

“Recommended For You”: A First Look at Content Recommendation Networks

Muhammad Ahmad Bashir
Northeastern University
Boston, MA
ahmad@ccs.neu.edu

Sajjad Arshad
Northeastern University
Boston, MA
arshad@ccs.neu.edu

Christo Wilson
Northeastern University
Boston, MA
cbw@ccs.neu.edu

ABSTRACT

One advertising format that has grown significantly in recent years are known as *Content Recommendation Networks* (CRNs). CRNs are responsible for the widgets full of links that appear under headlines like “Recommended For You” and “Things You Might Like”. Although CRNs have become quite popular with publishers, users complain about the low-quality of content promoted by CRNs, while regulators in the US and Europe have faulted CRNs for failing to label sponsored links as advertisements.

In this study, we present a first look at five of the largest CRNs, including their footprint on the web, how their recommendations are labeled, and who their advertisers are. Our findings reveal that CRNs still fail to prominently disclose the paid nature of their sponsored content. This suggests that additional intervention is necessary to promote accepted best-practices in the nascent CRN marketplace, and ultimately protect online users.

1. INTRODUCTION

As the online advertising ecosystem has grown and matured, publishers and advertisers have experimented with many different advertising formats. Although display ads (*a.k.a.* banner ads) and keyword ads (*a.k.a.* search ads) are perhaps the most well known ad formats, advertisers continue to experiment with new formats like native, social, and video ads (including 6 second mini-videos known as “bumpers” [11]).

One advertising format that has grown significantly in recent years are known as *Content Recommendation*

Networks (CRNs). CRNs are responsible for the widgets full of links that appear under headlines like “Recommended For You” and “Things You Might Like”. These widgets typically provide links to two types of content: 1) recommended content from the first-party publisher (i.e., the owner of the website that has embedded the widget), and 2) sponsored content from third-parties. Advertisers pay when users click on the sponsored links, and the revenue is split between the CRN and the publisher.

The growth of CRNs has been fueled by the compelling services they offer to publishers. First, CRNs provide ready-made recommendation engines that help publishers promote their own content to visitors. Second, publishers earn revenue when users click on third-party sponsored content. To put things in perspective: the largest CRNs, Outbrain and Taboola, claim to serve billions of recommendations per month to between 400–550M unique users, and earned \$240M and \$200M in revenue, respectively, in 2014 [9, 12].

However, CRNs have come under fire from users and regulators. Users have vocally complained about the “click-bait” headlines that are often promoted by CRNs, or by deceptive “bait-and-switch” sponsored links that appear to be content, but instead lead to product advertisements [9]. Some advertisers have even used CRNs to openly advertise scams (e.g., get rich quick schemes) [18]. Furthermore, regulators in the US and Europe have faulted CRNs for failing to prominently label sponsored links as advertisements, and encouraged them to change their business practices [7, 20].

In this study, we present a first look at CRNs, including their footprint on the web, how their recommendations are labeled, and who their advertisers are. We focus on five of the largest CRNs: Outbrain, Taboola, Revcontent, Gravity, and ZergNet. To study these services, we crawled CRN widgets from 500 websites in 2016, including 289 top publishers (e.g., CNN and USA Today) and 211 random sites from the Alexa Top-1M.

In total, we collected 130,996 ads and 53,202 recommendations from five CRNs, which form the basis of our study. Using this dataset, we make the following key observations:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IMC 2016, November 14 - 16, 2016, Santa Monica, CA, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4526-2/16/11...\$15.00

DOI: <http://dx.doi.org/10.1145/2987443.2987469>



Figure 1: Sponsored links with thumbnail images from Revcontent.

- CRNs serve 2.5× more advertisements than recommendations per page, on average. Furthermore, 12% of CRN widgets mix sponsored and non-sponsored links, which may confuse users.
- 11% of CRN widgets that contain ads do not have a headline, and of those that do, only 15% include words like “sponsored”, “promoted”, or “ad”. Furthermore, although 94% of CRN widgets include a nominal disclosure, the substantive quality of these disclosures varies greatly across CRNs.
- Advertiser quality varies widely across CRNs. Almost 60% of Gravity’s advertisers are in the Alexa Top-10K, while 40% of Revcontent’s advertisers registered their domain name <1 year ago.
- The most frequently advertised topics on CRNs include dubious financial services and salacious celebrity gossip. This confirms many of the content quality criticisms leveled against CRNs in the press [9, 17, 18].

