Market Power of News Aggregators*

Ruben Durante† Ruben Enikolopov‡ Daniil Mikhailov§
Andrey Simonov¶

March 20, 2025

Preliminary – Please do not circulate

Abstract

We measure the ability of news aggregators to impact news consumption and production through changes in their algorithms. Our empirical context is the online news market in Russia and the largest news aggregator in the country, Yandex News. Using the universe of articles published by major online news outlets and snapshots of Yandex's front page news citations in 2013-2018, we detect news events and estimate Yandex's news aggregation algorithm. We document that Yandex adjusted its algorithm in response to a Russian law in 2016 that made aggregators responsible for the content they cite, reducing references to independent media outlets by about 50%. This reduction in references was consequential for outlets' traffic: we estimate that Yandex can boost an outlet's traffic by 400 thousand daily visitors (a 23%-1,000% traffic increase for outlets in our sample) by referencing the outlet on the front page. The news aggregator faces little constraint on which topics and outlets it cites; the effectiveness of Yandex in generating traffic for news outlets does not change after the law and is similar across news topics and types (e.g. news censored by government-owned media). As a result, Yandex has the ability to substantially influence market shares of upstream news outlets and which news people read. We further show that news outlets removed from Yandex's references modify their production decisions, specializing in longer stories and investing less into optimizing their news for Yandex.

*This project is supported by the Chazen Institute for Global Business at Columbia Business School. Ruben Enikolopov acknowledges the financial support of grant PID2023-150768NB-I00 from the Spanish Ministerio de Ciencia e Innovación. We thank seminar participants at Carnegie Mellon University, Columbia University, Free University of Bozen-Bolzano, HEC Lausanne, LUISS, National University of Singapore, Stigler Center at the University of Chicago, UC Riverside, University of Colorado Boulder, Digital Competition and Tech Regulation at Harvard Business School, Conference on AI, ML, and Business Analytics at Yale SOM, Polarize Conference at Bergen, and 17th Digital Economics Conference at Toulouse School of Economics for helpful comments and suggestions.

†National University of Singapore, ICREA-UPF, CESifo, IZA, and CEPR. Email: rubendurante@gmail.com
‡ICREA-Universitat Pompeu Fabra, IPEG and CEPR. Email: Ruben.Enikolopov@upf.edu
§Lomonosov Moscow State University. Email: daniilmikhailov11@gmail.com
¶Columbia University and CEPR. Email: as5443@gsb.columbia.edu.

1 Introduction

News aggregators have become a vital component of the online news ecosystem, generating approximately 14% of all referral traffic to news publishers in 2022 (digiday.com, 2022). Existing academic work indicates that most of this traffic is incremental, especially for smaller and more niche news outlets (Calzada and Gil, 2020; Athey et al., 2021). As a result, the market expansion effect of news aggregators as a whole dominates the business stealing effect (Jeon and Nasr, 2016), alleviating concerns about the violation of misappropriation laws and copyrights by aggregators (e.g. Federal Trade Commission, 2010). Thus, on balance, aggregators are viewed as beneficial for the news industry and publishers.

In this paper, we take a different perspective and look at news aggregators as large attention bottlenecks that can hold considerable market and media power (Prat, 2018; Prat and Valletti, 2022). Such power stems from their ability to control and potentially manipulate the algorithm that ranks news events and articles. This commitment to an algorithm puts more constraints on aggregators compared to upstream publishers, who have discretion over which news to write. However, aggregation algorithms are non-transparent and can be adjusted in real time. Furthermore, news aggregators might be facing consumers who are less elastic to the composition of news events, since aggregators are frequently integrated into larger ecosystems (e.g. Google News, Apple News).

We investigate to what extent major news aggregators can influence news consumption and production through changes in their algorithms. We focus on the case of the Russian news aggregator Yandex News between 2013 and 2018. This case represents an ideal testing ground for various reasons. First, during this period, Yandex was one of the two largest search engines in Russia (with a market share similar to Google), and Yandex News was the largest news aggregator.[1] Second, over the same period, Yandex made major changes to its algorithms in response to changes in Russian legislation and pressure from the Russian government(Wijermars, 2021); this situation generates substantial variation useful to identify the impact of algorithms on news consumption and production. Finally, news stories ranked as the five most important of the moment on Yandex News are prominently featured on the main page of the Yandex search engine, which is visited by tens of millions of users daily. This amplifies the impact of Yandex News's ranking on the traffic directed to news websites.

Our analysis employs three novel datasets. First, we scrape snapshots of Yandex's homepage and of the top-5 section, using the Wayback Machine available from the Internet Archive. For the period 2013-2018, we collect 42,721 snapshots, with a median time of 29 minutes between

[1] In early 2023, Yandex still held leadership in terms of referral traffic, responsible for 19% of referrals to news outlets (Statista, 2023). In September 2022, Yandex sold its news service – together with its main domain, `yandex.ru` – to VK.com (Yandex, 2022); the main page with the news service was rebranded as `dzen.ru`. Throughout the paper, we will refer to "Yandex News" as "Yandex," unless we specify otherwise.

two snapshots. These records include the titles and links to articles cited in the top-5 section. We augment these data with snapshots of the overall Yandex News ranking (`news.yandex.ru`), including stories outside of the top-5.

Second, to measure news production and the overall composition of news events in Russian media, we acquire the universe of publication records covering 50 major Russian online news websites. The list of outlets includes almost all major federal news outlets in Russia between 2013 and 2018, both government-owned and independent. Overall, the data contains 15.1 million articles, including for each article the outlet's name, the article's title and text, the exact time of publication, and the corresponding URL.

Third, we use data from the traffic tracker `top.mail.ru` to measure news consumption of a subset of websites. Websites install this tracker to facilitate marketing analysis and can decide to make different types of data available to the public. We observe daily traffic for eleven major outlets from our sample, broken down by types of visitors (e.g., new/regular, mobile/desktop, age, gender), visit duration, and other factors. This data allows us to measure the effects of Yandex News references on the number of visitors to each outlet. In addition, for four outlets, we also observe a more detailed breakdown of the traffic by its referral source (e.g., search engines, social media), exact pages visited, etc. We use these traffic breakdowns to confirm the incrementality of the traffic from Yandex and measure spillover effects. We confirm the quality of this data using other public trackers.

We first reverse-engineer how Yandex's news aggregation algorithm works. The algorithm, which is similar to that used by other news aggregators, proceeds in two steps. In the first step, the algorithm ranks news events – i.e., clusters of recent news articles – based on their newsworthiness, and decides which events to cite. In the second step, the algorithm determines which outlets to reference for a given event.

To estimate Yandex's algorithm, we first group articles into news events using a topic detection and tracking algorithm (Allan et al., 1998). This method, used in previous work (e.g. Cagé et al., 2020), takes into account both the similarity in the article's text and the time of publication. We then map each event to the Yandex's citations using articles' titles and links.

Regarding the algorithm's choice of events in its first step, we find that events covered more extensively by news outlets are more likely to be cited by Yandex. The median number of articles is 29 for events cited by Yandex compared to 2 for other events. Yandex is also more likely to cite news events that are covered by more outlets, last longer, and develop faster. Interestingly, keeping everything else equal, Yandex tends to under-report news events that are covered primarily by independent but not by controlled outlets.

Regarding the choice of outlets conditional on news events, we find that an outlet is more likely to be referenced by Yandex if it publishes more articles about the event, does so faster, mentions more facts, has shorter titles, and tends not to copy previous news articles about the

same event excessively. However, even after accounting for these factors, we observe that a lot of the variation in references is explained by outlet weights, captured as fixed effects. These weights can change over time seamlessly and allow Yandex to adjust its algorithm.

We observe a major change in Yandex's algorithm in August 2016, shortly after Vladimir Putin signed a major law "on news aggregators." The law made news aggregators liable for the content they cite, with some exceptions based on the source of news (Wijermars, 2021). In a sudden and dramatic change, Yandex stopped referencing 9 out of 24 outlets that used to regularly appear in the "top-5" news slots before the change, primarily targeting independent outlets. As a result, the share of top-5 references to independent outlets dropped from 20% to 10%, with all independent outlets entirely excluded except for two business-focused outlets. Interestingly, this change only affected the distribution of outlets in Yandex's top-5 section, but not the type of events featured therein, which remained largely unchanged after the law came into effect.

Having estimated Yandex's algorithm, we turn to estimating the impact of Yandex top-5 references on the upstream outlets' traffic. Our identification strategy relies on day-to-day variation in Yandex references, driven by quasi-random realizations of the share of references an outlet gets on a given day. To isolate this variation, we include outlet-by-week, outlet-by-day-of-the-week, and day fixed effects in a regression of daily outlet traffic on the outlet's citation in Yandex's top-5. We validate this empirical strategy using referral data available for a subset of outlets, showing that Yandex references have no impact on an outlet's direct traffic or referral traffic from other sources.

Our results indicate that Yandex references have a large impact on traffic. In particular, a one standard deviation increase in the daily share an outlet appears in at least one top-5 slot increases outlet traffic by around 27,600 visitors. To put this into perspective, if all slots in Yandex's top-5 positions for an entire day were allocated to a single outlet, its traffic would grow by nearly 380,000 visitors. This represents a traffic increase of 23% for the largest outlet in our sample, and of nearly 1,000% for the smallest one.[2] The effect is larger for higher positions in the top-5 list, ranging from +61,000 viewers for the fifth slot to +96,000 viewers for the first. In contrast, appearing in positions 6-10 of Yandex News leads to effects on outlet traffic that are both smaller in magnitude and imprecise. This is due to the lower visibility of news events in positions 6-10, which do not appear on Yandex's front page.

We run additional tests to further validate our empirical strategy. First, we show that outlets referenced in Yandex's top-5 have a significant effect on traffic from Yandex referrals, but not on direct traffic, or on traffic mediated by other sources such as social media and other news aggregators. Second, appearing on Yandex's top-5 only affects an outlet's traffic on the day

[2]The effects are in traffic levels since the magnitudes are pinned down by the number of visitors to Yandex's front page and not by the outlet size. As a result, smaller outlets benefit disproportionately from Yandex's references relative to their existing traffic.

it is referenced but has no detectable effect on future traffic, ruling out strong intertemporal spillovers. Third, Yandex's references affect casual consumers – who do not regularly visit the outlet and make fewer page visits – more than the regular ones.

