure costs. Our crawls were used to measure the prevalence of compliance issues and misrepresentations. Our two dynamic crawls were not conducted concurrently to avoid stressing the web servers. Similarly, our additional static crawls for `ads.txt` and `sellers.json` were performed six months apart. While our crawlers did not follow the `robots.txt` directives (if present) on misinformation publishers, our crawling methodology is in line with ethical

and legal considerations of such crawling-based auditing systems [52]–[54]. Also note that our study did not involve human subjects or gather any personal information.

Advertising costs. To actually understand what brands are advertising on misinformation websites and what ad-exchange is responsible for showing that ad, we clicked on the ads shown during the page loads. The costs associated with our ad clicks are negligible (CPMs are in the order of fractions of cents and we clicked a total of 4247 ads). We believe these costs are justifiable given the benefit of understanding vulnerabilities in the ad-tech ecosystem.

4. Measuring Problematic Representations

In this section, we answer the question: *what is the prevalence of pooling and other problematic representations on misinformation websites?* Specifically, we focus on measuring the prevalence of misrepresentations that hinder end-to-end supply chain validation. In §4.1, we provide a broad overview of the types of misrepresentations commonly seen in `sellers.json` and `ads.txt` files. We compare the prevalence of these misrepresentations on control and misinformation websites. In §4.2, we present evidence of ad inventory pooling and highlight cases of dark pooling by misinformation websites.

4.1. Prevalence of misrepresentations

Certain types of misrepresentations in a publisher's `ads.txt` file or an AdX's `sellers.json` file may prohibit automated end-to-end verification of the ad inventory supply chain. We identify eight such problematic representations:

1) *Misrepresented direct relationships:* The Publisher claims that an AdX is a `DIRECT` seller of its inventory, but the AdX's `sellers.json` lists it as an `INTERMEDIARY` (reseller) relationship;
2) *Misrepresented reseller relationships:* The Publisher claims that an AdX account is a `RESELLER` of its inventory, but the AdX's `sellers.json` associates the corresponding account as a `PUBLISHER` (direct) entry;
3) *Fabricated seller IDs:* A publisher's `ads.txt` claims that an AdX is authorized to sell its inventory via some seller ID, but the AdX's `sellers.json` does not have *any* account associated with that specific ID;
4) *Conflicting relationships:* A publisher claims the same type of relationship(s) with more than one seller ID on a given AdX in their `ads.txt`, but the AdX only lists one of these relationships in their `sellers.json`;
5) *Invalid seller type:* The `sellers.json` does not use any of the three acceptable types (`PUBLISHER`, `INTERMEDIARY`, or `BOTH`) to describe the source of the inventory associated with a specific seller ID;

Index	Type	$\mathbf{C}_{\text{ranked}}$	$\mathbf{M}_{\text{ranked}}$
1	Misrepresented direct relationships	51%	64%
2	Misrepresented reseller relationships	47%	65%
3	Fabricated seller IDs	65%	83%
4	Conflicting relationships	33%	49%

TABLE 2: Prevalence of problematic representations in `ads.txt` from websites in $\mathbf{C}_{\text{ranked}}$ and $\mathbf{M}_{\text{ranked}}$.

Index	Type	No $\mathbf{M}_{\text{full}}$	$\geq$ 1 $\mathbf{M}_{\text{full}}$
5	Invalid seller type	0.7%	0%
6	Invalid domain names	0.8%	54.8%
7	Confidential sellers	0.1%	46.1%
8	Intermediaries w/o `sellers.json`	13.3%	49.8%
9	Non-unique seller IDs	62.6%	95.3%

TABLE 3: Fraction of `sellers.json` entries that contain different problematic representations from AdXs serving no $\mathbf{M}_{\text{full}}$ websites and at least one $\mathbf{M}_{\text{full}}$ website.

6) *Invalid domain names:* The `sellers.json` does not present a valid domain name[6] in the 'domain' field;
7) *Confidential sellers:* The `sellers.json` lists the domain associated with the seller ID as 'confidential'. It should be noted that this is not a violation of the `sellers.json` standard, but does prevent end-to-end supply chain verification because both the 'domain' and 'name' fields are redacted;
8) *Intermediaries without `sellers.json`:* An AdX's `sellers.json` lists intermediaries that do not have a `sellers.json`; and
9) *Non-unique seller IDs:* The `sellers.json` associates multiple publisher or seller domains with the same seller ID confounding the buyer's verification.

Table 2 compares the prevalence of misrepresentations in `ads.txt` files of $\mathbf{C}_{\text{ranked}}$ and $\mathbf{M}_{\text{ranked}}$ websites. We find a statistically significant difference in the number of errors present in `ads.txt` files from $\mathbf{C}_{\text{ranked}}$ and $\mathbf{M}_{\text{ranked}}$ websites (χ^2-test; $p < .05$). We find that misinformation websites are more likely to contain higher rates of `ads.txt` misrepresentations that result in failed supply chain validation, even when controlling for website rank. Table 3 compares the prevalence of misrepresentations in `sellers.json` of AdXs that serve $\mathbf{M}_{\text{full}}$ (344 AdXs) websites with the `sellers.json` from AdXs that do not serve any of our $\mathbf{M}_{\text{full}}$ websites (483 AdXs). Again, we see that the AdXs that engage with misinformation websites are significantly more likely to have misrepresentations in their `sellers.json` that result in the inability to perform supply chain validation. Taken together, our results highlight the lack of compliance with `ads.txt` and `sellers.json` standards and their current inability to allow end-to-end supply chain validation. This problem is especially pronounced for the ad inventory of misinformation publishers.

6. While a buyer may still rely on the 'name' field, it is not suitable for automated analysis because 'name' is a free text field. Automated analysis is crucial as bid requests need to be programmatically validated in real-time and at scale.

4.2. Prevalence of pooling

As described in §2.2, pooling is the practice of using a single AdX account to manage the inventory of multiple websites. This results in a single AdX-issued seller ID being associated with multiple websites. Although this practice enables more efficient management of advertising resources for publishers, it comes at the cost of increased opacity in the advertising ecosystem and reduces the effectiveness of the end-to-end supply chain validation mechanisms.

Gathering evidence of pooling with the $\mathbf{D}^{\text{static}}$ dataset. We begin by identifying evidence of pooling in the $\mathbf{C}_{100K}$ and $\mathbf{M}_{\text{full}}$ websites from our $\mathbf{D}^{\text{static}}$ dataset. We use this dataset of `ads.txt` files associated with the Tranco top-100K domains to identify all cases where multiple domains listed the same seller ID and AdX as a seller of their inventory. In total, we observed 79K unique pools — i.e., 79K unique (seller ID, AdX) pairs were observed to have been shared by multiple publisher domains. Of these 79K pools, 11% (8.7K) also included at least one of the misinformation websites in $\mathbf{M}_{\text{full}}$. We refer to these 79K pools identified through the $\mathbf{D}^{\text{static}}$ dataset as *static pools*. The size of these pools ranged from 2 to nearly 9K domains, with an average of 70 domains per pool.

Characteristics of pools identified in the $\mathbf{D}^{\text{static}}$ dataset. These above-reported pool sizes were certainly larger than what we anticipated and necessitated additional inspection for a better understanding of our findings. Specifically, we paid attention to the organizational relationships between pooled entities and whether pooling was occurring due to some ad-tech related mechanism.

Organizational homogeneity of pools. From a cursory manual inspection of our pools, we observed (rather unsurprisingly) that larger pools appeared to contain many organizationally unrelated domains — i.e., they were *heterogeneous*. To measure the prevalence of such types of pools at scale, we mapped each domain in a pool to their parent organization using the DuckDuckGo entity list [55] and labeled each pool as follows:

1) *Homogeneous:* Pools whose member domains could all be mapped to a single parent organization;
2) *Potentially homogeneous:* Pools for which the parent organizations of all domains could not be identified. However, all domains that could be mapped were found to have the same parent organization;
3) *Heterogeneous:* Pools whose member domains were owned by more than one parent organization; and
4) *Unknown:* Pools for which no domain could be mapped to a single parent organization.