Although CRNs have been criticized in the past for the quality of their content, and for failing to prominently disclose the paid nature of their sponsored content [7, 20], our findings reveal that these problems are still widespread. This suggests that additional intervention is necessary from industry trade groups and regulators to promote accepted best-practices in the nascent CRN marketplace, and ultimately protect online users.

Open Source. We have open sourced all the data from this project, which is available at:

<http://personalization.ccs.neu.edu/>

2. BACKGROUND AND RELATED WORK

We begin by setting the stage for our study. First, we discuss the broader online advertising ecosystem and highlight related work. Second, we introduce Content Recommendation Networks (CRNs).

2.1 The Online Ad Ecosystem

The online display and keyword ad ecosystem has been well-studied by researchers. Rodriguez et al. used anonymized data from a major European mobile carrier to study the mobile advertising space [21]. Similarly, Barford et al. used crawled data to map out the online *adscape* and characterize the footprint of major ad



Figure 2: Text-only recommendations and sponsored links from Outbrain.

networks [2]. Others have used controlled methodologies to examine the prevalence of behaviorally targeted ads, and the features that drive targeting [6, 10]. A long line of work has focused specifically on documenting the proliferation of tracking mechanisms used to fuel the ad industry [13–15, 19]. Finally, recent studies have shown that tracking data and revenue are highly skewed towards the largest ad networks [5, 8].

2.2 Content Recommendation Networks

CRNs, also known as *Content Discovery Networks*, first appeared in 2006 with the founding of Outbrain, closely followed in 2007 by Taboola. Today, Outbrain and Taboola’s widgets are embedded on thousands of websites, and both companies generate hundreds of millions of dollars in revenue [12]. However, competition in the CRN space is fierce: there are many incumbent services like Revcontent, Gravity, and ZergNet.

As the name suggests, CRNs are centered around HTML widgets that recommend content to users. Publishers embed these widgets into their pages, and they display two kinds of recommended content: 1) links to content from the publisher, and 2) sponsored content from third-parties. Figures 1 and 2 show typical widgets from Revcontent and Outbrain. CRNs personalize the recommendations shown to each individual to encourage engagement, although the specific mechanisms used by each CRN for personalization are unknown.

CRNs allow publishers to customize the widgets in various ways. This includes: choosing vertical and horizontal layouts; using CSS to style the widget; displaying thumbnail images or text-only links; and choosing how many and what types of links to include (i.e., the fraction of first- and third-party recommendations). Publishers may also choose the headline that is shown at the top of the widget. Figures 1 and 2 show three different headlines chosen by publishers.

For advertisers, CRNs offer a way to promote content across a wide range of publishers. Advertisers supply links with accompanying text and thumbnail images to CRNs, who then “recommend” these sponsored links to users via their widgets. Advertisers pay each time their sponsored links are clicked, and the revenue is split between the CRN and the publisher.

Controversy. Recently, CRNs have been at the center of two controversies. First, CRNs have been ac-

cused of not properly labeling sponsored links as advertisements, which may confuse and mislead users [9, 20]. For example, in Figure 1 the only disclosure is the tiny text “Sponsored by Revcontent” in the upper-right corner of the widget; in Figure 2, the only disclosure is the tiny “what’s this” in the bottom right of the box.

Non-uniform labeling of sponsored links is exacerbated by publishers’ ability to customize the CRN widgets. This leads to webpages that mix sponsored and non-sponsored links underneath misleading headlines. As of 2014, Outbrain claims to have added more prominent labeling to their sponsored links [7].

The second controversy concerns spammy links promoted by CRNs. In theory, CRNs are supposed to be used for “content marketing”, i.e., promotion of blog posts, articles, videos, *etc.* In practice, CRNs have been used to spread links leading to spam and scams [9, 17], which echoes similar issues on traditional ad networks [22]. As of 2012, Outbrain claims to have increased their pre-filtering of content to eliminate spam (causing their revenue to fall 25%) [18]. In contrast, Taboola relies on users to flag spammy content [16].

3. METHODOLOGY

In this study, we focus on five CRNs: Outbrain, Taboola, Revcontent, Gravity, and ZergNet. Our goal is to answer fundamental questions about these CRNs, including: *where do their widgets appear? How are ads disclosed to users? and Who advertises on these services?* To answer these questions, we need large-scale data gathered from CRN widgets.

In this section, we outline our data collection methodology. First, we discuss how we selected publishers to crawl, followed by how we gathered data from each site. Finally, we present definitions that we will use throughout our analysis.