Having measured the impact of Yandex's algorithm on upstream outlets' traffic, we check whether Yandex's ability to direct traffic is constrained by referenced outlets, cited topics, or types of visitors on their main page. We find little evidence for such constraints. First, the effectiveness of Yandex is maintained after the passing of the news aggregators law, after which Yandex penalized independent outlets. We see no evidence of users losing confidence in the aggregator and discount its recommendations.

Second, there are limited detectable differences in the effectiveness of Yandex in directing traffic to different types of news articles. We show that an average Yandex visitor is mostly indifferent between broad news topics (e.g. politics, sports) and news events censored by government-owned outlets that Yandex cites (e.g. political opposition, corruption). However, visitors are much (650%) more likely to read local news about their region, and 50% less likely to read news that have low newsworthiness based on our estimates of Yandex's citation algorithm. We conclude that Yandex has a lot of power over changing the composition of news diets of visitors but needs access to upstream news producers to correctly identify newsworthy events.

Third, we compare the ability of Yandex to direct traffic to news outlets across user characteristics. We find no difference in this ability across operating systems (e.g. iOS, Windows) and device types (e.g. mobile, desktop), and find stronger effects for visitors that are identified by the traffic tracker as older and male.

Given a limited response of consumers to the events that Yandex cites and outlets that it references, we evaluate two sets of counterfactual scenarios: a scenario where Yandex can direct all traffic to a single outlet, and a scenario where Yandex is required to keep the same algorithm as in step one (cite the same events) but otherwise can direct as much traffic to a single outlet as it likes. Without restrictions on which events to cite, Yandex would be able to increase the market shares of outlets from 23% (for ria.ru, the largest outlet in our sample) to 989% (for trud.ru, the smallest outlet). We expect that the revenues of these outlets would increase proportionally, given that none of the outlets use a subscription model and instead rely on display advertising for monetization. If Yandex had to cite the same news events as they are citing now, they could increase the market shares of outlets from 13% to 330%. The latter numbers highlight the ability of news aggregators to impact the consumption of upstream news producers even if they commit to using a particular aggregation algorithm.

Finally, we document how news outlets adjust their production decisions in response to changes in Yandex's algorithm. For this, we leverage the change in Yandex's algorithm in response to the law on news aggregators that removed 9 out of 24 outlets from the top-5 references. Comparing the evolution of the news coverage of these 9 outlets with the remaining outlets, we

show that the "removed" outlets changed their production in both consequential and inconsequential ways. First, these outlets started to write fewer "spot" news cited by the aggregator and focused on longer "feature" news. Second, they started to optimize their titles less, with the length of titles increasing by 7-8 symbols and having more punctuation, while Yandex's aggregator prefers articles with shorter titles and less punctuation. It took around 10 months for the news outlets to adjust their production decisions to a new level.

To our knowledge, this paper presents the first large-scale empirical evidence on the ability of news aggregators to impact news consumption and production through the control of their citation and referral algorithms. Existing work has shown that the general presence of aggregators and social media platforms increases news consumption from upstream producers (Sismeiro and Mahmood, 2018; Calzada and Gil, 2020; Athey et al., 2021) and has used natural experiments to show that adding or removing outlets from news aggregators impacts the outlets' traffic (Chiou and Tucker, 2017; George and Hogendorn, 2020). The related work used case studies of changes in the algorithms of social media and search engines to show their effect on news readership and engagement (Calzada et al., 2023; Dujeancourt and Garz, 2023; Moehring, 2024). Our contribution to this stream of literature is in reverse-engineering and characterizing the algorithm that aggregators use and showing that consumers are inelastic to the algorithm adjustments of the aggregator, giving aggregators the power to preference particular outlets in their ranking and direct traffic to these outlets. We estimate Yandex's algorithm and show that it prioritizes particular outlets and down-weights government-sensitive news stories, a behavior similar to platform bias in other contexts (Reimers and Waldfogel, 2023).

We show that news aggregators impact upstream news production due to the incentives of editors to produce and package news content in a way that will be referenced by the aggregator; these results relate to the theoretical work on product differentiation of outlets due to news aggregators (Dellarocas et al., 2013; Jeon and Nasr, 2016) and to the empirical work on the incentives behind news production (e.g. Sen and Yildirim, 2016; Cagé et al., 2022; Boxell and Conway, 2022; Sen et al., 2023; Garz and Szucs, 2023). Our work is also related to the growing literature that examines the effects of adjusting algorithms or recommendation systems of platforms on user engagement, both in the context of news (Peukert et al., 2023) and more broadly (e.g. Anderson et al., 2020; Holtz et al., 2020; Aridor et al., 2022).

This paper also contributes to the literature on the political economy of mass media. We show how legislation in Russia has forced a news aggregator to adjust its algorithm and remove a set of outlets from its references, providing empirical evidence on media capture of a news aggregator (e.g. see Enikolopov and Petrova, 2015, for an overview).

The rest of the paper is organized as follows. Section 2 describes our data sources. Section 3 explains how we classify news articles into events, as well as describes the events. Section 4 presents the estimates of Yandex's news aggregation algorithm. Section 5 estimates the effect of

referencing outlets in Yandex's top-5 section on the outlets' traffic and evaluates the outlets' traffic under counterfactual algorithms of Yandex. Section 6 shows the effect of Yandex's algorithm on upstream news production. Section 7 concludes.

2 Data

We combine three novel data sources to characterize the Yandex News algorithm and measure its effect on news consumption and production: the citations of Yandex News, publication records from major news outlets in Russia, and records of news consumption.

2.1 Yandex News

The first dataset that we create contains historical citations of news by Yandex News in Russia. We construct these citations using the Internet Archive, a digital library containing historical snapshots of various web pages. Our primary interest is in the five most important stories of the moment that are shown on the main page of `yandex.ru`; Figure 1 shows an example of this web page. The top-5 news stories are global (similar for all visitors) and are frequently updated; a click on a news title directs a user to the referenced news outlet.[3]

Since Yandex's front page is visited by tens of millions of users daily, the Internet Archive has a very detailed collection of its snapshots (the number of snapshots is proportional to the popularity of the web page). Our sample contains 42,721 snapshots of the front page covering the period between 2013 and 2018. The median time between two snapshots is 29 minutes. From each snapshot, we extract the top-5 news titles and the underlying URL links. We augment the data with snapshots of the news rankings from the Yandex News (`news.yandex.ru`) page.

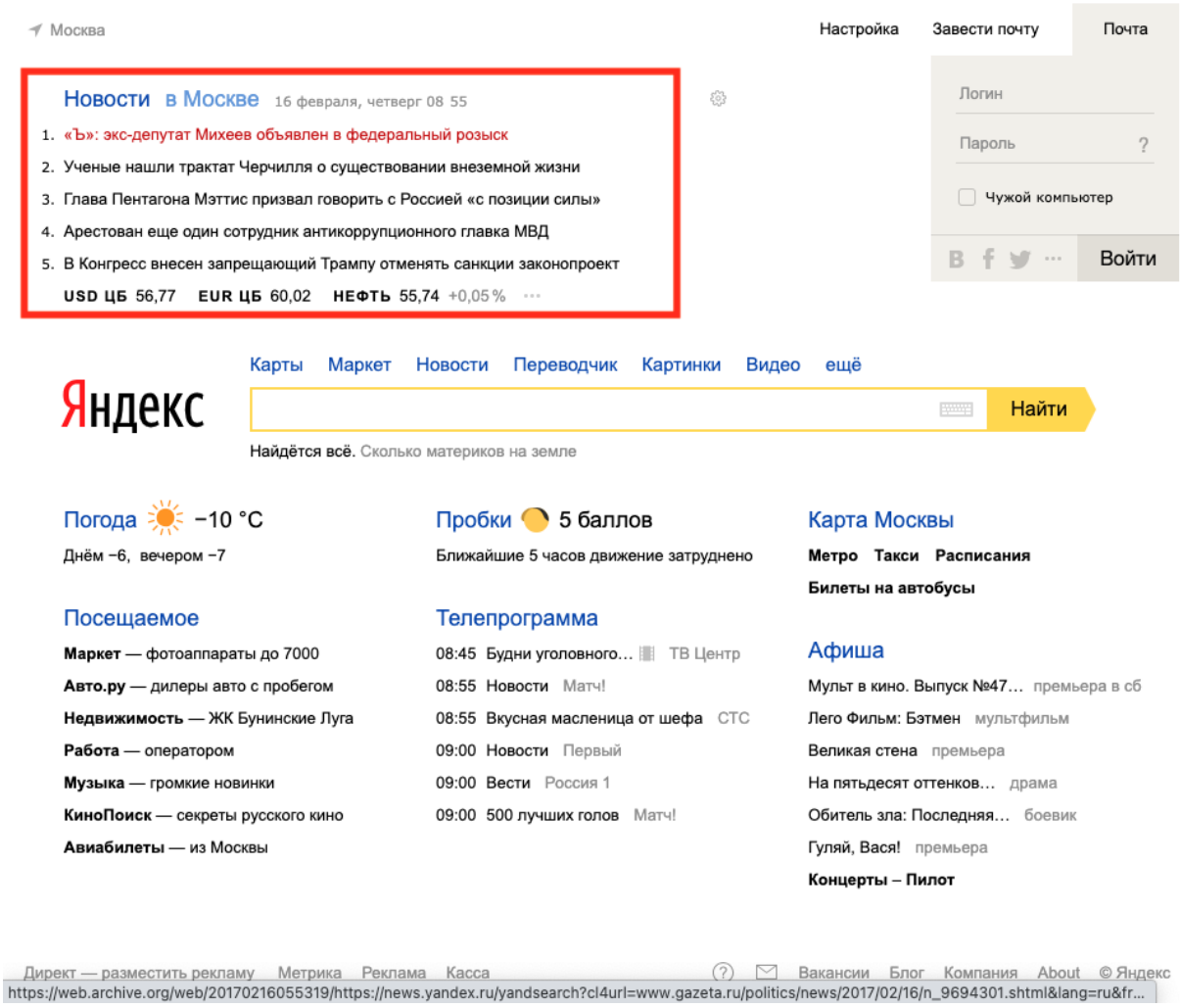
2.2 Publication Records

We complement the Yandex citation records with the publication records of 50 major online news outlets in Russia at the time. Data were provided by Medialogia, a public relations company that collects the universe of social media posts and media publications in Russia (`mlg.ru`). The publication records include outlet names, article titles and texts, the publication time and URLs. The data contains 15,149,664 news articles published during the period from 1 January 2013 to 31 October 2018.