Figure 5 illustrates how pools are categorized into homogeneous and heterogeneous. Table 4 provides a breakdown of the prevalence of different types of pools. We make three key observations. First, we notice that *heterogeneous pools comprise a large fraction of all pools* — a deviation from the expectation that pools are allowed in order to facilitate resource sharing between sibling domains. The high incidence rates of heterogeneous pools in non-misinformation

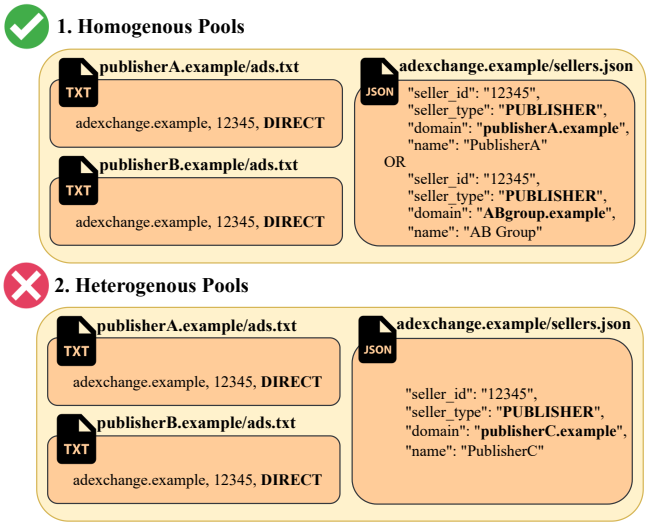


Figure 5: Categorization of pools based on the relation of the publishers with the domain owner organization. In the homogeneous pool, *PublisherA* and *PublisherB* authorize seller account 12345 on *adexchange.example* as their direct seller. The `sellers.json` of *adexchange.example* recognizing 12345 as an account owned by either *PublisherA*, *PublisherB*, or *AB Group* represent all valid cases of pooling assuming *PublisherA* and *PublisherB* are related (in this case owned or operated by *AB Group*). If the `sellers.json` of *adexchange.example* shows that the seller account 12345 is owned by *PublisherC* and *PublisherA* or *PublisherB* are unrelated to *PublisherC*, then this represents a case of heterogeneous pool, which we consider a dark pool.

websites also *suggests* that there may be legitimate (i.e., not ill-intentioned) mechanisms that facilitate seller ID sharing between organizations. Second, *pools containing misinformation websites are statistically significantly more likely to be heterogeneous* (85%) than pools without misinformation websites (41%) [χ^2-test; $p < .05$]. Finally, we see that *pools containing misinformation websites are statistically significantly larger* (412.1 websites/pool) than pools without misinformation websites (20.3 websites/pool) [2-sample t-test: $p < .05$; u-test: $p < .05$]. Taken together, the latter two findings lend credence to the thesis that misinformation websites are effectively "laundering" their ad inventory by participating in mechanisms that facilitate large heterogeneous pools.

Pools facilitated by authorized ad-tech mechanisms. Our findings about the high rate of heterogeneous pools of large sizes, even among non-misinformation websites, suggest that there are ad-tech mechanisms that organically facilitate pooling. After further investigation we found that many of the heavily pooled (seller ID, AdX) pairs appeared to be issued by a small number of AdXs whose `sellers.json` file indicated that the issued seller IDs were not associated with specific publishers but instead other ad platforms (AdXs or SSPs). In other words, the seller ID issuing AdX's `sellers.json` file indicated that the 'owner domain' of

Pool Type	Pools w/ M_{full}		Pools w/o M_{full}	
	# Pools	μ_{size}	# Pools	μ_{size}
Homogeneous	40 (0.4%)	2.6	6.7K (9.6%)	2.6
Po. Homogeneous	913 (9.1%)	18.8	18.4K (26.6%)	7.0
Heterogeneous	8.6K (85.0%)	482.5	28.4K (41.0%)	42.2
Unknown	563 (5.6%)	4.3	15.7K (22.7%)	3.9
All pools	8.7K	412.1	70.5K	20.3

TABLE 4: **Prevalence of pools from D^{static} in C_{100K}.** Pools are broken down by organization homogeneity and whether they contained a misinformation website from the M_{full} dataset. μ_{size} denotes the average (mean) number of websites in a pool.

the pooled seller ID was another AdX/SSP — suggesting that these pooling mechanisms might be authorized by the AdX platforms themselves for aggregating and reselling ad inventory of different publishers. Table 5 shows that three of the most commonly pooled owner domains belong to large AdXs (`google.com`, `justpremium.com` owned by GumGum, and `townnews.com`). Most notably, nearly 25% and 12% of the pools that used GumGum- and Google-owned seller IDs also contained known misinformation websites. For example, `100percentfedup.com`, a website that promoted anti-vax and stolen-election theories, received ads through pools using Google-owned seller IDs issued by the AdX 'Index Exchange'. In contrast, TownNews, an advertising firm focused on serving local media organizations did not have a single pool containing known misinformation websites.

To investigate the prevalence of pooling, we looked for AdX-sanctioned programs that might require pooling — i.e., is there public documentation of *authorized* programs to allow unrelated publishers to pool their inventory through intermediaries. Notably, we found public documentation of Google's Multiple Customer Management (MCM) program that allows 'Google MCM-partner' organizations to manage the inventory of multiple client publishers through a single account [56]. As a result, all the publishers that are managed by an MCM partner are served ads via the same seller ID of the intermediary MCM organization. Our results show that misinformation websites are able to monetize their ad inventory by being part of these MCM networks. *Our results highlight a violation of Google's own policies regarding advertising on websites 'making unreliable claims' or 'distributing manipulated media'* [57]. However, public documentation does not clearly state whether Google delegates all website and content verification responsibilities to their MCM partners and therefore it remains unclear if the violation is a failure of Google's own verification practices or those of their MCM partners. Similarly, the pooled misinformation websites using GumGum-owned seller IDs were also in violation of GumGum's content policy [58].

Pools using seller IDs with hidden or unknown owner domains. During our investigation, we also discovered that many AdX's `sellers.json` files did not allow identification of the owner domain of the seller ID that was used. This comprised nearly half of all identified pools. The breakdown

Type	Domain	Pools	Pools w/ M_{full}
Owner of sellerID	google.com	5.1K	598
	gannett.com	370	5
	justpremium.com	337	84
	townnews.com	313	0
	hearst.com	219	1
AdX issuer of sellerID	google.com	10.3K	461
	taboola.com	6.6K	132
	freewheel.com	3.9K	625
	pubmine.com	3.6K	2
	openx.com	2.4K	524

TABLE 5: **Most pooled domains and AdXs from D^{static}.** The top five rows represent the most frequently observed domains whose seller IDs were used in pools. The bottom five rows represent the most frequently observed AdXs who issued the seller IDs that were used for pooling.

of reasons for this is provided in Table 6. Here, we see that the most common reasons for failed identification of the owners of seller IDs being used in pooling are: (1) the seller ID is itself unlisted in the issuing AdX's `sellers.json` file and (2) the unavailability of a public `sellers.json` from the owner domain that owned the AdX-issued seller ID (when owner domain is not a PUBLISHER type entry). It is important to note that any of the reasons shown in Table 6 would result in the impossibility of any end-to-end supply chain verification. Interestingly, we find no statistical differences (χ^2-test; $p < .05$) between the reasons for failed identification of owners of non-misinformation and misinformation pools. This suggests that the issues of poor compliance with end-to-end supply chain verification procedures are industry-wide and no specific cause for these failures is exploited by misinformation websites.

Reason	All pools	Pools w/ M_{full}
Total pools	79K	8.7K
seller ID unlisted	20.9K	2.5K
`sellers.json` not public	16.5K	2.0K
Owner not listed	2.6K	135
Owner is *confidential*	3.4K	86

TABLE 6: **Pools from D^{static} using IDs of unknown owners.** Reasons for failed identification of the owners of seller IDs used in pools.