3.1 Choosing Publishers

The first step in our study is choosing websites to crawl. CRNs cater to large publishers by helping them recommend their content to users (in addition to being a source of ad revenue). Armed with this intuition, we collected the list of 1,240 publisher websites that appear in Alexa’s 8 “News and Media” categories in February 2016. Example categories include *News*, *Business News and Media*, and *Health News and Media*. We crawled all 1,240 websites to identify publishers that may embed CRN widgets. We randomly visited five pages per website¹ and analyzed the generated HTTP requests. Out of 1,240 publishers, 289 contacted at least one of the five CRNs.

However, examining large publishers alone may result in a biased sample of CRN widgets. Thus, we also analyzed the HTTP requests generated by all sites in the Alexa Top-1M that we had collected for an earlier

¹We only included pages from the same domain.

study [3]. Out of 5,124 websites from the Alexa Top-1M that contacted a CRN, we randomly sampled 211.

We focus on these 500 publishers in this study: 211 randomly sampled from Alexa Top-1M, and all 289 from our Alexa “News and Media” categories crawl². This sample gives us a balance of top media and news websites, as well as lower-ranked websites.

3.2 Crawling and Parsing

Now that we have chosen 500 publishers, our next step is to collect samples of CRN widgets. To do this, we manually developed a set of XPath queries that correspond to specific widgets from our five target CRNs. These XPaths serve the dual purpose of allowing us to *detect* the presence of widgets in webpages, as well as *extract* specific information from the widgets. In total, we developed 12 XPaths, with most (7) targeting Outbrain, since they have the widest diversity of widgets. Two example XPath queries are:

- Outbrain: `//a[@class='ob-dynamic-rec-link']`
- ZergNet: `//div[@class='zergentity']`

Our crawler works as follows: we visit the homepage of a publisher p , and then proceed to crawl links that point to p until either all links on the homepage are exhausted, or we find 20 pages that include CRN widgets. We also crawl one additional link that points to p from each of the 20 pages, to add another level of depth to our traversal. Finally, our crawler refreshes all 41 pages (i.e., homepage, depth-one, and depth-two pages) three times, to ensure that we enumerate all ads and recommendations offered by the CRNs [10]. The crawler saves all HTML from traversed pages. The crawl was done between February 26–March 4, 2016.

Definitions. Using our XPath queries, we extract specific pieces of data related to CRNs from our raw HTML. This includes the number of widgets per page, and all links in each widget. We label each link as *recommended* if it points to the publisher hosting the widget, and as an *ad* if it points to a third-party (i.e., it is a sponsored recommendation). We also extract the *headline* of each widget (e.g., the “Trending Today” text in Figure 1) as well as any *disclosures* (e.g., the “Sponsored by Revcontent” text in Figure 1). Note that not all widgets have headlines or disclosures.

4. ANALYSIS

In this section, we analyze our dataset in order to gain a better understanding of the CRN ecosystem. *First*, we present general statistics about the online footprint of CRNs and the functionality of their widgets. *Second*, we examine the headlines of CRN widgets, and whether paid sponsorships are being clearly disclosed. *Third*, we briefly look at how CRNs target advertisements based on context and geography. *Fourth*, we investigate the

²Note that the 211 sites do not overlap the 289 sites. We also excluded pornographic sites from the samples.

CRN	Publishers	Total Ads	Total Recs	Average Ads/Page	Average Recs/Page	% Mixed	% Disclosed
Outbrain	147	57,447	35,476	5.6	3.8	16.9	90.8
Taboola	176	56,860	15,660	7.9	1.5	9.0	97.1
Revcontent	29	576	16	6.5	1.3	0	100.0
Gravity	13	744	2,054	1.1	9.5	25.5	81.6
ZergNet	14	15,375	0	6.0	0	0	24.1
<i>Overall</i>	334	130,996	53,202	6.8	2.7	11.9	93.9

Table 1: Overall statistics about our five target CRNs.

# of CRNs	# of Publishers	# of Advertisers
1	298	2,137
2	28	474
3	7	70
4	1	8

Table 2: Number of CRNs used by publishers and advertisers.

relationship between ads and advertisers, and *finally* we examine the content that is advertised via CRNs.