Both government-controlled and independent news outlets coexisted in the Russian media space at the time. We use outlets' ownership structure, location, and information on the independence of their editorial boards to classify news outlets into four groups: government-controlled

[3]This design changed in April 2018, when Yandex started to direct users to their subdomain that describes the news story. We focus on the period before this change.

Figure 1: An Example of the Front Page of Yandex



The snapshot depicts the front page of `yandex.ru` on February 16, 2017, at 8:55 am GMT+3. Links in the red rectangle are global top-5 news of the moment, similar across all locations. These links refer a user directly to the news website that published the referred article, as shown at the bottom of the image (the cursor hovers over the first link, making it red).

(GC), directly owned and controlled by the government; independent but potentially-influenced, privately owned but with close tie to the government or information that editorial board is not independent; independent, which are privately owned and have independent editorial boards; and international, which are based outside of Russia but focus their news on the Russian audience. This classification closely follows Simonov and Rao (2022).

We match the publication records to the news that Yandex cite in two steps. First, for Yandex, we can match published and cited news articles directly by the article URLs, since they are observed on the front page of Yandex and in the publication records. Second, we cluster publication records into news events and match Yandex-cited news to the corresponding events using the distance between the titles of the cited articles and the center of the cluster.

2.3 Consumption Data

The third dataset that we create measures the daily news consumption of a subset of major news websites in our sample. We assemble this dataset using a traffic tracker of `top.mail.ru`. Websites install this tracker to help with marketing analysis and can choose to make different types of data public. Eleven out of fifty outlets in our sample keep their overall traffic open for the public; this includes daily records of the total visitors, visitors split by the intensity and frequency of visits (e.g. no visits in the last 31 days), total pageviews, desktop and mobile visitors, among other metrics. For a subset of these websites, we observe more detailed statistics, such as daily visits of particular webpages, location of visitors, their browser, and operation system. In particular, four websites provide data on the source of their traffic, allowing us to separate traffic from Yandex, other news aggregators, or other referral domains.

We confirm the quality of data from `top.mail.ru` using other public data trackers in Russia. In particular, we use `liveinternet.ru` and `rambler.ru`; while these domains do not allow us to pull historical data from 2013-2018, we examine the correlation in the traffic records from these trackers and `top.mail.ru` in the more modern data.

3 Classification of News Events

Before proceeding to the main analysis, we use publication records to detect and characterize news events that occurred in Russia during the period of interest.

We detect news events from texts and titles of news articles using the topic detection and tracking algorithm (TDT), a common tool developed by Allan et al. (1998) and used to detect ongoing news events (e.g. see Cagé et al., 2020). We process the texts by removing the html tags, stemming, and tokenizing the texts and titles. We then translate them into numeric vectors. For this, we treat texts and titles as bags of words, weighting words by TF-IDF (term frequency-inverse document frequency). The weighted counts of words represent the news articles.

We then cluster news articles based on the resulting numeric representations of the news articles. The documents were clustered from the bottom up. First, for all pairs of documents that cooccur in a two-day period, we compute the cosine similarity of their representations. Two documents are assigned to a similar news event (cluster) if the cosine similarity of their representations is lower than the clustering distance threshold.[4] We apply this classification to all subsequent two-day partitions of the data; if there is a news article that belongs to two different clusters across the overlapping time periods, we redefine the news event as a union of these overlapping clusters. As a result, the news event is finalized if no new similar news articles

[4]In our main event clusters, we use DBSCAN clustering with a 0.35 clustering distance threshold. We have also tested alternative algorithms, HDBSCAN (McInnes et al., 2017) and agglomerative (hierarchical) clustering with several linkage criteria (ull, ward, and average).

are published the next day.

The resulting classification produces 1.33 million news events, where an event is defined if it has at least two news articles clustered together. In total, 50.9% of 15.1 million news articles are classified into events; the rest are treated as standalone articles. Table 1 presents the summary statistics of the events; a median event has two articles from two different outlets and lasts 3.4 hours (taking the difference in the time of publication of the first and last articles). The distribution of the number of articles per event is heavily skewed to the right, with around 49 thousand events having at least 20 articles, and the largest detected event having 6,615 articles and covered by 49 outlets in our sample.

Table 1: Summary Statistics: All Events

	Min	25%	Median	Mean	75%	Max
articles per topic	2	2	2	5.79	4	6615
outlets per topic	1	2	2	3.81	4	49
topic length (hours)	0	1.07	3.35	11.12	13.70	3168.13

Based on the 1.33 million news events detected from raw articles data; an event is defined if at least 2 articles form a cluster.

Table 2 summarizes the events from our publication records data that appear in Yandex's top-5 news at least once in our data. We identify the events cited by Yandex that match the article links in Yandex citations and the publication records data. As expected, the events cited by Yandex are larger and last longer than regular events. For instance, the median event cited by Yandex has 29 articles about it, is covered by 22 outlets, and lasts for 23 hours.

Table 2: Summary Statistics: Yandex Top-5 Events

	Min	25%	Median	Mean	75%	Max
articles per topic	2	12	29	75.15	65	6615
outlets per topic	1	11	22	20.89	30	49
topic length (hours)	0	7.25	23.07	46.09	57.57	2640.7

Based on the 24 thousand news events that are cited in top-5 news on Yandex at least once.

We also classify events as whether they are sensitive and censored by government-controlled outlets. For this, we follow the procedure in Simonov and Rao (2022) to detect events that receive systematically more coverage from the majority of independent outlets compared to the majority of controlled outlets. The procedure ranks news outlets in terms of their reporting and takes the average rank score difference between controlled and independent outlets on each news event. We apply the procedure only to events that were reported by more than 10 outlets. Tables 10 and 11 in Appendix A.1 present the most over- and under-reported events by independent outlets. Examples of stories reported more by independent media are related to the activities of Alexey Navalny and changes in the editorial policies and management of independent media;

government-controlled media report more on the Russia support of Donbass and the construction of the bridge to Crimea.

Finally, we group news articles into general news topics. For this, we estimate a structural topic model (Roberts et al., 2013) based on article texts and detect 100 different topics. We grouped these topics into 9 broader categories which were used by Yandex.News.

4 News Aggregation Algorithms

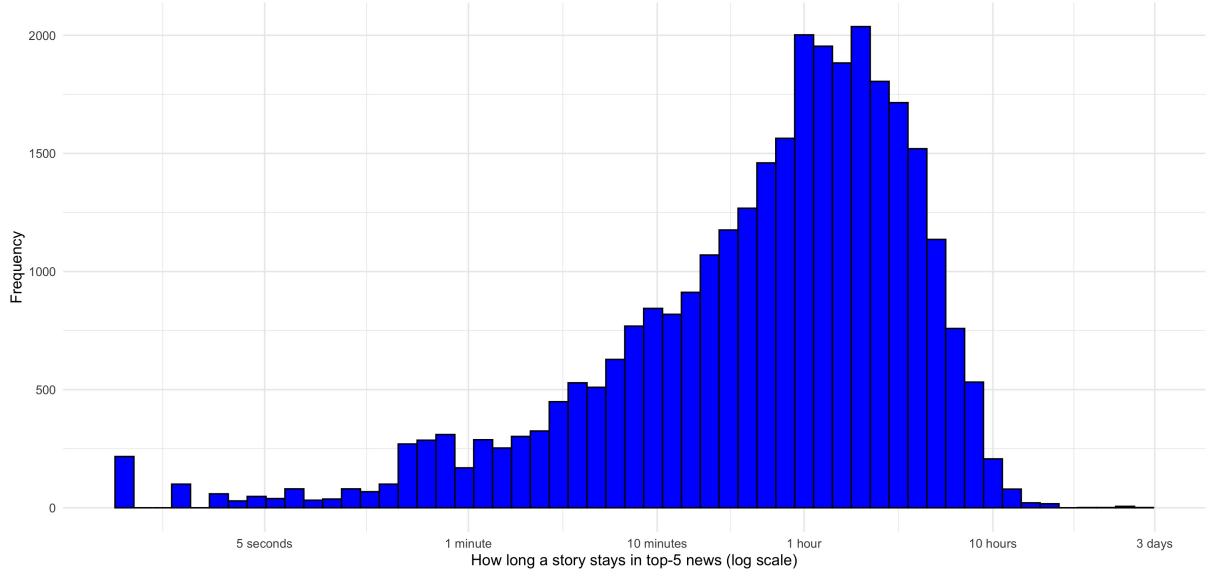
Building on the classification of news events, we now proceed to reverse-engineer and describe the news aggregation and citation algorithm employed by Yandex.

4.1 Descriptive Evidence

We start by establishing some facts about the Yandex citation process.

First, we observe a lot of rotation of news articles in the top-5 news. On a typical day, we observe 60-70 news articles in the top-5 news section of Yandex. Figure 2 presents a distribution of the duration of the news articles remaining in the top-5 section, based on the articles that we observe in at least two snapshots.[5] A median story in this set remains in the top-5 for 56 minutes, and almost no articles are observed in the top-5 for more than 10 hours.