Finding occurrences of pooling with the D^{crawls} dataset. Because of the high rates of misrepresentations, unreliability of publisher-sourced `ads.txt` files, and the incompleteness of AdX-sourced `sellers.json` files, it is important to note that our analysis of the D^{static} can only be used as evidence that suggests the widespread practice of *potential* dark pooling. In order to confirm a dark pool's existence with certainty we need to observe it in a live page load. To this end, we leverage the set of all (publisher domain, AdX, owner domain) triples recorded in our D^{crawls} dataset (cf. §3.2). Since these were obtained from actual ad-related metadata from crawls of the M_{full} dataset, they provide concrete evidence of pooling actually being leveraged by known misinformation websites (i.e., dark pooling). In total,

we gathered 2.8K (publisher domain, AdX, owner domain) triplets through two crawls of $\mathbf{M}_{\text{full}}$ websites from which we identified 297 pools across 38 ad exchanges. These 297 pools are depicted in Figure 6 Of these, 218 pools (73.4%) overlapped with those identified in our analysis of the $\mathbf{D}^{\text{static}}$ dataset and 79 were new. The non-existence of 79 pools in the $\mathbf{D}^{\text{static}}$ dataset prevented us from classifying them and this once again highlights the ad industry's poor compliance with `ads.txt` and `sellers.json` standards.

Google and PubMatic were found to be the issuers of the seller IDs associated with 120 and 48 pools, respectively. These pools enabled advertising supply chains for 127 (Google) and 67 (PubMatic) misinformation websites. 33Across and Gourmet Ads were found to be the owners of seller IDs that were shared by the most number of misinformation websites (28 and 23 websites, respectively). Both seller IDs were issued by PubMatic. Other notable AdXs (and count of the number of seller IDs issued by them which were pooled by misinformation websites) include Rubicon Project (now Magnite) (34), ContextWeb (now PulsePoint) (30), Amazon (28), and media.net (25).

Homogeneity of $\mathbf{D}^{crawls}$ *pools.* From 297 distinct pools, we were able to identify the presence of 15 homogeneous and 203 heterogeneous pools. The homogeneity of the remaining pools could not be determined. The largest homogeneous pool shared a seller ID issued to `funkedigital.de` by PubMatic. This pool included nine websites such as `principia-scientific.org`, `allnewspipeline.com`, `russia-insider.com` — Media Bias/Fact Check identified all the nine websites as 'Conspiracy Theory' or 'Propaganda' related with 'Low' factual reporting and having 'Right' to 'Extreme-Right' bias. We identified stories related to climate change denial, vaccination misinformation, and pro-insurrection views — all in violation of PubMatic's own content guidelines for publishers [59]. Incidentally, a seller ID on Pubmatic was also associated with the largest heterogeneous pool with 47 unique misinformation websites, including `drudgereport.com` and `worldtruth.tv`. Unfortunately, PubMatic's `sellers.json` file did not list the seller ID associated with this heterogeneous pool, suggesting that it was employing fabricated or unlisted ID for pooling.

$\mathbf{D}^{crawls}$ *pools and the Google MCM program.* In order to identify occurrences of pooling in Google's MCM program, we identified pools associated with the seller IDs issued by Google to MCM partners. Of the 203 unique heterogeneous pools identified, a vast majority were labeled as confidential in Google's `sellers.json` [60] but we were able to link 15 to Google's MCM program based on public documentation. In total, Google's MCM partners were associated with 27 misinformation websites. Some of these MCM partners whose Google-issued seller IDs were pooled by misinformation websites include Adnimation, Ezoic, etc. Misinformation websites supported by Google's MCM program included `369news.net` (pseudoscience or anti-vaxx theories) and `truthandaction.org` (extreme-right propaganda and/or misinformation), amongst other

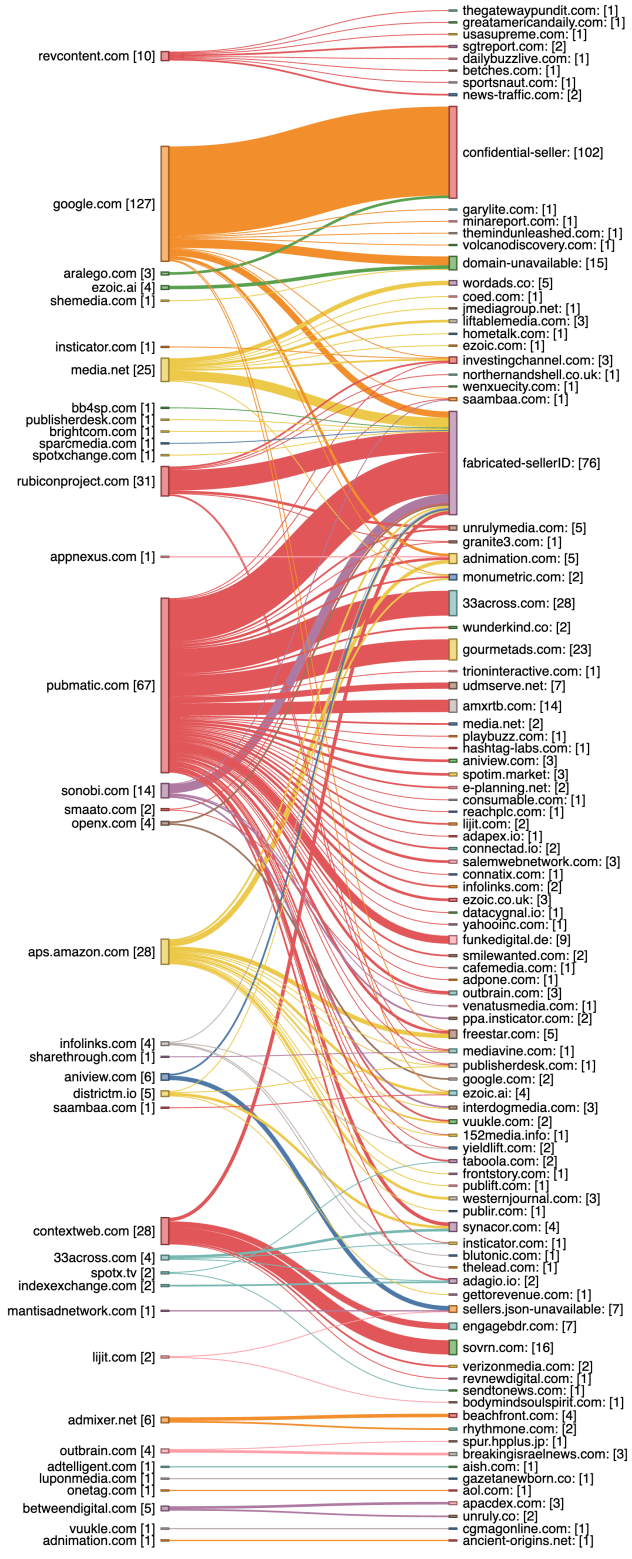


Figure 6: Dark pooling relationships between AdXs (left) and owner domains of AdX-issued pooled IDs (right) for 297 unique pools observed during the crawls of $\mathbf{M}_{\text{full}}$ websites. The counts represent the number of distinct misinformation websites pooled.

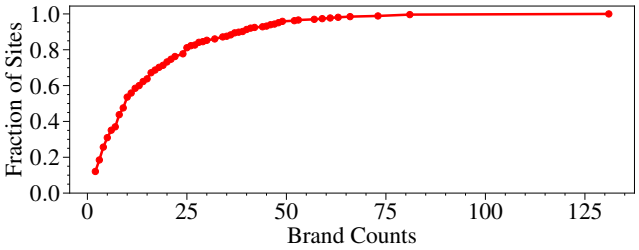


Figure 7: Cumulative distribution of the number of distinct brands across different misinformation websites

similar websites. The MCM partners most frequently found to be using their Google-issued seller ID for pools containing misinformation websites were Monumetric (5 pools) and Freestar (4 pools).