4.1 High-Level Statistics

We begin by presenting a broad overview of our five target CRNs in Table 1. As expected, Outbrain and Taboola are embedded in an order-of-magnitude more publishers than their smaller competitors. We also see that only 334 of our 500 publishers have embedded widgets from CRNs, and yet all 500 request at least one resource from a CRN (see § 3.1). The 166 missing publishers include trackers from CRNs, but do not embed recommendation widgets in their pages.

Table 1 shows the total number of ads (i.e., recommendations sponsored by third-parties) and recommendations we observe from each CRN, as well as average ads and recommendations per page. Four of the CRNs serve more ads than recommendations; ZergNet is a special case, since it *only* serves ads. Outbrain serves $1.5\times$ more ads per page than recommendations on average, while Taboola and Revcontent serve $5\times$ more ads on average. Gravity is the sole exception to this trend: we observe $2.7\times$ more recommendations from Gravity overall, and only ~ 1 ad per page.

Interestingly, we observe cases where a given publisher will embed widgets from competing CRNs into their website. As shown in Table 2, this situation is relatively rare: only 36 publishers use ≥ 2 CRNs. The Huffington Post actually embeds widgets from Outbrain, Taboola, Gravity, and Revcontent. Publishers with multiple CRN widgets may be attempting to capture additional revenue by presenting users with more ads, or the publisher may be conducting an A/B test to compare revenue from competing CRNs.

Mixed Recommendations. Most of the CRN widgets in our dataset include ads *or* recommendations. This makes sense from a user interface perspective: mixing sponsored and organic recommendations in a single container may confuse users. However, 11.9% of CRN widgets in our dataset do not behave this way: the “%

Recommendation Headline	%	Ad Headline	%
you might also like	17	around the web	18
featured stories	12	promoted stories	15
you may like	7	you may like	15
we recommend	7	you might also like	6
more from variety	5	from around the web	2
more from this site	4	trending today	2
you might be interested in	2	we recommend	2
trending now	1	more from our partners	2
more from hollywood life	1	you might like from the web	1
more from las vegas sun	1	more from the web	1

Table 3: Top-10 headlines used for labeling recommendation and ad widgets.

Mixed” column in Table 1 shows the percentage of widgets from each CRN that include ads *and* recommendations. We observe that Taboola, Outbrain, and Gravity all allow publishers to mix ads and recommendations inside a single widget, with 26% of Gravity widgets behaving this way.

We manually examined the widgets with mixed recommendations and found that it was often unclear which links were ads. In some cases, the ads have a small icon next to them that links to a disclosure (similar to the AdChoices  icon). In other cases, the target of each link is stated in parenthesis (see Figure 2 for an example), which explicitly informs the user that the link directs to a third-party, but does not explicitly inform the user that the link is a paid promotion.

4.2 Headlines and Disclosures

Our observations about widgets with mixed recommendations leads directly to our next set of questions: *are CRNs explicitly labeling sponsored links as advertisements?* To investigate this issue, we examine the headlines and disclosures in CRN widgets.

Headlines. Overall, we observe that 88% of CRN widgets have headlines. Of those that do not, 11% contain ads. Table 3 shows the top-10 most common headlines for recommendation and ad widgets³, as well as the percentage of widgets with those headlines. Surprisingly, three of the top-10 headlines are identical for recommendation and ad widgets, and none of the three explicitly state that the links may be advertisements. The most common ad widget headline also does not in-

³Many widgets have headlines that differ by exactly one word, e.g., “You May Like” and “You Might Like”. We cluster these headlines together in Table 3.

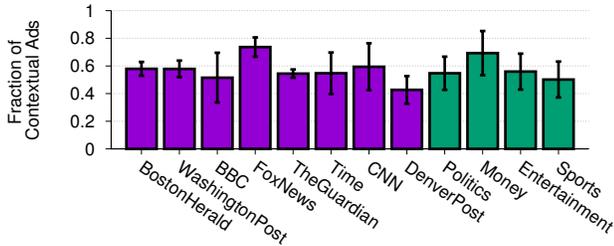


Figure 3: Average contextual ads per Outbrain widget for different publishers and topics.

indicate the presence of sponsored links. Across all headlines for ad widgets, only 12% include the word “promoted”, 2% include “partner”, 1% include “sponsored”, and <1% include “ad” or “advertiser”.

Disclosures. Besides headlines, some CRN widgets include small text snippets, images, or links that disclose the presence of ads (see Figures 1 and 2 for examples). As shown in Table 1, overall, we observe that 94% of CRN widgets include such disclosures. However, this behavior varies significantly across CRNs: for example, Revcontent includes disclosures in 100% of widgets, but ZergNet only includes disclosures in 24%.