Figure 2: Duration of News Articles in the Top-5 Yandex News

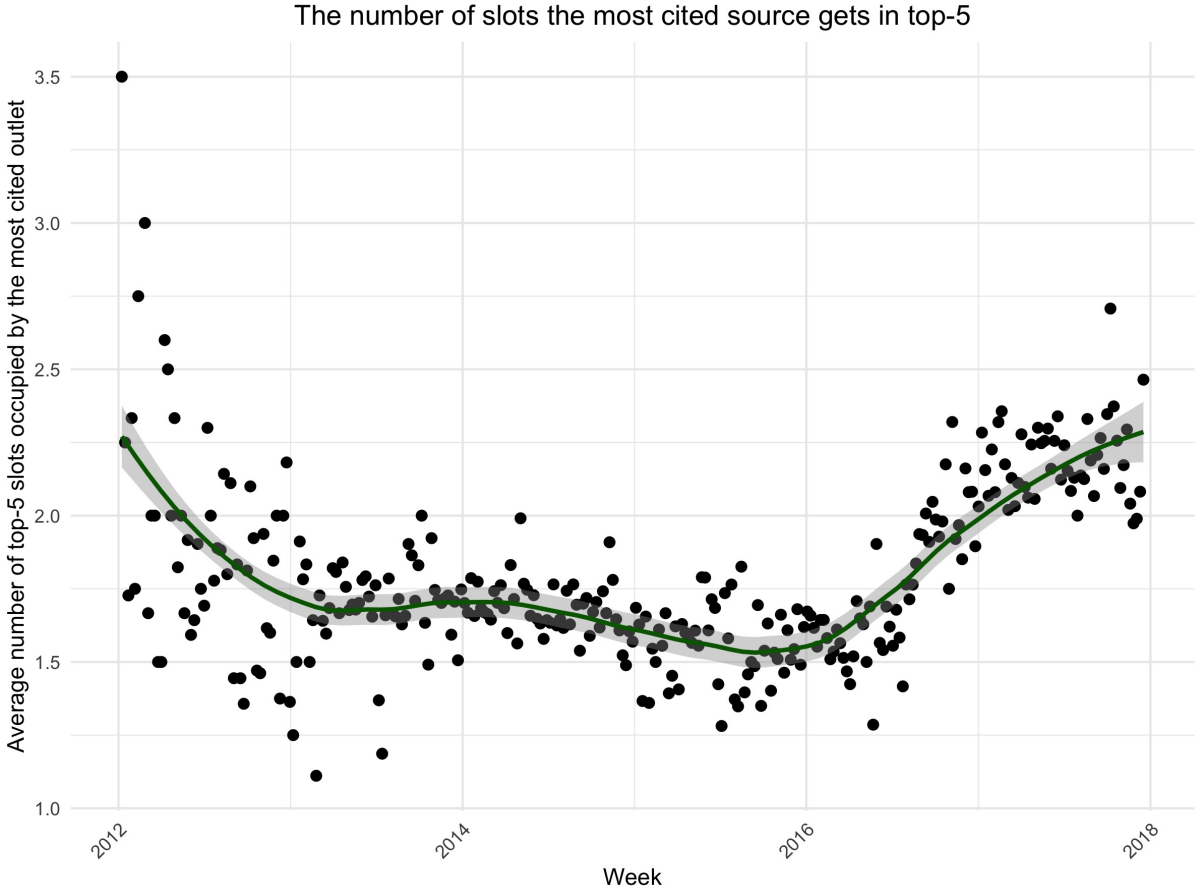


A histogram of the duration of stories in the top-5 section of Yandex. We exclude articles that we observe only on one snapshot of Yandex.

[5]This implies that an actual rotation of news articles is even faster.

Second, the same outlet can appear in the top-5 news more than once. Figure 3 presents the number of slots that the most cited outlet of the moment occupies throughout the weeks in our sample. In an average snapshot, the outlet most cited occupies 1.84 slots. However, there is a visible increase in the probability that the same outlet occupies multiple slots after mid-2016, with an average of 2.13 slots after that.

Figure 3: The Number of Top-5 Slots Occupies by the Most Cited Outlet



Dots represent a weekly average of the number of top-5 slots occupied my the most cited outlet at that moment. For instance, the value of two means that in an average snapshot two out of five slots are occupied by the same outlet (e.g. ria.ru).

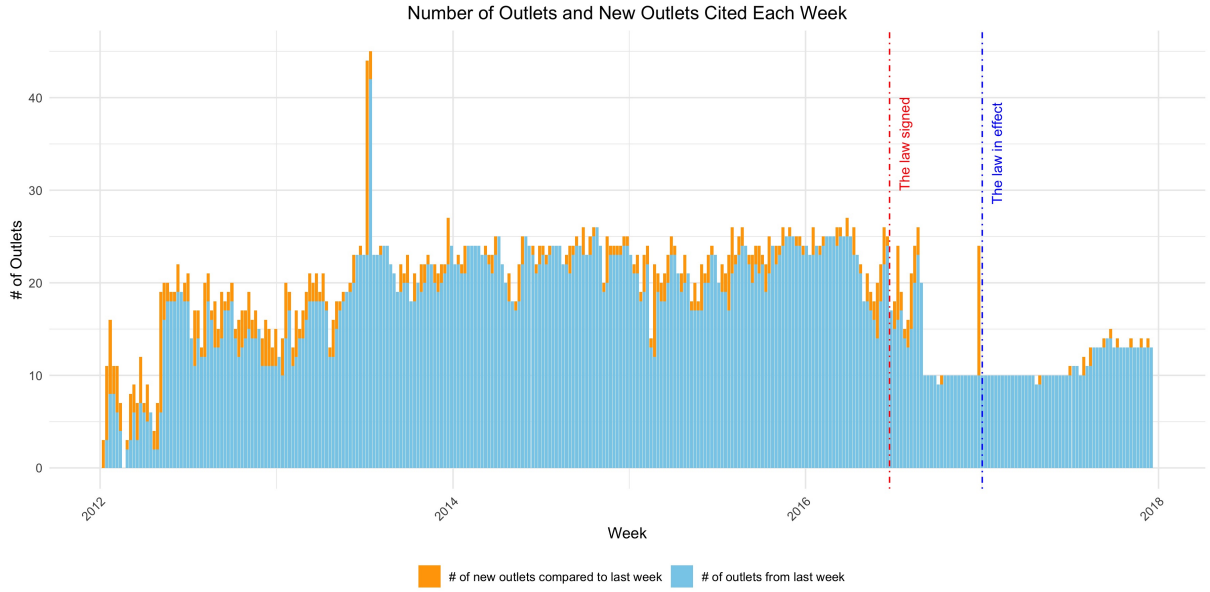
On 23 June 2016, Vladimir Putin signed a new federal law on news aggregators (Wijermars, 2021), which became effective on 1 January 2017. Before the law, news aggregators were not held liable for the information they aggregated and distributed, with the legal responsibility carried by news publishers.[6] The law changed this, making aggregators liable for spreading "fake news" unless the information came from the media registered with the government.

Shortly after the law was signed, Yandex sharply changed its citation algorithm. Figure 4

[6]This is similar to the legislation in other jurisdictions, e.g. Section 230 of the Communications Decency Act (CDA) in the United States that provides legal immunity to online platforms for distributing third-party content.

makes this change visible by plotting the number of unique outlets that Yandex cites in its top-5 slots on a given week. During a regular week of 2016 before the law was signed, around 24 different outlets appeared in Yandex's top-5 slots. In contrast, several weeks after the law was signed, the number of unique outlets cited by Yandex drops to 10 and stays around 10 for nearly a year. We see a slight increase in the number of outlets cited by Yandex by the end of 2017.

Figure 4: The Number of Unique Outlets Cited in Top-5 Slots

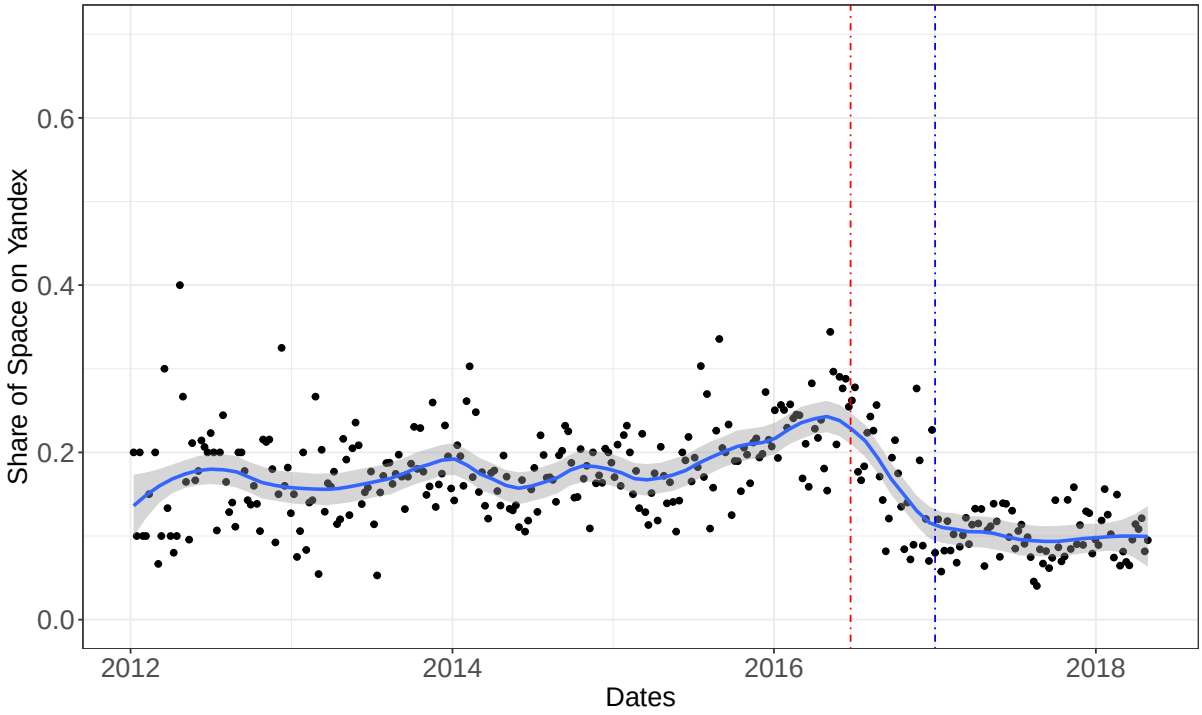


Bars represent the number of unique outlets Yandex cites in top-5 slots that week. Blue color counts the outlets that were also cited last week; yellow color counts the outlets that were not cited last week. The red vertical line corresponds to June 23, 2016, when the law on news aggregators has been signed. The blue vertical line corresponds to January 1, 2017, when the law on news aggregators came to the effect.

The algorithm change came primarily at the expense of independent outlets. Yandex stopped citing several prominent independent media that were frequently cited before the adjustment (e.g., Forbes Russia, Novaya Gazeta, TV Rain). Only two of the ten remaining outlets were independent, both with a specialization in business news (Rbc and Vedomosti). As a result, the share of independent news outlets cited by Yandex decreased from around 20% before the law – which corresponded to the overall share of news produced by independent outlets – to around 10% after the law (see Figure 5).