Takeaways. Our analysis shows a widespread failure to adhere to the `ads.txt` and `sellers.json` standards and the compliance is even more weaker amongst misinformation websites (§4.1). This poor adherence has one major consequence: end-to-end validation of the ad-inventory supply chain is not always possible, particularly in the case of misinformation websites. Further compounding supply chain validation challenges, we find that the pooling of seller IDs by unrelated publishers is also widespread (§4.2). Misinformation websites, which violate the publisher content policies of many AdXs, are able to monetize their ad inventory through these pools. In fact, we find that in many cases they are able to leverage the authorized programs of the same AdXs whose policies they violate.

5. Brand Analysis

In this section, we analyze the display ads loaded on misinformation websites to identify the advertisers/brands that end up buying their ad inventory.

Data collection. We curate $\mathbf{D}^{\text{brands}}$ by crawling each of the 669 misinformation websites ten times as discussed in §3.2. We are able to collect a total of 4,246 ads belonging to 2,068 distinct brands. Figure 7 plots the distribution of the number of distinct brands across misinformation websites. We find that a non-trivial fraction of misinformation websites are able to get ads from tens of distinct brands. Specifically, 23 misinformation websites have ads from at least 41 distinct brands each while 142 misinformation websites have ads from at most 10 distinct brands each.

Reputable brand classification and prevalence. To assess whether these ads are from reputable brands, we use their Tranco ranks as a rough proxy for their reputation. Specifically, we classify brands with top-1K Tranco ranking as "reputable". Figure 8 shows the number of distinct reputable and non-reputable brands across top-20 misinformation websites that contain ads from the highest distinct brands. Perhaps surprisingly, we find that Breitbart – a well-known misinformation website – is able to attract ads from the highest number of distinct brands. The two reputable brands with ads on Breitbart include Forbes and GoDaddy.

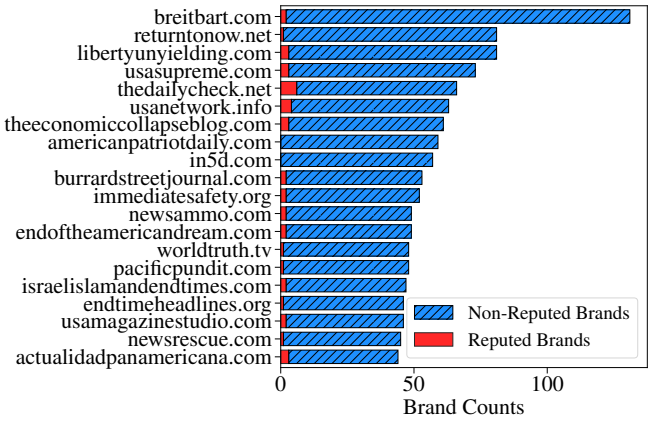


Figure 8: Distribution of reputable and non-reputable brands among the top-20 misinformation websites with the highest number of distinct brands advertising on their website.

In total, we observe ads from 55 reputable brands including Forbes, GoDaddy, Harvard, Intel, Microsoft, Nike, Samsung, Tumblr, Yahoo!, Verizon, and Wayfair. We note that these top-20 misinformation websites tend to have more ads from reputable brands on average as compared to the remaining misinformation websites. Specifically, the average number of reputable brands is 2.05 for the top-20 misinformation websites in Figure 8 and 0.78 for the remaining misinformation websites.

Correlation between ad inventory misrepresentation and number of brands. Next, we investigate whether the misrepresentation of ad inventory by misinformation websites impacts their ability to sell their ad inventory.

Figure 9 plots the distribution of the distinct brand counts of all the brands advertising on misinformation websites with/without `ads.txt`. Note that we are looking for the existence of `ads.txt`.[7] We find that misinformation websites with `ads.txt` are able to attract ads from twice as many brands on an average as compared to the websites without `ads.txt`. We conclude that some brands do avoid advertising on misinformation websites without `ads.txt`.

Figure 10 plots the conditional probabilities of observing reputable brands across misinformation websites with/without dark pools. We find that more than half of the misinformation websites part of one or more dark pools get ads from reputable brands. In contrast, less than one-third of the misinformation websites part of no dark pools get ads from reputable brands. This nearly 20% difference in the conditional probability shows that dark pooling significantly increases the chances of ads from reputable brands ending up on misinformation websites.

Brand disclosures. It is reasonable to assume that reputable brands generally do not want to advertise on misinformation websites [29], [30], [32]. Taking the example of Breitbart, there is ample evidence that reputable brands did not want

7. The mere existence of `ads.txt` does not guarantee the veracity of its content. A misinformative publisher could have misrepresented `ads.txt` entries to bypass brand checks. Advertisers are recommended by IAB to perform `ads.txt` checks against the data observed in the bid requests before making a bid.

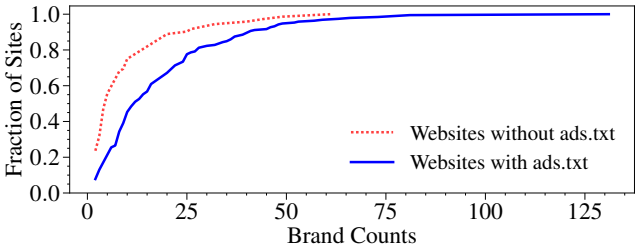


Figure 9: Cumulative distribution of the number of distinct brands on misinformation websites with or without `ads.txt`

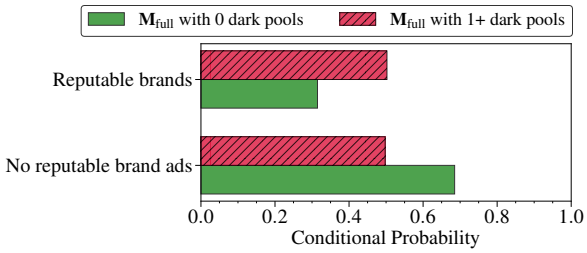


Figure 10: Conditional probabilities of ads from reputable brands in presence or absence of dark pooling

their ads shown on Breitbart [32], [61], [62]. DSPs and AdXs typically provide brand safety features [63] to help brands avoid buying the ad inventory of low-quality websites. Brand safety features allow brands to block unwanted ad inventory through a block list of domains or seller IDs [60]. One would expect that reputable brands would attempt to avoid buying the ad inventory of misinformation websites through these brand safety features. Since brand safety is not externally measurable, we conduct individualized disclosures to these 55 reputable brands and specifically ask them (a) whether they want their ads on misinformation websites or not and (b) whether they employ brand safety features to this end.

To perform disclosures, we first attempted to find advertising-related email addresses for each reputable brand from their website. If we were unsuccessful, we included generic email addresses from their "About Us" and "Contact Us" pages. In our disclosures, we listed the misinformation websites where the ads of the reputable brand were observed. We included full-page screenshots showing the brand's ad creative on the misinformation website as well as the full HTTP Archive (HAR) recording of the network traffic. We also asked them whether they were aware of or intended to have their ads on the misinformation websites and whether/which brand safety service they used.

We received responses from 11 reputable brands. 8 brands confirmed that they were unaware and did not intend to advertise on these misinformation websites. For example, one brand responded that "*We don't advertise on the site. It was an unintentional oversight related to automated advertising and the ad was immediately pulled when discovered. We always aim to advertise on sites that are aligned with our mission and values and we apologize if this upset any of our customers.*" Another brand mentioned that "*We will not want to see our ads on misinformation websites.*" Regarding the deployment of brand safety features, we received confirmation from 4 brands that they indeed used a brand safety service but it did not adequately detect or prevent their ad from appearing on the misinformation website. One brand told us that it used Google Display Network's built-in brand-safety measure while two brands employed the brand safety service provided by Integral Ad Science (IAS). One brand told us that "*the misinformation website disclosed by you [...] is neither present in the logs provided to us by our DSP partner nor is flagged by IAS. We think that it is being misplaced on the misinformation website due to dark pooling.*" Another brand told us that "*About the brand-safety service, please understand we are not able to tell any detail*" presumably due to a confidentiality agreement.