Although it sounds heartening that 94% of CRN widgets include disclosures, we observe that the substantive quality of these disclosures varies widely. As shown in Figure 1, Revcontent has the most explicit and uniform disclosures, i.e., the text “Sponsored by Revcontent”. In the 97% of cases where Taboola discloses, they also do so explicitly by including the AdChoices  icon in their widget. In contrast, Outbrain’s disclosures are non-uniform and problematic: some widgets hide the disclosure behind an opaque link (e.g., the “[what’s this]” text in Figure 2), while others include an image stating that the links are “recommended”, rather than paid advertisements (e.g., ).

Summary. Our findings reveal that the vast majority of CRN widgets do not explicitly state that they contain ads. In many cases, widgets do not contain any headline or disclosure. When the widgets do have headlines, they rarely include words that are associated with paid promotion. Furthermore, Outbrain’s disclosures merely reveal that the links are recommended, not that the links are sponsored promotions.

4.3 Ad Targeting

Next, we investigate how Outbrain and Taboola target ads. We focus on these two CRNs because they contribute the vast majority of ads in our dataset. Both CRNs claim to use machine learning to recommend content that each individual is likely to click on, and refine their models based on engagement [9,16]. Additionally, both CRNs give advertisers some flexibility to target their ads: for example, we examined Outbrain’s docu-

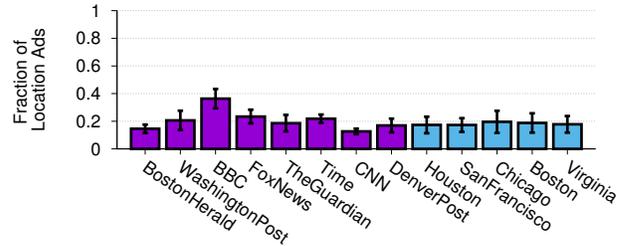


Figure 4: Average location ads per Outbrain widget for different publishers and topics.

mentation, and found that advertisers can specify geographic regions for their ads.

Context. To get some idea of how Outbrain and Taboola target users, we conducted two experiments. First, we investigate whether the CRNs contextually target ads. To examine this, we manually selected four broad topics (*Politics*, *Money*, *Entertainment*, and *Sports*) and eight top-publishers that 1) embed Outbrain or Taboola widgets, and 2) have sections of their websites devoted to all four topics. Next, we manually selected 10 articles in each topic on each publisher (320 total articles), and crawled each article three times to collect data from the CRN widgets.

To identify targeted ads, we compute the difference between the set of ads that appear in articles in a specific topic and the set of ads that appear in all other articles. Intuitively, ads that only appear on articles for a specific topic are likely to be contextually targeted.

Figure 3 shows the fraction of ads from Outbrain that were contextually targeted on each publisher, as well as the fraction of contextual ads for each topic aggregated across publishers (along with standard deviation error bars). We observe that >50% of ads from Outbrain are contextually targeted, with the *Money* topic seeing the heaviest targeting.

We observe similar trends for Taboola: all topics see >50% contextually targeted ads, with the *Sports* topic leading with 64%. We omit these results for brevity.

Location. Second, we investigate whether Outbrain and Taboola target ads based on location. To examine this, we used the Hide My Ass! VPN service to obtain IP addresses in nine major American cities. Using these IPs, we recrawled the 10 political articles we previously selected on all eight top-publishers (we focus on a single topic to control for contextual effects). As before, all 80 pages were refreshed three times.

Figure 4 shows the fraction of ads from Outbrain that are targeted based on location on each publisher, as well as averaged across publishers for a subset of our locations. We observe that only ~20% of ads are location-dependent, with BBC being the exception; we hypothesize that this may be due to the international nature of their audience. For Taboola, we observe slightly higher

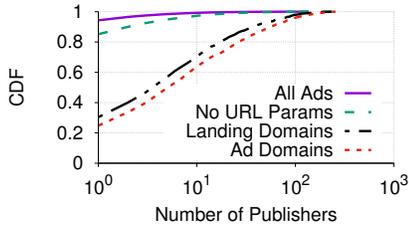


Figure 5: Number of publishers for each ad.