Although Yandex has removed most independent media outlets from its top-5 section, the composition of events cited by Yandex in the top-5 news remained the same before and after the law. Figure 6 shows the share of news events that are classified as censored by government-controlled outlets in the overall composition of the news events on that day and in the share of the events reported by Yandex. If anything, Yandex reports more censored events than an average in the media space, and there is no visible change in the reporting after the law was

Figure 5: Share of news in Yandex's top-5 from independent outlets



Each dot corresponds to an average share of independent news in Yandex's top-5 news on a given week. The red vertical line corresponds to June 23, 2016, when the law on news aggregators has been signed. The blue vertical line corresponds to January 1, 2017, when the law on news aggregators came to the effect.

signed in effect. This suggests that while Yandex has changed the outlets to cite in the top-5 slots, it did not change the underlying algorithm that scores the most important events of the moment.

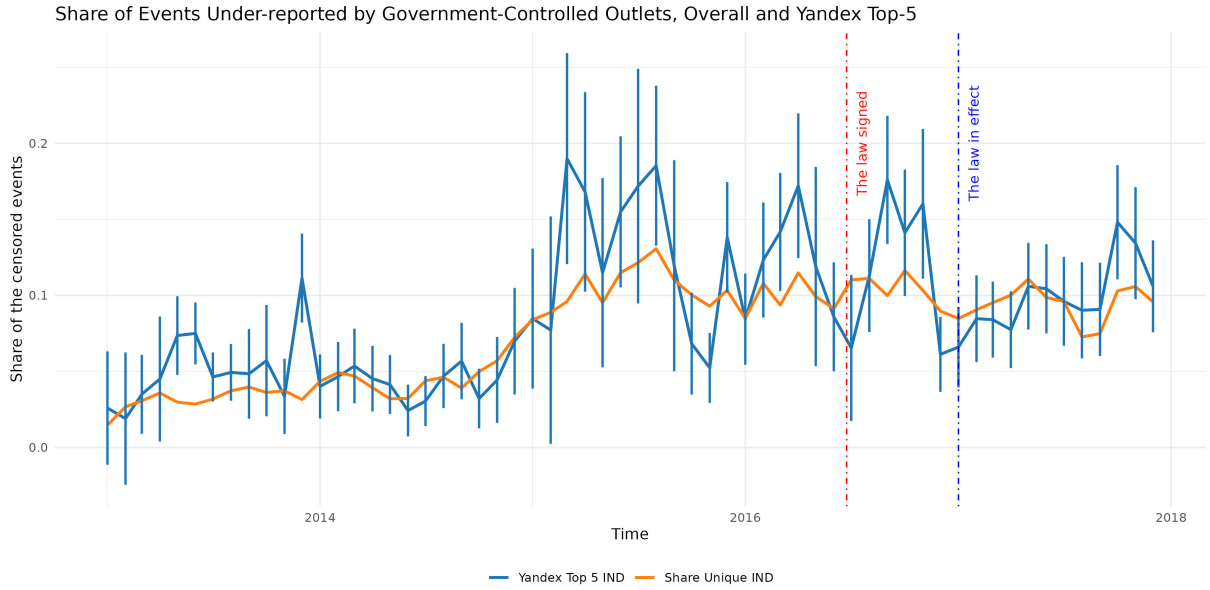
4.2 First Step: Ranking of News Events

Yandex's news aggregation and citation algorithm proceeds in two steps. In the first step, an algorithm ranks news events – inferred from clusters of arriving news articles – by their newsworthiness. The newsworthiness is measured through the characteristics of the event, such as the number of articles about the event, what outlets write about the event, the speed of the arrival of articles, and other factors.

We estimate the Yandex's algorithm using the data on news events and their citations. Every time t, there is a set of news events E_t that are realized. A news event e_t is characterized by its newsworthiness, n_e, which is not observed by the researcher. A news outlet j chooses the events to report in period t, $r_{jt} \in E_t$, given the preferences of the news outlets for the topics of the events and the newsworthiness of the events.

Every t, Yandex selects the five main news events, $Y_{1t}, \ldots, Y_{5t}$, trying to detect the events

Figure 6: Share of Event Under-Reported by Government-Controlled Outlets



Each point corresponds to a monthly share of events that are under-reported by the government controlled outlets (compared to independent outlets). Bars correspond to the 95% confidence interval, clustering standard errors on the day-hour level. We restrict the sample of events to the events covered by at least 10 outlets. The red vertical line corresponds to June 23, 2016, when the law on news aggregators has been signed. The blue vertical line corresponds to January 1, 2017, when the law on news aggregators came to the effect.

with the highest n_e. The algorithm does not observe n_e directly; instead, the top 5 news stories are inferred based on the coverage of news events by the outlets. More formally, we assume that Yandex cites news events in its top 5 section based on the inferred event scores,

$$s_{et} = X'_{ejt}\beta + \epsilon_{et} \tag{1}$$

where X_{ejt} is a vector of event characteristics that captures the volume of coverage of this event, overall and by various outlets, the length of the event coverage, and the speed with which it develops. The term ϵ_{et} is an idiosyncratic error that shifts the score of the event e at period t.

We use the realized rankings of news events on Yandex and the underlying set of events to estimate β parameters in the above scoring model. Table 3 presents the results. Events that are covered by more articles, more outlets, and have a higher fraction of news covered in the first hour (a proxy for the speed of coverage) are more likely to be cited. Interestingly, the regression fit is relatively high, with three variables explaining 28.1% of the variation in Yandex's citation decisions. In contrast, day (column 2) and news topic fixed effects (column 3) add little explanatory power.

The lack of explanatory power in day fixed effects suggests that the first stage of Yandex's algorithm does not change much over time. We confirm that there is little variation in the size of

Table 3: Factors Correlated with Event Citations by Yandex

Dependent Variable:	I(Event in Yandex Top-5)				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Constant	-0.0512***				
	(0.0008)				
# of outlets	0.0155***	0.0156***	0.0156***	0.0160***	
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	
# of articles	0.0004***	0.0004***	0.0004***	0.0004***	0.0003***
	(3.74×10^{-5})	(3.73×10^{-5})	(3.72×10^{-5})	(3.65×10^{-5})	(2.7×10^{-5})
Share of coverage in first hour	0.0207***	0.0209***	0.0197***	0.0198***	0.0201***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Under-reported by controlled outlets				-0.0789***	-0.0944***
				(0.0062)	(0.0059)
Under-reported by controlled outlets x After law				-0.0209**	0.0049
				(0.0094)	(0.0082)
Over-reported by controlled outlets				0.0531***	0.0500***
				(0.0096)	(0.0090)
Over-reported by controlled outlets x After law				-0.0182	-0.0106
				(0.0154)	(0.0141)
Fixed-effects					
Day		Yes	Yes	Yes	Yes
Topic			Yes	Yes	Yes
Outlets					Yes
Fit statistics					
Observations	1,234,213	1,234,213	1,234,213	1,234,213	1,234,213
R^2	0.28094	0.29018	0.29083	0.29361	0.32831
Within R^2		0.28296	0.28205	0.28486	0.32000

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Clustered (week) standard-errors in parentheses. News topics correspond to nine general news topics. Estimates based on the events during days with at least 10 observed events that are cited by Yandex.

residuals over time. For this, we rerun the model with fourth-order polynomials of all explanatory variables and examine squared residuals of this model month by month. In particular, there is no substantial change in the model fit before and after the law on news aggregators was signed in effect.

We also examine whether Yandex's algorithm treats news events that are under-reported by the government-controlled outlets differently than other similar events. Although Figure 6 shows that these events are reported proportionally to their volume in the media space, under-reported events may be more or less newsworthy along other dimensions compared to regular events. Columns (4) and (5) of Table 3 check if the algorithm treats news events that are under- and over-reported by government-controlled outlets differently from regular events. We find that the probability of being cited in the top-5 news of Yandex is 7.9-9.4 percentage points lower if the event is under-reported by the government-controlled media, keeping everything else constant. In contrast, events that government-controlled outlets overreport are around 5 percent more likely to be cited. There is little difference in the estimates before and after the law on news aggregators, suggesting that the change in the first stage of the algorithm was limited.

4.3 Second Step: References of News Outlets

In the second step of the algorithm, after identifying the top-5 news events, Yandex selects the outlet to reference in the top-5 section for each event.

For each event e, there is a subset of outlets J_e that cover this event, potentially with more than one article per outlet. The algorithm of Yandex picks one of the articles from this subset to cite, taking into account various characteristics of the article and outlet. Based on the public information disclosed by Yandex and our interviews with media professionals who worked at news outlets at the time, Yandex takes into account the volume of coverage of this article by an outlet, the delay of the article coverage compared to the first article that broke the news, number of unique facts in the article, properties of the title of the article (e.g. number of symbols, punctuation, and how well it reflects the event), the degree to which a news article copies the original source of the news event (with articles that copy a lot of text verbatim getting a penalty), and other related characteristics of the news coverage. Beyond these factors, Yandex can manually adjust outlet weights, increasing or decreasing the probability that these outlets appear in the top-5 news section.

To validate the factors that reportedly affect references of Yandex and test whether our data has enough statistical power to detect these correlations, we regress the outlet references of Yandex on the characteristics of these outlets coverage

$$c_{je} = \tilde{X}'_{ej}\tilde{\beta} + \gamma_e + \alpha_j + \tilde{\epsilon}_{ej} \tag{2}$$

where $\tilde{X}_{ej}$ captures various characteristics of outlet j's coverage of event e, γ_e and α_j are event and outlet fixed effects, and $\tilde{\epsilon}_{ej}$ is an idiosyncratic shock that affects references of Yandex.

Table 4 presents the estimates of $\tilde{\beta}$ from equation 2. All specifications include event fixed effects, to control for the newsworthiness of each event. The estimates confirm the description of Yandex's algorithm by industry professionals. An outlet is more likely to be cited for an event if it covers the event early, has more named entities in the story (a proxy for the number of facts),[7] has a shorter title with less punctuation, has a title closer to the core message of the event, and is sufficiently differentiated from the first publication on the topic.[8] Based on our interviews, Yandex should particularly penalize news articles that copy 40% or more text from the first article; to test for this, in Column (2) we break down the measure of text distance from the first publication into quintiles. Indeed, we see that being in the lowest quantile (highest similarity) has a strong negative effect on the probability of citation.

[7] Interestingly, beyond the number of named entities, the length of articles does not matter for citation probability.

[8] To measure the difference of the article text compared to the first publication about the news event, we assign each article a score (from 0 to 1) based on the cosine similarity of its text to the text of the original article (we weight words by their TF-IDF scores). We assign the value of 1 to the first article about the event, reflecting the originality of this article.