Takeaways. In summary, our results show that the misrepresentation of ad inventory by misinformation websites seems to be correlated with their ability to monetize their ad inventory through reputable brands. We found that the ad inventory of misinformation websites that use dark pooling is more likely to be bought by reputable brands. The limited responses from reputable brands suggest that they do not want to advertise on misinformation websites and employ brand safety features to this end.

6. Related Work

Examining the online advertising ecosystem. In recent years, there have been many research efforts to bring transparency to the mechanisms of online advertising. A large number of these have focused on studying personal data collection and sharing to deliver personalized ads [64]–[69]. Our work instead focuses on the prevalence of inventory fraud, pooling, and its impact on brands.

Inventory fraud. There have been a few measurements related to `ads.txt` standard and related inventory fraud since its introduction. However, no work has focused on `sellers.json` or the ad-fraud that emerges by the combined failure of `ads.txt` and `sellers.json`. In 2019, Bashir et al. [21] gathered and conducted a longitudinal analysis of `ads.txt` files. They found that these files were riddled with syntactic errors and inconsistencies that made them difficult to process in an automated fashion. Tingleff [70] and Pastor et al. [71] highlighted flaws of the `ads.txt` standard that undermines its effectiveness in preventing ad fraud, albeit without measurements to support their hypotheses. Some of these identified flaws are, however, supported by measurements from Papadogiannakis et al. [72]. These findings, suggesting that the `ads.txt` standard is not effectively enforced, are corroborated by our study. Our work complements these efforts by undertaking a measurement study of both the standards of `ads.txt` and `sellers.json` for the first time to measure inventory fraud as well as prevalence of pooling, which allows low-quality publishers to launder their ad inventory.

Brand safety. There have been many studies that have highlighted the impact of ads (and the websites on which they appear) on the reputation of a brand [28], [63], [73], [74]. In fact, several activist efforts have successfully leveraged brand safety concerns to demonetize misinformation. Notable among these are the efforts of Check My Ads and Sleeping Giants [75], who successfully used public campaigns to pressurize 820 brands to add Breitbart News' domain to their advertising block lists. Our cataloging of brands found on known misinformation websites can supplement these ad-hoc efforts and increase pressure on ad-tech to enforce its own `ads.txt` and `sellers.json` standards more effectively. Other work has focused on measuring or improving the effectiveness of mechanisms for identifying 'brand safe' web content. Most recently, Vo et al. [76] built an image-based brand-safety classifier to prevent ad placement on inappropriate pages. Numerous products from major ad-tech firms such as DoubleVerify [77], Integral Ad Science [78], and Oracle [79] have also recently started promoting their 'brand safety' features.

Funding infrastructure of misinformation. Ours is not the first work to consider the role of the online advertising ecosystem in funding misinformation. In fact, it has been known for several years that online advertising provides the primary revenue stream for misinformation websites [80]–[84]. Han et al. [85], in their study, focused on network infrastructure, also explored the revenue streams on misinformation websites and identified disproportionately high reliance on advertising and consumer donations.

Bozarth et al. [86] showed that although there is a unique ecosystem of 'risky' AdXs that partner with publishers of misinformation, there is also a heavy presence of mainstream AdXs (e.g., Google) in the misinformation ecosystem.

Braun & Eklund [87] take a qualitative approach to understand the role of the advertising ecosystem in increasing revenues of misinformation and the dismantling of traditional journalism. Their work, along with numerous others [88]–[90], has highlighted the need for additional transparency to realize the promise of market-based strategies to curb funding of misinformation.

Considering another angle, several studies have also examined how deceptive ads are used to promote and fund harmful products [91]–[93] and ideologies [43], [94], [95].

At a high-level, our work complements all these efforts to better understand how the misinformation ecosystem is funded by online advertising by uncovering and analyzing the exploitation of specific advertising-related vulnerabilities such as pooling and relationship misrepresentations by the misinformation ecosystem.

7. Concluding Remarks

Our work shows how the opacity of the ad-tech supply chain is exploited by misinformation publishers to monetize their ad inventory. Through our measurements, we demonstrate a widespread lack of compliance with the IAB's `ads.txt` and `sellers.json` standards, ad inventory pooling by misinformation publishers, and reputed brands who end up buying this ad inventory of misinformation publishers. Taken all together, our results point to specific gaps that need to be further explored by the ad-tech and security research communities.

Trust delegation in advertising partner programs. One of our key findings is that a small number of ad exchanges are responsible for a majority of dark pooling. In many cases, we see evidence that this dark pooling is achieved through the use of legitimate partner programs made available by ad exchanges (e.g., Google's MCM partner program [56]). These programs serve an important purpose — to help reduce the management burdens on small publishers. However, as we see in our study, this expanded access facilitated via advertising partners results in new vulnerabilities. Specifically, publishers who are in clear violation of the policies set by an ad exchange are still able to obtain seller IDs issued by the exchange through their partners. One perspective of this problem is that there is a fundamental breakdown of trust delegation — i.e., partners are delegated the rights to assign and manage seller IDs on behalf of exchanges, but without being properly delegated the responsibilities for vetting publishers and verifying their compliance with ad exchange policies. While this work is the first to uncover this delegation of trust in the form of verification responsibilities in the ad-tech ecosystem, it is not new to the security community. Indeed, this type of trust delegation is a central theme in the public key infrastructure [96], app stores [97], and other domains. From these prior efforts to delegate verification responsibilities, it is clear that success is only possible with effective mechanisms to monitor compliance and revoke delegated trust. A key difference from prior efforts, however, is that it is not publicly known how these trust delegation processes work within specific ad exchanges. Without public documentation of these processes or research studies that uncover them, we anticipate that identifying weaknesses and causes for failure will remain an open challenge.

Supply chain transparency and compliance with industry standards. The programmatic advertising supply chain is complex because of the large number of entities involved between publishers and advertisers. Further complicating matters, our study shows that these entities are frequently out of compliance with even basic standards such as `ads.txt` and `sellers.json`. In fact, many of the concerning findings of our work could be addressed if advertisers were able to trace the provenance of ad inventory using the existing 'Supply Chain Object' (SCO) ad-tech standard. Unfortunately, our analysis of SCOs in §B shows that less than a quarter of bid requests actually include the SCO. Further, even when the SCO is included in bid requests, they are often incomplete and missing information would make end-to-end verification of the supply chain difficult. We further find that even major ad exchanges implement digital advertising standards in a way that hinders external independent audits. Notably, Google's widespread

use of confidential `sellers.json` entries [60] makes it challenging to identify Google AdX's partners who are not doing adequate compliance verification for the publishers whose inventory they list. IAB has recently released new and updated digital advertising standards [98], [99] to improve end-to-end validation of the ad-tech supply chain. However, these are not widely adopted yet. Therefore, in its current state, to mitigate ad fraud and reduce ad-tech's inadvertent funding of misinformation, it is crucial that adoption and compliance with new and existing digital advertising standards such as SCO, `ads.txt`, and `sellers.json` improve. However, a key challenge is the absence of incentives for achieving compliance with these standards. It remains to be seen if recent US regulatory efforts will improve compliance. Notably, the Digital Services Oversight and Safety Act (DSOSA) [100] and Advertising Middlemen Endangering Rigorous Internet Competition Accountability Act (AMERICA) [101] introduce new requirements related to online advertising transparency. In addition, we are currently engaged in conversations with members of US Congress seeking to draft additional legislation specifically to strengthen compliance with ad-tech industry standards, improve transparency around the ad-tech supply chain, and mitigate ad fraud.

Effective notification and vulnerability reporting mechanisms. The ad-tech industry is currently lacking mechanisms through which supply chain vulnerabilities may be reported. This absence has resulted in several community-organized efforts such as the Check My Ads Institute [31] that monitor ads on misinformation websites and use social media to report on the brands or ad-tech loopholes that fund these publishers. While these efforts have been successful at mitigating some of the harms from the opacity of the supply chain, they are not systematic reports and rely on amplification via social media in order to reach their intended targets. Further, like our study, they generally focus on specific harms caused by the opacity of ad-tech (e.g., funding of misinformation). There is a need to develop more generalized and systematic mechanisms for reporting supply chain vulnerabilities and non-compliance with existing industry standards.