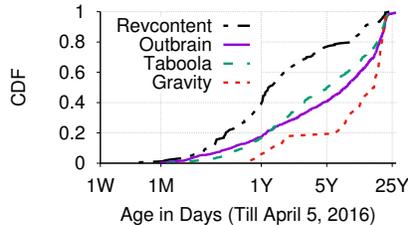


Figure 6: Age of landing domains based on Whois records.

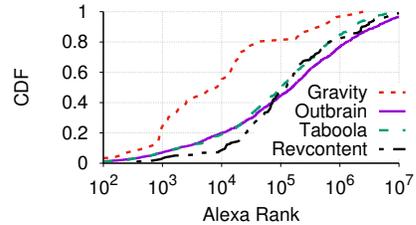


Figure 7: Alexa ranks of landing domains.

(~26%) location-dependent ads (not shown in Figure 4). These results agree with prior work showing that location has a relatively minor impact on online display ads [2, 10].

4.4 Down the Funnel

In this section, we analyze the advertising funnel, starting with ads themselves, and ending with the pages that users are brought to after clicking on the ads.

We begin by examining the uniqueness of ads served by CRNs. The “All Ads” line in Figure 5 shows the distribution of publishers per ad URL in our dataset, i.e., *on how many publishers did a given ad URL appear on?* Of the 131K total ad URLs in our dataset, 94% only appear on a single publisher, i.e., they are unique. However, this statistic is somewhat misleading: we see many ad URLs that include unique IDs in their parameters, which we suspect are used by publishers to implement conversion tracking and A/B testing. The “No URL Params” line plots the distribution of publishers per ad when we filter out URL parameters, and we observe that the percentage of unique ads drop to 85%.

Next, we aggregate ads based on the domain name they point to. The “Ad Domains” line in Figure 5 plots the distribution of publishers per *advertised domain*. We observe that unlike ad URLs, 50% of advertised domains appear on ≥ 5 publishers. The total number of unique advertised domains in our dataset is 2,689, which provides a rough lower-bound on the number of advertisers that were using the five target CRNs between February 26–March 4, 2016 (when we performed our crawls). This finding demonstrates that the predominant strategy used by advertisers on CRNs is to flood them with many unique ads.

# Redirected Sites	# Ad Domains
1	466
2	193
3	97
4	51
≥ 5	42

Table 4: Number of advertised domains that always redirect to other sites.

As shown in Table 2, 79% of advertised domains only appear in widgets from a single CRN. We only observe eight advertisers that leveraged four CRNs. This reveals that advertisers prefer to work with a single platform to distribute their ads.

Redirection. Of course, just because an ad links to a domain d does not mean d is the final destination: d may redirect the user to another domain entirely. To investigate the *landing domains* that users are directed to, we crawled all 131K ads in our dataset. For this crawl, we used a highly instrumented browser that records all information about redirects, even when they are initiated by JavaScript or Flash [1].

Note that when we visit the ad URLs, we bypass the initial redirect through the CRN, meaning that *the advertiser will not be billed by the CRN for our impressions*. We were able to avoid visiting the CRNs due to a quirk in their implementations. All five CRNs embed advertisers’ URLs into their HTML; however, they dynamically replace the advertiser URL with a link pointing to the CRN when a user clicks the link. In our case, we do not click on advertiser URLs, and thus never trigger the dynamic redirects. Rather, we extract the advertiser URLs from the HTML and visit them separately using our instrumented browser.

The “Landing Domains” line in Figure 5 plots the distribution of publishers per landing domain. Surprisingly, we see an increase in the number of unique landing domains compared to ad domains (from 25% to 30%). This reveals that some of the ad domains redirect users to other sites. Table 4 shows that there are 466 ad domains that always redirect to a specific landing domain, while 383 ad domains redirect to >1 landing domain. The ad domain with widest fanout in our dataset is DoubleClick, which redirected to 93 different landing domains.

4.5 Advertisers and Content

Now that we have traveled down the advertising funnel, we turn our attention to the advertisers themselves, and the content that is being advertised via CRNs.

Advertiser Quality. We use two metrics to assess the “quality” of advertisers on CRNs: the age and Alexa rank of their landing domains. Intuitively, domains that were registered recently have not had time to

Topic	Example Keywords	% of Landing Pages
Listicles	improve, scams, experience	18.46
Credit Cards	credit, card, interest	16.09
Celebrity Gossip	Kardashians, sexiest, caught	10.94
Mortgages	mortgage, HARP, loan	8.76
Solar Panels	solar, energy, panel	6.29
Movies	Hollywood, Batman, Marvel	5.90
Health & Diet	diabetes, fat, stomach	5.62
Investment	Dow, dividend, stocks	1.57
Keurig	coffee, Keurig, taste	1.21
Penny Auctions	auction, bid, pennies	1.15

Table 5: Top-10 most frequent topics extracted from landing pages.

build up a positive reputation. Similarly, we would not expect scammers or shady businesses to achieve high Alexa ranks, which are based on visitor volume.