Table 4: Factors Correlated with Outlet References by Yandex

Dependent Variable:	I(Outlet j in Yandex Top-5)		
Model:	(1)	(2)	(3)
Variables			
Share of articles on this event by j	-0.0306	-0.0316	0.0238***
	(0.0249)	(0.0253)	(0.0083)
log(hours published after first event article)	-0.0187***	-0.0188***	-0.0049***
	(0.0056)	(0.0057)	(0.0009)
# of words in event articles by j	-8.97×10^{-6}	-9.03×10^{-6}	-5.3×10^{-6}
	(1.07×10^{-5})	(1.06×10^{-5})	(4.65×10^{-6})
# of named entities in event articles by j	0.0015**	0.0015**	0.0004***
	(0.0006)	(0.0006)	(0.0001)
Title length (symbols)	-0.0003*	-0.0003*	-0.0004***
	(0.0002)	(0.0002)	(0.0001)
# of punctuation marks in the title	-0.0032*	-0.0032*	-0.0017**
	(0.0016)	(0.0016)	(0.0007)
Title-to-event-texts distance	-0.0755***	-0.0735***	-0.0146***
	(0.0225)	(0.0224)	(0.0054)
Difference from the first publication	0.1040***		
	(0.0308)		
Difference from the first publication: 0-20%		-0.0581***	-0.0345***
		(0.0130)	(0.0066)
Difference from the first publication: 20-40%		-0.0197***	-0.0094***
		(0.0058)	(0.0019)
Difference from the first publication: 60-80%		0.0118**	0.0049***
		(0.0048)	(0.0018)
Difference from the first publication: 80-100%		0.0269**	0.0164***
		(0.0107)	(0.0051)
Fixed-effects			
Event	Yes	Yes	Yes
Outlet	No	No	Yes
Fit statistics			
Observations	1,030,204	1,030,204	1,030,204
R^2	0.06787	0.06727	0.13484
Within R^2	0.01399	0.01336	0.00225

Clustered (outlet) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Observations correspond to news articles about the events that were cited by Yandex in top-5 slots at least once.

Finally, in Column (3) of Table 4 we add outlet fixed effects to the regression. With outlet fixed effects, we observe a similar relationship between the characteristics of news coverage and

the probability of citation of a news outlet by Yandex. If anything, the relationships become more statistically significant. However, interestingly, the model fit increases substantially, with R^2 increasing from 6.7% to 13.5%. This suggests that while the characteristics of coverage are important in Yandex's citation decisions, outlet weights also play an important role.

To understand the role of the outlet weights in Yandex's citation decisions and to examine how these weights changed over time, we estimate a discrete choice model of Yandex's citation decisions. The benefit of a discrete choice model is that we can explicitly account for the fact that news outlets cover only a subset of events.

We estimate the following multinomial logit model:

$$Pr(\text{Outlet } j \text{ in Yandex Top-5}_e | j \in J_e) = \frac{exp(\alpha_j + \alpha'_j I(\text{post-law})_e + \tilde{X}'_{ej}\tilde{\beta}}{\sum_{j' \in J_e} exp(\alpha_{j'} + \alpha'_{j'} I(\text{post-law})_e + \tilde{X}'_{ej'}\tilde{\beta}} \tag{3}$$

where α_j is outlet j fixed effect, α'_j is a change in outlet j fixed effects after the law, and $\tilde{\beta}$ include the same controls as in equation 2.

Table 5: Implied Probabilities of Citation Based on Outlet Weights: By Outlet Type

	Probabilities of Citation (in %)			
Type	Equal Weights	Before the Law	After the Law	Difference
International	10	0.08	0.01	-0.08
Independent	28	29.53	16.19	-13.34
Potentially-Influenced	42	28.29	32.80	4.50
Government-controlled	20	42.09	51.00	8.91

Implied probabilities of references based on outlet weights α_j and α'_j from equation 3. The probabilities are first computed for each outlet and then summed up by outlet type. The "equal weights" column is proportional to the number of outlets of each type in the sample.

Table 5 summarizes the estimates of α_j and α'_j by presenting the implied probabilities of citation of the outlets only based on the outlet weights. To make the results more readable, we present the citation probabilities by outlet types; Table 12 in Appendix A.1 presents the citation probabilities for all outlets in our sample. The first column serves as a benchmark; these are probabilities of references if Yandex weighted outlets equally.

The second column of Table 5 captures the outlets' weights in the citation probabilities of Yandex before the law. Independent outlets have weights proportional to their share in the market, while potentially-influenced and international outlets are cited less than their share. The government-controlled outlet have higher-than-average weights, with an implied probability of citation of 42%.

The outlet weights change dramatically after the law on news aggregators is signed. Independent outlets experience the greatest change, with their probability of references decreasing from 29.53% to 16.19%. The model estimates show that this drop is due to a number of independent outlets receiving a strongly negative weight after the law and getting an implied citation

probability of zero.[9]

The estimates of outlet weights confirm the suggestive results in Figures 4 and 5: the drop in citation of independent outlets after the law on news aggregators is due to lower weights that Yandex assigns to these outlets in the second stage of their algorithm, and not due to other factors that enter the algorithm (e.g., events that these outlets choose to cover). This change in outlet weights highlights the ability of an aggregator to seamlessly change their algorithm and decide to cite a different set of outlets. We next explore the implications of such control of the algorithm.

5 Effects on News Consumption

5.1 Effects on Outlet Traffic

We now transition to measuring the implications of the ability of news aggregators to control and adjust their citation algorithm. As we saw above, Yandex can exercise a lot of discretion over which news outlets it references in the links in the top-5 section. A natural implication of such discretion is that Yandex is potentially able to refer traffic to news outlets of its choosing.

We evaluate the ability of Yandex to direct traffic to news outlets by estimating the impact of a Yandex referral on the outlet's traffic. For this, we leverage day-to-day data on the news traffic from `top.mail.ru` and estimate the impact of the outlet's references in the links in the top-5 section on its traffic

$$y_{jd} = aX(j \text{ in top-5})_d + bI(j \text{ in top-6})_d + \rho_{jw} + \delta_{j,\text{wd}} + \theta_d + \xi_{jd} \tag{4}$$

where y_{jd} is the total number of visitors (in logs or levels) to outlet j on day d and $X(j \text{ in top-5})_d$ is the share of Yandex top-5 slots occupied by j throughout the entire day d (It equals 1 if j held all top-5 slots for the whole day). All regressions include ρ_{jw}, $\delta_{j,\text{wd}}$, and θ_d, which are outlet-week, outlet-day-of-the-week, and day fixed effects, respectively. The term ξ_{jd} is an idiosyncratic shock that impacts outlet traffic.

For robustness, in some specifications we control for $I(j \text{ in top-6})_d$, an indicator of whether outlet j appears in the "top-6" slot on d. While the top-6 slot does not exist on the main page of Yandex, we can infer which news events "almost made it" to the top-5 slots by looking at the top news on Yandex News by news topic subcategories. We use the top story in each news topic subcategory as a contestant to the top-5, and we label these articles as stories in the "top-6" slot. To construct this measure, we augment out data scrapped from the front pages of Yandex with the data scrapped from `news.yandex.ru` pages, which has more sparse records on the Wayback Machine due to a lower traffic volume to this page. As a result, we have fewer observations once

[9] See Table 12 in Appendix A.1 for outlet-level results.

we control for $I(j \text{ in top-6})_d$.

The primary parameter of interest is a, a causal effect of Yandex's reference of a news outlet's links on the outlet's traffic. Our identification strategy relies on the day-to-day variation in Yandex references, driven by quasi-random realizations of the share of references an outlet gets on a given day. To isolate this variation, we include outlet-by-week, outlet-by-day-of-the-week, and day fixed effects into the regression. In particular, we rely on γ_{jw} to control for any differences in outlets' weights and adjustments of these weights over time. Below, we provide supportive evidence for this empirical strategy by showing the robustness of our estimates to alternative parametric specifications of the treatment variable. We also use referral traffic data available for a subset of outlets to show the null effects of Yandex's citation on the outlets' direct traffic and referrals from other sources (e.g. other news aggregators).

Table 6: Effect of Outlet References in Yandex's Top-5: Overall Traffic

	Dependent variable:			
	$log(y_{jd})$		y_{jd}	
	(1)	(2)	(3)	(4)
$X(j \text{ in top-5})_d$	0.362***	0.446***	252,803.200***	374,025.800***
	(0.026)	(0.036)	(26,066.930)	(38,134.070)
$I(j \text{ in top-6})_d$		0.044***		10,636.180***
		(0.006)		(3,432.030)
Fixed Effects:				
- Week x Outlet	Y	Y	Y	Y
- Week Day x Outlet	Y	Y	Y	Y
- Date	Y	Y	Y	Y
Clustered S.E.	Weeks	Weeks	Weeks	Weeks
Observations	27,952	20,084	27,952	20,084
Adjusted R^2	0.983	0.981	0.955	0.953
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6 presents the estimates of a using specifications with logs and levels of the total traffic navigating to news outlets. If an outlet occupies all slots in Yandex's top-5 positions for an entire day, its traffic grows by 56.2% (column 2), or 374 thousand visitors (column 4). In contrast, the effect of appearing in the "top-6" slot is much lower. Our preferred specification is in levels (column 4), since the magnitudes of referrals from Yandex are pinned down by the number of visitors to Yandex's front page and not by the outlet size.