For reproducibility and to foster follow-up research, our dataset is available at https://osf.io/hxfkw/?view_only=bda006ebbd7d4ec2be869cbb198c6bd5

Acknowledgment

This work is supported in part by the National Science Foundation under grant numbers 2103439, 2103038, and 2138139. We want to thank the anonymous shepherd and reviewers for their constructive feedback that helped improve the work.

References

[1] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the seventh international workshop on data mining for online advertising*, 2013.

[2] OpenRTB Guidelines. IAB: https://www.iab.com/guidelines/openrtb/, 2022.

[3] Display ad-tech Lumascape. https://lumapartners.com/content/lumascapes/display-ad-tech-lumascape/.

[4] What's the Difference Between Waterfall Auctions & Header Bidding? https://clearcode.cc/blog/difference-waterfall-header-bidding/, 2022.

[5] Sumayah A. Alrwais, Alexandre Gerber, Christopher W. Dunn, Oliver Spatscheck, Minaxi Gupta, and Eric Osterweil. Dissecting ghost clicks: Ad fraud via misdirected human clicks. In *Proceedings of the 28th Annual Computer Security Applications Conference (ACSAC)*, 2012.

[6] Brett Stone-Gross, Ryan Stevens, Apostolis Zarras, Richard Kemmerer, Chris Kruegel, and Giovanni Vigna. Understanding fraudulent activities in online ad exchanges. In *Proceedings of the ACM Internet Measurement Conference*, 2011.

[7] Jonathan Crussell, Ryan Stevens, and Hao Chen. Madfraud: Investigating ad fraud in android applications. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, 2014.

[8] Vacha Dave, Saikat Guha, and Yin Zhang. Measuring and fingerprinting click-spam in ad networks. In *Proceedings of the ACM SIGCOMM Conference*, 2012.

[9] Apostolis Zarras, Alexandros Kapravelos, Gianluca Stringhini, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proceedings of the ACM Conference on Internet Measurement Conference*, 2014.

[10] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Shady paths: Leveraging surfing crowds to detect malicious web pages. In *Proceedings of the ACM SIGSAC Conference on Computer & Communications Security*, 2013.

[11] Kurt Thomas, Elie Bursztein, Chris Grier, Grant Ho, Nav Jagpal, Alexandros Kapravelos, Damon Mccoy, Antonio Nappa, Vern Paxson, Paul Pearce, Niels Provos, and Moheeb Abu Rajab. Ad injection at scale: Assessing deceptive advertisement modifications. In *IEEE Symposium on Security and Privacy*, 2015.

[12] Shishir Nagaraja and Ryan Shah. Clicktok: Click fraud detection using traffic analysis. In *Proceedings of the Conference on Security and Privacy in Wireless and Mobile Networks*, 2019.

[13] Suibin Sun, Le Yu, Xiaokuan Zhang, Minhui Xue, Ren Zhou, Haojin Zhu, Shuang Hao, and Xiaodong Lin. Understanding and detecting mobile ad fraud through the lens of invalid traffic. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

[14] Mobin Javed, Cormac Herley, Marcus Peinado, and Vern Paxson. Measurement and analysis of traffic exchange services. In *Proceedings of the ACM Internet Measurement Conference*, 2015.

[15] Shehroze Farooqi, Guillaume Jourjon, Muhammad Ikram, Mohamed Ali Kaafar, Emiliano De Cristofaro, Zubair Shafiq, Arik Friedman, and Fareed Zaffar. Characterizing key stakeholders in an online black-hat marketplace. In *APWG Symposium on Electronic Crime Research (eCrime)*, 2017.

[16] Manolis Chalkiadakis, Alexandros Kornilakis, Panagiotis Papadopoulos, Evangelos Markatos, and Nicolas Kourtellis. The rise and fall of fake news sites: A traffic analysis. In *ACM Web Science Conference*, 2021.

[17] Inside the Macedonian Fake News Complex. https://www.wired.com/2017/02/veles-macedonia-fake-news/, 2022.

[18] How teens in the balkans are duping trump supporters with fake news. https://www.buzzfeednews.com/article/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo#.fu2okXaeKo.

[19] How Facebook powers money machines for obscure political news sites. https://www.theguardian.com/technology/2016/aug/24/facebook-clickbait-political-news-sites-us-election-trump, 2016.

[20] IAB Brand-safety. https://www.iab.com/topics/brand-safety/, 2022.

[21] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. A longitudinal analysis of the ads. txt standard. In *ACM Internet Measurement Conference*, 2019.

[22] What is domain spoofing? https://www.adpushup.com/blog/what-is-domain-spoofing/, 2017.

[23] Domain spoofing remains a huge threat to programmatic. https://digiday.com/marketing/domain-spoofing-remains-an-ad-fraud-problem/, 2017.

[24] The Sentencing of The King Of Fraud and the Birth of Collective Protection. https://www.humansecurity.com/learn/blog/the-sentencing-of-the-king-of-fraud-and-the-birth-of-collective-protection, 2021.

[25] The four types of domain spoofing. https://integralads.com/insider/the-four-types-of-domain-spoofing/, 2018.

[26] IAB ads.txt Specifications. https://iabtechlab.com/ads-txt/, 2017.

[27] IAB sellers.json Specifications. https://iabtechlab.com/sellers-json/, 2019.

[28] Chunsik Lee, Junga Kim, and Joon Soo Lim. Spillover effects of brand safety violations in social media. *Journal of Current Issues & Research in Advertising*, 2021.

[29] New ias report uncovers how consumer perception of misleading content impacts brand favorability. https://integralads.com/news/misinformation-consumer-research/, 2022.

[30] New ias report uncovers how misleading content impacts digital advertising. https://integralads.com/news/new-ias-report-uncovers-how-misleading-content-impacts-digital-advertising/, 2022.

[31] Check My Ads Institute. https://checkmyads.org/, 2022.

[32] List of advertisers that demonetized breitbart. https://twitter.com/slpng_giants/status/1200473586886205440, 2020.

[33] Cvs blocks breitbart. https://mobile.twitter.com/slpng_giants/status/808381433173577730, 2016.

[34] Further investigation into the "dark pool sales house" phenomenon. https://deepsee.io/blog/non-unique-pub-ids, 2021.

[35] Murky ad-tech tactics: What you should know about dark pool sales houses. https://www.adweek.com/media/dark-pool-sales-houses-what-you-need-to-know/, 2021.

[36] Google openrtb integration. https://developers.google.com/authorized-buyers/rtb/openrtb-guide.

[37] Vungle openrtb integration. https://support.vungle.com/hc/en-us/articles/360045953431-Vungle-Exchange-OpenRTB-2-5-Integration-Guide#3-2-2-source-ext-schain-nodes-object-0-12, 2022.

[38] Doubleclick for publishers. https://www.adpushup.com/blog/google-dfp-doubleclick-for-publishers/, 2021.

[39] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco list. https://tranco-list.eu. Accessed on 27th Oct, 2021.

[40] Maciej Szpakowski and Renato. Fake news corpus. https://github.com/several27/FakeNewsCorpus, Feb 2018.

[41] Media Bias/Fact Check Team. Media bias/fact check: The most comprehensive media bias resource. https://mediabiasfactcheck.com.

[42] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. Identifying disinformation websites using infrastructure features. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2020.

[43] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. Bad news: Clickbait and deceptive ads on news and misinformation websites. In *Workshop on Technology and Consumer Protection (ConPro)*, 2020.

[44] Syed Suleman Ahmad, Muhammad Daniyal Dar, Zareed Zaffar, Narseo Vallina-Rodriguez, Rishab Nithyanand, et al. Apophanies or epiphanies: How crawlers can impact our understanding of the web. In *The Web Conference*, 2020.