Figure 6 shows the age of landing domains in our dataset, based on Whois records. We calculate age relative to April 5, 2016. We observe that Revcontent’s advertisers have the youngest domains, while Gravity’s have the oldest. Note that we do not analyze ZergNet because *all* of the ads they serve point back to the ZergNet homepage, which is simply a launchpad for third-party, promoted content.

Figure 7 plots the Alexa ranks of landing domains in our dataset. We observe the same trends as in Figure 6: Gravity’s advertisers have the highest ranks, while Revcontent’s have the lowest.

The results in Figures 6 and 7 reveal that Gravity caters to older, more established web properties. Gravity is owned by AOL, and thus it is not surprising that it tends to advertise well-known, AOL-owned properties like aol.com and techcrunch.com. In contrast, Revcontent serves ads for obscure websites like Buzzfeed-knockoff thebuzzstuff.com. Outbrain and Taboola fall somewhere in the middle, advertising a small number of reputable properties and a long tail of unknown properties.

Ad Content. Next, we investigate the landing pages’ content associated with 131K ads in our dataset, to answer the question *what is being advertised?* To answer this question, we used Latent Dirichlet Allocation (LDA) [4] to extract topics from our corpus of landing pages. LDA uses statistical sampling to identify k groups of words that frequently co-occur in documents; each group represents a coherent topic. In our analysis, we experimented with $20 \leq k \leq 100$, but found that $k = 40$ produced the most succinct topics.

Table 5 shows the top-10 topics extracted from the landing pages, sorted by frequency. We observe that $\sim 20\%$ of all landing pages are about the *Mortgage* or *Credit Cards* topics, epitomized by words like “mortgage”, “credit” and “loan”. 19% of landing pages are listicle-style articles (e.g., “8 Pro-Tips For Improving Your IMC Review Scores!”). Other frequent topics include celebrity gossip, “miracle” diets, investment ad-

vice, and penny auctions. Overall, these 10 topics cover 51% of the landing pages in our dataset (note that some pages may fall under multiple topics, e.g., a listicle about weight loss).

The results in Table 5 confirm many of the concerns about CRNs that have been identified in the press [9, 17, 18]. Specifically, we observe that many of the most commonly advertised topics are not “content”, but commercial offers related to financial services, penny auctions, and medical services. Other topics are “click-bait” centered around bombastic celebrities (e.g., Kardashians).

5. CONCLUDING DISCUSSION

In this paper, we present the first evaluation of Content Recommendation Networks (CRNs). CRNs have become so ubiquitous that their headlines are clichés; indeed, we find that CRN widgets are embedded in 23% of the most popular publishers from Alexa’s “News and Media” categories.

However, CRNs have also been a source of controversy. In 2014, Outbrain and Taboola (the leading CRNs) were told by government regulators and industry trade groups to prominently disclose the presence of promoted links in their widgets [7, 20]. Similarly, CRNs in general have been repeatedly faulted in the press for recommending spammy content and scams [9, 17, 18].

Using our dataset, we find that these issues have not been fully rectified. Only $\sim 15\%$ of CRN widgets have headlines stating that content is “sponsored” or “promoted” (see Table 3), and only two CRNs in our study (Taboola and Revcontent) consistently include an informative disclosure in their widgets (see Table 1). With respect to content quality, we observe that CRNs continue to serve ads for dubious financial services, celebrity gossip, diet schemes, and penny auctions.

Our findings point to the need for further intervention in the CRN market by government regulators and industry groups. At a minimum, CRNs should conform to accepted best-practices like the Adchoices program, as Taboola already does. CRNs could also make progress towards correcting disclosure problems by making their widgets more uniform, as Revcontent already does. Finally, CRNs could remove or restrict publishers’ ability to customize widget headlines, and enforce clear labels like “Paid Content”. We reached out to the five companies examined in this study via their public press contacts, but none responded to our inquiries.