To measure separate effects of Yandex's top-5 positions, we break down $X(j \text{ in top-5})_d$ into five position-specific variables. Table 7 presents the estimates. As expected, higher positions

Table 7: Effect of Outlet References in Yandex's Top-5: By Position

	Dependent variable:			
	y_{jd}			
	(1)	(2)	(3)	(4)
$X(j \text{ in position } 1)_d$	71,224.160***	95,052.020***	97,472.670***	97,255.460***
	(11,904.520)	(18,903.270)	(18,766.620)	(18,754.490)
$X(j \text{ in position } 2)_d$	73,026.640***	95,340.820***	96,855.890***	96,824.440***
	(11,397.670)	(18,845.710)	(18,752.850)	(18,789.290)
$X(j \text{ in position } 3)_d$	31,314.410***	67,196.380***	68,601.270***	68,607.890***
	(9,800.035)	(15,668.120)	(15,614.530)	(15,541.340)
$X(j \text{ in position } 4)_d$	41,081.190***	59,287.360***	61,290.650***	61,645.760***
	(9,547.727)	(15,717.560)	(15,683.990)	(15,696.940)
$X(j \text{ in position } 5)_d$	37,951.000***	57,089.880***	58,628.420***	58,439.710***
	(10,828.240)	(15,696.010)	(15,689.010)	(15,715.920)
$I(j \text{ in top-6})_d$		10,566.070***		
		(3,428.774)		
$\sum_{p=6}^{10} X(j \text{ in position } p)_d$			17,692.910*	
			(9,334.369)	
$X(j \text{ in position } 6)_d$				28,764.830
				(24,277.970)
$X(j \text{ in position } 7)_d$				9,545.342
				(22,701.370)
$X(j \text{ in position } 8)_d$				18,049.660
				(26,390.310)
$X(j \text{ in position } 9)_d$				4,978.436
				(25,461.100)
$X(j \text{ in position } 10)_d$				28,099.380
				(25,847.220)
Fixed Effects:				
- Week x Outlet	Y	Y	Y	Y
- Week Day x Outlet	Y	Y	Y	Y
- Date	Y	Y	Y	Y
Clustered S.E.	Weeks	Weeks	Weeks	Weeks
Observations	27,952	20,084	20,084	20,084
Adjusted R^2	0.956	0.953	0.953	0.953
Note:				*p<0.1; **p<0.05; ***p<0.01

bring more traffic to an outlet than lower positions. For instance, occupying position 1 for an entire day brings an outlet 97 thousand visitors, whereas occupying position 5 for an entire day brings only 58.4 thousand visitors. If an outlet occupies all slots in Yandex's top-5 positions for an entire day, its traffic grows by nearly 383 thousand visitors (using estimates from column 4). Estimates are robust across specifications with alternative controls for articles in the "top-6" – "top-10" positions.[10]

Table 8: Effect of Outlet References in Yandex's Top-5: By Type of Visitors

	Dependent variable:				
	y_{jd}	y_{jd}^{New}	y_{jd}^{Regular}	$y_{jd}^{\text{Pageviews}}$	$y_{jd}^{\text{Av. Pageviews}}$
	(1)	(2)	(3)	(4)	(5)
$X(j \text{ in top-5})_d$	374,025.800***	157,061.000***	3,405.561	681,298.800***	-0.756***
	(38,134.070)	(15,415.500)	(2,786.359)	(88,099.370)	(0.085)
$I(j \text{ in top-6})_d$	10,636.180***	7,402.519***	-212.421	15,513.410	-0.065***
	(3,432.030)	(1,727.682)	(229.638)	(10,237.020)	(0.012)
Fixed Effects:					
- Week x Outlet	Y	Y	Y	Y	Y
- Week Day x Outlet	Y	Y	Y	Y	Y
- Date	Y	Y	Y	Y	Y
Clustered S.E.	Weeks	Weeks	Weeks	Weeks	Weeks
Observations	20,084	18,655	18,655	18,655	18,655
Adjusted R^2	0.953	0.881	0.998	0.952	0.963

*p<0.1; **p<0.05; ***p<0.01

New visitors are defined as visitors who came to this site for the first time in 91 days. Regular visitors are defined as visitors who visited this site for at least 7 days in the last 28 days.

We check whether the traffic referred by Yandex is driven by new outlet visitors – who are more likely to be incremental – or returning visitors. For this, we re-estimate the regression 4 with the number of "new" and "regular" visitors to an outlet. The tracker `top.mail.ru` defines new visitors as those who came to the site for the first time in three months; regular visitors are defined as those who visited the site for at least 7 days during the last 28 days.

Columns (2)-(3) in Table 8 present the estimates for new and regular visitors. A full-day reference in all of the top-5 links of Yandex brings in 157 thousand new visitors, but no regular visitors. This supports the result that the referral traffic from Yandex's top-5 is incremental.

In Columns (4)-(5) of Table 8, we further examine the effect of Yandex's outlet references on the number of page views that an outlet gets. The count of page views is important for the

[10]We construct articles in fictional "top-7" ... "top-10" positions in the same way as we did for "top-6": we take articles from the ranking of Yandex News by topic subcategories. For instance, to construct articles in the "top-7" slot, we take news articles in the second slots in Yandex's rankings for news topic subsections; to construct articles in the "top-8" slot, we take news articles in the third slots in Yandex's rankings for news topic subsections; etc.

outlets, as none of the outlets used a subscription model at the time and instead relied on display advertising for monetization.[11] The 374 thousand new visitors bring in 681 thousand more page views (column 4). Interestingly, we see that the daily average number of page views per visitor on the oultet falls when there is more traffic from Yandex (column 5). Taking into account estimates from Tables 6 and 8, the back-of-the-envelope calculation shows that marginal visitors have much fewer page views (1.8) compared to inframarginal visitors (around 4 pageviews for an average website).[12]

Table 9: Effect of Outlet References in Yandex's Top-5: By Source

	Dependent variable:			
	$log(y_{jd})$	$log(y^{Y}_{jd})$	$log(y^{-Y}_{jd})$	$log(y^{\text{ext}-Y}_{jd})$
	(1)	(2)	(3)	(4)
$X(j \text{ in top-5})_d$	0.685^{***}	1.156^{***}	0.081	0.067
	(0.104)	(0.182)	(0.094)	(0.091)
$I(j \text{ in top-6})_d$	0.063^{***}	0.162^{***}	0.018^{**}	0.020^{**}
	(0.010)	(0.020)	(0.009)	(0.008)
Fixed Effects:				
- Week x Outlet	Y	Y	Y	Y
- Week Day x Outlet	Y	Y	Y	Y
- Date	Y	Y	Y	Y
Clustered S.E.	Weeks	Weeks	Weeks	Weeks
Observations	5,716	5,716	5,716	5,716
Adjusted R^2	0.976	0.955	0.978	0.974
Note:				*p<0.1; **p<0.05; ***p<0.01

Finally, we use a subset of four outlets for which we observe referral traffic to check whether Yandex references have an effect on direct traffic or traffic from other sources. For this, we re-estimate the regression 4 with only the traffic referred by Yandex (y^{Y}_{jd}), all traffic net of Yandex referrals (y^{-Y}_{jd}), and all external traffic net of Yandex referrals ($y^{\text{ext}-Y}_{jd}$). We use specifications with traffic in logs to improve the precision of our estimates (since we are relying on a subset of outlets).

Table 9 presents the estimates. First, in Column (1), we replicate the a estimate based on

[11] The only exception is a small indepedendent media `slon.ru`, which introduced subscriptions in mid-2014. Later in 2016, it changed its name to `republic.ru`.

[12] The pageviews-per-visitor for marginal consumers is computed by dividing the incremental number of pages views (681 thousand) on the incremental number of visitors (374 thousand). The pageviews-per-visitor for inframarginal consumers is backed out using the estimates from Table 8 and the result from Table 6, which showed that that 374 thousand consumers corresponds to 56.2% of visitors for an average outlet.

a full sample of our websites (Column 2 in Table 6). While the estimate is less precise – it has a standard error of 0.104 instead of 0.036 – the magnitudes of the coefficient estimates are comparable, confirming that Yandex affects the traffic of this subset of four websites in the same way as it affects a broader set of outlets.

In Columns (2)-(4) of Table 9, we split the total traffic to the outlets by its referral source. A full-day reference in some of the top-5 links increases the traffic coming to outlets from Yandex by 217.7% (column 2). In contrast, we see no effect of references in the Yandex's top-5 section on other traffic to outlets, both if we include (column 3) or exclude (column 4) direct traffic. Once again, the estimates confirm that Yandex brings in incremental news readers to outlets.

We further show that the incremental visitors coming to outlets because of the Yandex's references are not likely to become loyal readers and return to the same outlets later. For this, we re-run the regression 4 using daily total traffic to outlets from days $d+1, d+2, ...d+14$ instead of d. Figure 10 in Appendix A.1 visualizes the estimates. On most of the $d+1, ..., d+14$ days, we can reject an increase in the number of visitors of 100 thousand, compared to the 380 thousand effect on day d. We also do not observe any trend in the estimates.

5.2 Effectiveness of Yandex's Top-5 References

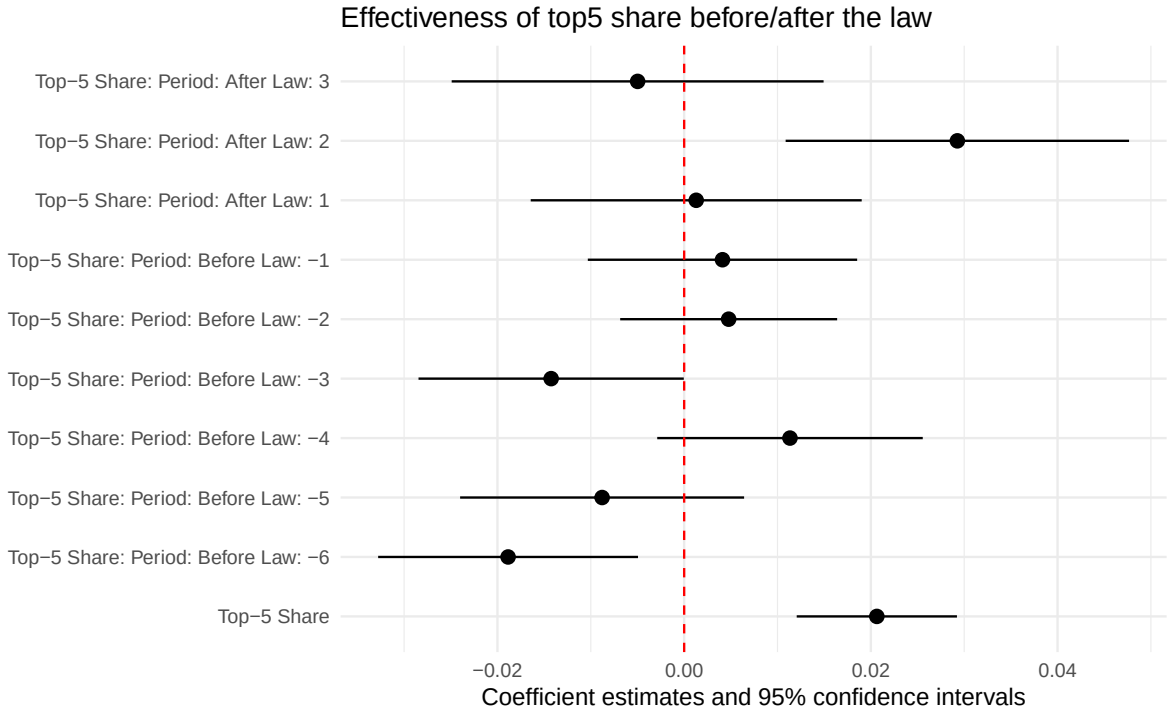
In the above estimates, we examine the volume of traffic that Yandex is able to direct to news outlets given their current algorithm. However, we are interested in assessing the ability of Yandex to preference particular outlets by changing its algorithms, both through changes in the events that Yandex cites and the outlets that it references in the top-5 links.