[45] Jordan Jueckstock, Shaown Sarker, Peter Snyder, Aidan Beggs, Panagiotis Papadopoulos, Matteo Varvello, Benjamin Livshits, and Alexandros Kapravelos. Towards realistic and reproducibleweb crawl measurements. In *Proceedings of the Web Conference*, 2021.

[46] John Cook, Rishab Nithyanand, and Zubair Shafiq. Inferring tracker-advertiser relationships in the online advertising ecosystem using header bidding. *Proceedings on Privacy Enhancing Technologies*, 1, 2020.

[47] Maaz Bin Musa and Rishab Nithyanand. Atom: Ad-network tomography. a generalizable technique for inferring tracker-advertiser data sharing in the online behavioral advertising ecosystem. *Proceedings on Privacy Enhancing Technologies*, 2022.

[48] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. Adgraph: A graph-based approach to ad and tracker blocking. In *IEEE Symposium on Security and Privacy*, 2020.

[49] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In *Proceedings of the 2017 Internet Measurement Conference*, 2017.

[50] Menlo report. https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf, 2012.

[51] Belmont report. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html, 1979.

[52] Robots.txt meant for search engines don't work well for web archives. https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives, 2017.

[53] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014), 2014.

[54] Madison Addicks. Van buren v. united states: The supreme court's ruling on the fate of web scraping-" access" to discovery or detention? *Tul. J. Tech. & Intell. Prop.*, 24, 2022.

[55] Github — duckduckgo tracker-radar/entities. https://github.com/duckduckgo/tracker-radar/tree/main/entities.

[56] Overview of MCM - Google AdMob Help. https://support.google.com/admob/answer/9142605?hl=en.

[57] Misrepresentation - Advertising Policies Help. Google Advertising Policies, https://support.google.com/adspolicy/answer/6020955?hl=en&ref_topic=1626336, 2022.

[58] GumGum — Prohibited Content Policy for Buyers and Sellers. https://gumgum.com/terms-and-policies/buyer-policy, 2022.

[59] Supply policy. Pubmatic, 2022.

[60] Porn, piracy, fraud: What lurks inside google's black box ad empire. https://www.propublica.org/article/google-display-ads-piracy-porn-fraud, 2022.

[61] Sunrun blocks breitbart. https://twitter.com/slpng_giants/status/1024019694523695111, 2018.

[62] Sagenamerica blocks breitbart. https://twitter.com/slpng_giants/status/913821166304763904, 2017.

[63] Steven Bellman, Ziad HS Abdelmoety, Jamie Murphy, Shruthi Arismendez, and Duane Varan. Brand safety: the effects of controversial video content on pre-roll advertising. *Heliyon*, 4(12), 2018.

[64] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016.

[65] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, Phillipa Gill, et al. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *The 25th Annual Network and Distributed System Security Symposium (NDSS 2018)*, 2018.

[66] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. Tracing information flows between ad exchanges using retargeted ads. In *USENIX Security Symposium*, 2016.

[67] Aaron Cahn, Scott Alfeld, Paul Barford, and Shanmugavelayutham Muthukrishnan. An empirical study of web cookies. In *Proceedings of the 25th international conference on world wide web*, 2016.

[68] Yash Vekaria, Vibhor Agarwal, Pushkal Agarwal, Sangeeta Mahapatra, Sakthi Balan Muthiah, Nishanth Sastry, and Nicolas Kourtellis. Differential tracking across topical webpages of indian news media. In *ACM Web Science Conference*, 2021.

[69] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *International Workshop on Traffic Monitoring and Analysis*. Springer, 2014.

[70] S Tingleff. The Three Deadly Sins of Ads. Txt and How Publishers Can Avoid Them. https://iabtechlab.com/blog/the-three-deadly-sins-of-ads-txt-and-how-publishers-can-avoid-them, 2019.

[71] Antonio Pastor, Rubén Cuevas, Ángel Cuevas, and Arturo Azcorra. Establishing trust in online advertising with signed transactions. *IEEE Access*, 9, 2020.

[72] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. Who funds misinformation? a systematic analysis of the ad-related profit routines of fake news sites. In *Proceedings of the ACM Web Conference 2023*, pages 2765–2776, 2023.

[73] Edlira Shehu, Nadia Abou Nabout, and Michel Clement. The risk of programmatic advertising: Effects of website quality on advertising effectiveness. *International Journal of Research in Marketing*, 38(3), 2021.

[74] Sophie Bishop. Influencer management tools: Algorithmic cultures, brand safety, and bias. *Social Media+ Society*, 7(1), 2021.

[75] Claudia Pereira Ferraz. Sleeping giants and indirect boycotts against the far right in united states of america. *Aurora.*, 14(40), 2021.

[76] Quan Minh Vo, Nhan Thi Cao, and An Hoa Ton-That. Unsafe image classification using convolutional neural network for brand safety. In *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020.

[77] Doubleverify unveils expanded brand safety & brand suitability integration with facebook - doubleverify. https://doubleverify.com/newsroom/doubleverify-unveils-expanded-brand-safety-brand-suitability-integration-with-facebook/.

[78] Integral ad science — brand safety & suitability solutions. https://integralads.com/solutions/brand-safety-suitability/.

[79] Oracle moat measurement — oracle advertising. https://www.oracle.com/cx/advertising/measurement/.

[80] Nir Kshetri and Jeffrey Voas. The economics of "fake news". *IT Professional*, 19(6), 2017.

[81] Arvind Hickman. Advertisers spend $2.6bn on misinformation websites, study finds. https://www.campaignlive.com/article/advertisers-spend-26bn-misinformation-websites-study-finds/1725293, Aug 2021.

[82] Global Disinformation Index. Cutting the funding of disinformation: The ad-tech solution. https://disinformationindex.org/wp-content/uploads/2019/05/GDI_Report_Screen_AW2.pdf, May 2019.

[83] Global Disinformation Index. The quarter billion dollar question: How is disinformation gaming ad tech? https://disinformationindex.org/wp-content/uploads/2019/09/GDI_Ad-tech_Report_Screen_AW16.pdf, Sep 2019.

[84] Augustine Fou. Big advertisers still fund hate and disinformation outside of facebook. https://www.forbes.com/sites/augustinefou/2020/07/06/big-advertisers-still-fund-hate-and-disinformation-outside-of-facebook/?sh=383aa3376f78, July 2020.

[85] Catherine Han, Deepak Kumar, and Zakir Durumeric. On the infrastructure providers that support misinformation websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2022.

[86] Lia Bozarth and Ceren Budak. Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. In *The International AAAI Conference on Web and Social Media*, 2020.

[87] Joshua A Braun and Jessica L Eklund. Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism. *Digital Journalism*, 7(1), 2019.

[88] Joel Timmer. Fighting falsity: Fake news, Facebook, and the first amendment. *Cardozo Arts & Ent. LJ*, 35, 2016.

[89] Damian Tambini. Fake news: public policy responses. 2017.

[90] Norman Vasu, Benjamin Ang, Terri-Anne Teo, Shashi Jayakumar, Muhammad Raizal, and Juhi Ahuja. *Fake news: National security in the post-truth era*. S. Rajaratnam School of International Studies., 2018.

[91] Yelena Mejova and Kyriaki Kalimeri. Advertisers jump on coronavirus bandwagon: Politics, news, and business. *arXiv preprint arXiv:2003.00923*, 2020.

[92] Amelia M Jamison, David A Broniatowski, Mark Dredze, Zach Wood-Doughty, DureAden Khan, and Sandra Crouse Quinn. Vaccine-related advertising in the facebook ad archive. *Vaccine*, 38(3), 2020.

[93] Vanessa Boudewyns, Brian G Southwell, Kevin R Betts, Catherine Slota Gupta, Ryan S Paquin, Amie C O'Donoghue, and Natasha Vazquez. Two awareness of misinformation in health-related advertising: A narrative review of the literature. *Misinformation and mass audiences*, 2021.

[94] Eric Zeng, Miranda Wei, Theo Gregersen, Tadayoshi Kohno, and Franziska Roesner. Polls, clickbait, and commemorative $2 bills: problematic political advertising on news and media websites around the 2020 US elections. In *Proceedings of the 21st ACM Internet Measurement Conference*, 2021.