Acknowledgements

We thank our shepherd, Georgios Smaragdakis, and the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants IIS-1408345 and IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

6. REFERENCES

- [1] ARSHAD, S., KHARRAZ, A., AND ROBERTSON, W. Include me out: In-browser detection of malicious third-party content inclusions. In *Proc. of Intl. Conf. on Financial Cryptography* (2016).
- [2] BARFORD, P., CANADI, I., KRUSHEVSKAJA, D., MA, Q., AND MUTHUKRISHNAN, S. Adscape: Harvesting and analyzing online display ads. In *Proc. of WWW* (2014).
- [3] BASHIR, M. A., ARSHAD, S., ROBERTSON, W., AND WILSON, C. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *Proceedings of the 25th USENIX Security Symposium* (2016).
- [4] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3 (2003).
- [5] CAHN, A., ALFELD, S., BARFORD, P., AND MUTHUKRISHNAN, S. An empirical study of web cookies. In *Proc. of WWW* (2016).
- [6] CARRASCOSA, J. M., MIKIANS, J., CUEVAS, R., ERRAMILI, V., AND LAOUTARIS, N. I always feel like somebody's watching me: Measuring online behavioural advertising. In *Proc. of ACM CoNEXT* (2015).
- [7] DAVIS, W. Better business bureau tells taboola to make 'sponsored content' disclosures more prominent. *OnlineMediaDaily*, May 2014. <http://www.mediapost.com/publications/article/226254/better-business-bureau-tells-taboola-to-make-spon.html>.
- [8] GILL, P., ERRAMILI, V., CHAINTREAU, A., KRISHNAMURTHY, B., PAPAGIANNAKI, K., AND RODRIGUEZ, P. Follow the money: Understanding economics of online aggregation and advertising. In *Proc. of IMC* (2013).
- [9] GRIFFITH, E. How taboola and outbrain are battling a bad reputation... and each other. *Fortune*, August 2014. <http://fortune.com/2014/08/18/taboola-outbrain-battle-bad-reputation-each-other/>.
- [10] GUHA, S., CHENG, B., AND FRANCIS, P. Challenges in measuring online advertising systems. In *Proc. of IMC* (2010).
- [11] HA, A. Youtube introduces six-second bumper ads. *TechCrunch*, April 2016. <http://techcrunch.com/2016/04/26/youtube-bumper-ads/>.
- [12] HIRSCHAUGE, O. Outbrain, taboola make their mark on online advertising industry. *The Wall Street Journal*, March 2015. <http://on.wsj.com/1FkpIGR>.
- [13] KRISHNAMURTHY, B., NARYSHKIN, K., AND WILLS, C. Privacy diffusion on the web: A longitudinal perspective. In *Proc. of WWW* (2009).
- [14] KRISHNAMURTHY, B., AND WILLS, C. Privacy leakage vs. protection measures: the growing disconnect. In *Proc. of W2SP* (2011).
- [15] KRISHNAMURTHY, B., AND WILLS, C. E. Generating a privacy footprint on the internet. In *Proc. of IMC* (2006).
- [16] LAWLER, R. Taboola now lets you filter out content recommendations that you don't want to see. *TechCrunch*, September 2013. <http://techcrunch.com/2013/09/04/taboola-choice/>.
- [17] MARSHALL, J. Content marketing's got a quality problem. *Digiday*, May 2013. <http://digiday.com/publishers/content-marketings-got-a-quality-problem/>.
- [18] RAY, J. D. Outbrain expects 25% revenue hit as it cuts off spammy content marketers. *AvertisingAge*, November 2012. <http://adage.com/article/digital/outbrain-cuts-spammy-marketers-expects-revenue-hit/238200/>.
- [19] ROESNER, F., KOHNO, T., AND WETHERALL, D. Detecting and defending against third-party tracking on the web. In *Proc. of NSDI* (2012).
- [20] SWENEY, M. ASA ruling on outbrain link heightens 'native advertising' debate. *The Guardian*, June 2014. <http://www.theguardian.com/media/2014/jun/18/asa-outbrain-native-advertising-link>.
- [21] VALLINA-RODRIGUEZ, N., SHAH, J., FINAMORE, A., GRUNENBERGER, Y., PAPAGIANNAKI, K., HADDADI, H., AND CROWCROFT, J. Breaking for commercials: Characterizing mobile advertising. In *Proc. of IMC* (2012).
- [22] ZARRAS, A., KAPRAVELOS, A., STRINGHINI, G., HOLZ, T., KRUEGEL, C., AND VIGNA, G. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proc. of IMC* (2014).