[95] Yingying Chen and Luping Wang. Misleading political advertising fuels incivility online: A social network analysis of 2020 US presidential election campaign video comments on YouTube. *Computers in Human Behavior*, 2022.

[96] Laurent Chuat, AbdelRahman Abdou, Ralf Sasse, Christoph Sprenger, David Basin, and Adrian Perrig. Sok: Delegation and revocation, the missing links in the web's chain of trust. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2020.

[97] Fuqi Lin, Haoyu Wang, Liu Wang, and Xuanzhe Liu. A longitudinal study of removed apps in iOS app store. In *Proceedings of the Web Conference*, 2021.

[98] Ads.cert 2.0. https://iabtechlab.com/ads-cert/, 2022.

[99] ads.txt Version 1.1. https://iabtechlab.com/wp-content/uploads/2022/04/Ads.txt-1.1.pdf, 2022.

[100] H.R.6796 - Digital Services Oversight and Safety Act of 2022. https://www.congress.gov/bill/117th-congress/house-bill/6796/text, 2022.

[101] S.1073 - Advertising Middlemen Endangering Rigorous Internet Competition Accountability Act or the AMERICA Act. https://www.congress.gov/bill/118th-congress/senate-bill/1073/text, 2023.

Appendix A.
Longitudinal Analysis of `sellers.json`

Various campaigns have highlighted the role of AdXs in monetizing the misinformation ecosystem, pressuring them to remove their support for these domains [31]. To understand the effectiveness of these campaigns, we monitored changes to the `sellers.json` files present in our $\mathbf{D}^{\text{static}}$ dataset for a three-month period (from Oct'21 to Feb'22). Of the 470 AdXs found to support misinformation websites (by listing them as publishers) on October 2021, 39 (8.3%) AdXs delisted at least one misinformation website by February 2022.

Bashir et. al. [21] performed this analysis on `ads.txt` of Alexa Top-100K websites in their work. However, our study is on misinformation websites, whose `ads.txt` should not be trusted. Hence, we perform this analysis on `sellers.json` files of trusted AdXs.

We observed 470 `sellers.json` supporting at least one misinformation website as per October's crawl – 46 of which support 10 or more misinformation outlets. The ones that support the highest misinformation websites are *revcontent.com* (204), *liveintent.com* (56), *outbrain.com* (56), *pixfuture.com* (39), and *lijit.com* (now part of *Sovrn*) (30). From Oct'21 to Feb'22, only 39 AdXs de-list at least 1 misinformation website, while 53 `sellers.json` include at least 1 misinformation website in their files. Table 7 shows the top AdXs and their longitudinal support for the misinformation websites in their `sellers.json`.

Ad exchange	Misinformation Website Counts			
	Oct'21	Feb'22	Added	Dropped
revcontent.com	204	73	2	133
outbrain.com	56	35	0	21
9mediaonline.com	20	1	0	20
stitchvideo.tv	14	1	0	13
adtelligent.com	26	28	13	11
infolinks.com	23	14	2	11
publisherdesk.com	14	3	0	11
mgid.com	20	32	13	1
nextmillennium.io	7	9	3	1
vidazoo.com	5	8	3	0
pixfuture.com	39	41	2	0
lijit.com	30	30	0	0

TABLE 7: AdXs that add and drop the most misinformation websites from their `sellers.json` between Oct'21 and Feb'22. The table is arranged in descending order of the dropped counts.

Upon further investigation of RevContent, we observed that it dropped ∼87% of the total publisher domains from their `sellers.json` in mid-December 2021 (Oct'21: 4727 domains to Feb'22: 621 domains) and we speculate that their primary aim might not have been to drop misinformation websites, but they ended up de-listing a few of misinformation websites too as a result of their bulk drop. There has always been a constant peer-pressure and criticism from activists (e.g., [31]) forcing RevContent to remove their support for misinformation websites. There were active discussions on social media speculating RevContent's intent behind this massive drop. However, RevContent did this silently and never publicly addressed this action. Even after the drop, RevContent still potentially funds the most misinformation websites in our data. Other than RevContent, other AdXs that continued their support for the highest misinformation websites in Feb'22 are *LiveIntent* (56), *Pixfuture* (41), *Outbrain* (35), and *MGID* (32).

Additionally, the misinformation outlets which were added by the most AdXs are *rearfront.com*, *vidmax.com*, and *thetruereporter.com*. The former 2 outlets are agents of spreading viral and misleading content, while the latter publishes politicized news, commentary and analysis. These were added by 6 different AdXs. Similarly, *lifezette.com*, *waynedupree.com*, and *news18.co* were dropped by 6, 6, and 5 AdXs respectively.

Appendix B.
Supply Chain Object Analysis

If adopted and implemented correctly, Supply Chain Objects (SCOs) can aid overall validation and provide transparency into all the entities involved in (re-)selling of a particular ad-inventory. In absence of SCOs, a buyer has visibility into only the immediate upstream seller but not the entire path of (re-)sellers that were involved before the upstream seller. It is the job of each seller to append its seller object in the existing SCO and forward the bid request further. A buyer extracts the SCO object from the bid request and parses the list of all seller nodes represented by key `nodes`. Higher the index of a node in this list, the more recent the seller. When an AdX forwards the bid request for a publisher, it associates the publisher with dictionary key `asi` and the account identifier for that publisher in its network with the key `sid`.

In order to analyze the adoption and correctness of SCOs in our data, we use our custom SCO parser (based on the IAB guidelines) on all the bid requests captured in the $\mathbf{D}^{\text{crawls}}$ dataset. Despite SCOs being introduced by IAB since July 2019, only 20.5% (3796) bid requests have included SCOs, all of which comprised only a single seller node. To verify the correctness of SCOs, we extracted `sid` and `asi` associated with the seller node and performed `sid` lookup in the `sellers.json` file of the `asi` to obtain the upstream seller domain with which the ad-inventory is associated as per the SCO. Next, we checked if this website domain matched the actual website's domain on which the current bid request was captured during the dynamic crawl. Let's call this boolean result – A. We also validated all 3796 SCO-based paths (upstream website → `asi` seller → ad-request domain) against 3-hop static paths involving each misinformation website generated from the `sellers.json` files. Let's call this boolean result – B. The cases where A and B were True are cases where we could verify the correctness of the SCOs. The rest cases were SCO misrepresentations. We observed only 18.94% (719) of 3796 bid requests with correct implementation of SCOs.

Appendix C.
Meta-Review

C.1. Summary

This paper investigates how dark pooling of ad inventory leads to misrepresentation of the true property on which ads are displayed, bypassing brand safety measures used by major ad exchanges. The authors perform a large-scale measurement study to identify and characterize how this dark pooling leads to misinformation sites selling ads to major brands.

C.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research
- Provides a New Data Set For Public Use
- Establishes a New Research Direction
- Provides a Valuable Step Forward in an Established Field

C.3. Reasons for Acceptance

1) This paper provides independent confirmation of important results with limited prior research. While the issue of ad inventory misrepresentation was known, its prevalence was not well-characterized. The work's large-scale measurement demonstrates the real-world impact of this issue.
2) The paper provides a new data set for public use and establishes a future research direction. Measurement study data is being shared in a format that will enable future analysis of ad inventory misrepresentation. New research in ad inventory transparency, including several suggested in the paper, can leverage this data.
3) The paper provides a valuable step forward in an established field. The paper provides a systematic review of the interaction between ad inventory transparency and the real-world behavior of advertisers and website operators.

C.4. Noteworthy Concerns

1) The vulnerabilities discussed in the paper are not novel. It is already understood that ad standards are not effectively enforced and that dark pooling can undermine the validation of ad supply chains.
2) While this paper performed a detailed study on ad fraud, the measurement methodology is straightforward and not technically challenging.
3) Because detected advertisers are a lower-bound, some correlations presented in the paper may not hold if this undercounting is not uniform.
4) Because crawling was done from a single IP address, IP-based tracking could have biased measured advertisements.