To assess the ability of Yandex to direct traffic in such counterfactual scenarios, we estimate and compare its effectiveness in directing traffic to news outlets for different sets of outlets, events, and groups of consumers. We measure the effectiveness of Yandex by normalizing the volume of traffic referred to outlet j on day d, y_{jd}, by the total number of visitors to Yandex that day (y_{jd}^{Yandex}). This normalization is important to control for trends and other changes in the volume of traffic visiting the Yandex front page. We then use these normalized traffic variables as an outcome variable in regression 4.

Effectiveness of Time First, we evaluate whether the effectiveness of Yandex has changed before and after the law on news aggregators. In response to the law, Yandex has adjusted the second stage of its algorithm, reducing the number of outlets it cited from around 24 to 10 (and later adjusting the number to 15). We are interested whether the effectiveness of top-5 references has decreased because Yandex restricted the outlet sample.

Figure 7 presents the estimates of Yandex's effectiveness in referring traffic based on semi-annual periods before and after the law on news aggregators. Overall, a full-day reference of an outlet in all top-5 links allows Yandex to send around 2% of its front page visitors to the outlet

Figure 7: Effectiveness of Outlet References in Yandex's Top-5: Before and After the Law



Points correspond to the a estimates from regression (1) with y_{id} normalized by the volume of traffic arriving on `yandex.ru` on day d. Error bars correspond to the 95% confidence intervals.

(the last estimate). There is no systematic decrease in the ability of Yandex to direct traffic to news outlets after they had to change their algorithm in response to the law; if anything, the effectiveness of Yandex is slightly higher in the second 6-month period after the law was signed. This suggests that reducing the number of outlets that Yandex references in the top-5 links did not decrease its ability to direct traffic to news outlets.

Effectiveness by News Types We check whether the effectiveness of Yandex is different for news events that fall under different news topics. For this, we estimate a structural topic model (Roberts et al., 2013) based on article texts and detect 100 different topics. We grouped these topics into 9 broader categories which were used by Yandex.News. Interacting $X(j \text{ in top-5})_d$ with the category of the referenced article, we find that Yandex's ability to direct traffic to outlets is relatively similar across topics: the only estimate significant at 5% level is the interaction of $X(j \text{ in top-5})_d$ and the topic of Sports (Yandex is less effective in sending traffic to outlets through news stories about sports). Table 13 in Appendix A.1 presents the result. It suggests that the ability of Yandex to refer traffic to news outlets is not limited by the nature of the events that they cite.

Similarly, Yandex is equally effective in directing traffic to news that are and are not censored and sensitive for the government. To detect such news, we apply the classification from

Simonov and Rao (2022) and find topics that were significantly under- and over-reported by the government-controlled outlets, compared to the independent media. We apply this procedure to the events that had enough – at least 10 – outlets writing about them. A typical news story under-reported by government-controlled outlets correspond to events related to political opposition and government corruption (for more details, see Simonov and Rao, 2022). Table 14 in Appendix A.1 shows that there are no significant differences in Yandex's ability to direct its front page visitors to both stories that are under- and over-reported by government-controlled media, implying that Yandex is able to increase or decrease citizen's exposure to censored news events.

While visitors of Yandex are seemly indifferent between various news topics and news types, it is not because they do not have a preference over particular news stories. To show this, we re-estimate the effectiveness of Yandex in directing traffic to news outlets on the level of traffic from 83 regions ("federal subjects") of Russia. We stack the observations from all regions and interact all control variables from regression (4) with region fixed effects (e.g. getting week by outlet by region fixed effects). We further classify news titles in Yandex's top-5 news by whether they mention news about the region of the Yandex's visitor (e.g. a hockey game in Saint-Petersburg when the visitor is from Saint-Petersburg) and check if visitors have a taste for their local news. Table 15 in Appendix A.1 shows that consumers have very strong tastes for news about their region – the share of visitors navigating to those news stories increases from 2.4 percentage points for a typical news story to 18 percentage points, a 650% increase. However, the taste for censored political news is still estimated as a relatively precise null (-1 percentage point with a 1 percentage point standard error).

We further show that the newsworthiness of articles matters for the share of visitors that navigate to news outlets. For this, we classify news titles by their predicted probability to be cited by Yandex in top-5 news from regression (1). We then check if articles with a higher predicted probability generate more traffic to news outlets. Table 16 in Appendix A.1 presents the estimates, splitting news events into the highest and lowest quartiles by newsworthiness. News events in the lowest newsworthiness quartile generated around 50% (1.4 instead of 2.8 percentage points) less traffic then the rest of the articles. This indicates the value of correct newsworthiness prediction for Yandex, which relies on news production of upstream outlets to correctly identify newsworthy events.

Effectiveness by Visitor Characteristics We further explore the heterogeneity in Yandex's ability to direct traffic for different types of customers. First, we measure the effectiveness of Yandex for news readers of different operation systems. Table 17 in Appendix A.1 present the results. We see that there are no statistically significant differences across operating systems (column 1) or device types (column 2).

Second, we compare the effectiveness of Yandex differs for news readers of different demo-

graphics. For this, we re-estimate the effectiveness of Yandex's references on traffic split by the age and gender of visitors. Tables 18 and 19 in Appendix A.1 present the results. We find that Yandex's references are around two times more effective in directing traffic to news outlets for older visitors (e.g. the effectiveness of 0.034 for visitors older than 50) than for younger visitors (e.g. 0.014 for visitors between 19 and 24). Yandex is also more effective in directing visitors identified by the tracker as male (0.029) than female (0.018) to news outlets.

5.3 Counterfactual References of Yandex

Having confirmed that the ability of Yandex to direct traffic to news outlets is not restricted by which outlets it references and which topics it cites, we now turn to evaluating the impact of Yandex's counterfactual algorithms on the market shares of news outlets.

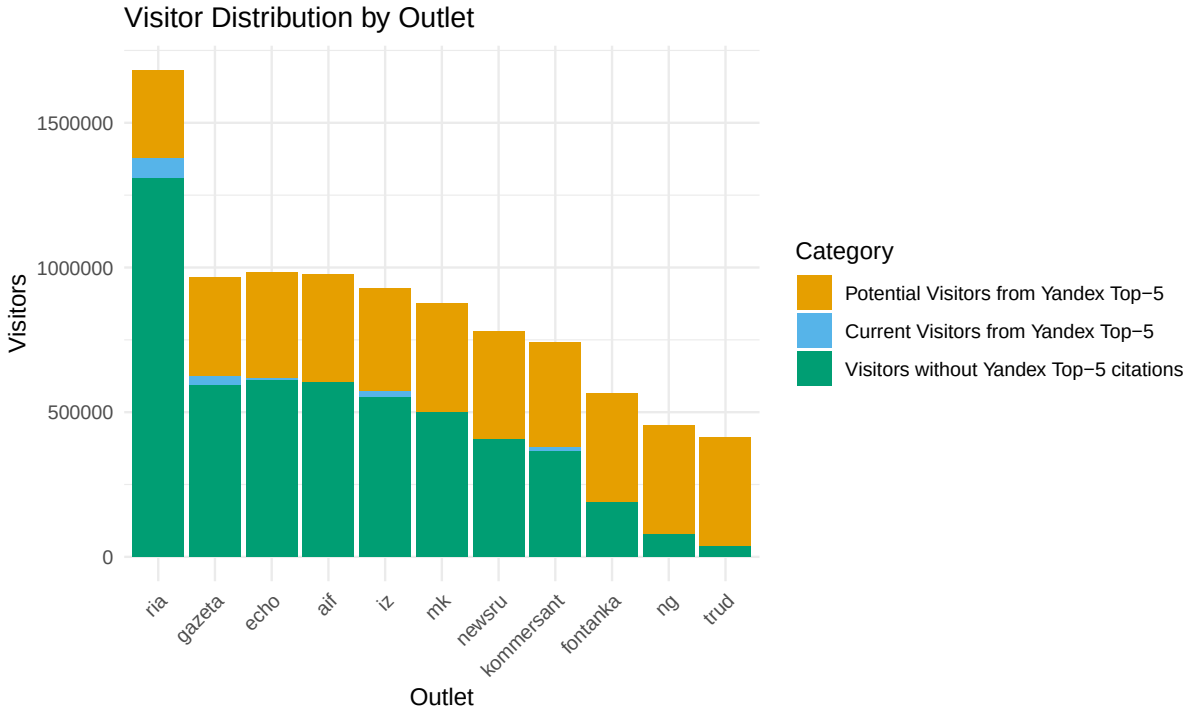
First, we consider a scenario where Yandex is not restricted either by which events it needs to cite or by which outlets it needs to reference. As a result, Yandex has an unrestricted ability to preference a particular news outlet in its top-5 section.

To measure the ability of Yandex to direct traffic to particular outlets, we simulate scenarios where Yandex fully preferences one of the news outlets in our sample, referencing only them in all top-5 slots. Figure 8 presents the number of visitors to 11 outlets for which we have traffic data, separating traffic into (i) the current traffic that an outlet receives without Yandex references, (ii) the current traffic due to Yandex references, and (iii) the counterfactual traffic that an outlet would receive if Yandex would reference it all the time. Under the current algorithm, news outlets get from close to zero (e.g. fontanka.ru) to 5.36% (gazeta.ru) of their traffic due to references in Yandex's top-5 slots. In contrast, without any restrictions on which topics to cite, Yandex would be able to increase the market shares of outlets from 23% (for ria.ru, the largest outlet in this sample) to 989% (for trud.ru, the smallest outlet).

The scenario above assumes that Yandex has a full discrepancy over which news events to cite. The second scenario we consider makes an opposite assumption: that Yandex is fully committed to the first stage of the algorithm and cannot make adjustments to the cited news events. In this scenario, the power of Yandex to direct traffic to news outlets is limited by whether they have covered the event that Yandex needs to cite due to the event's importance.

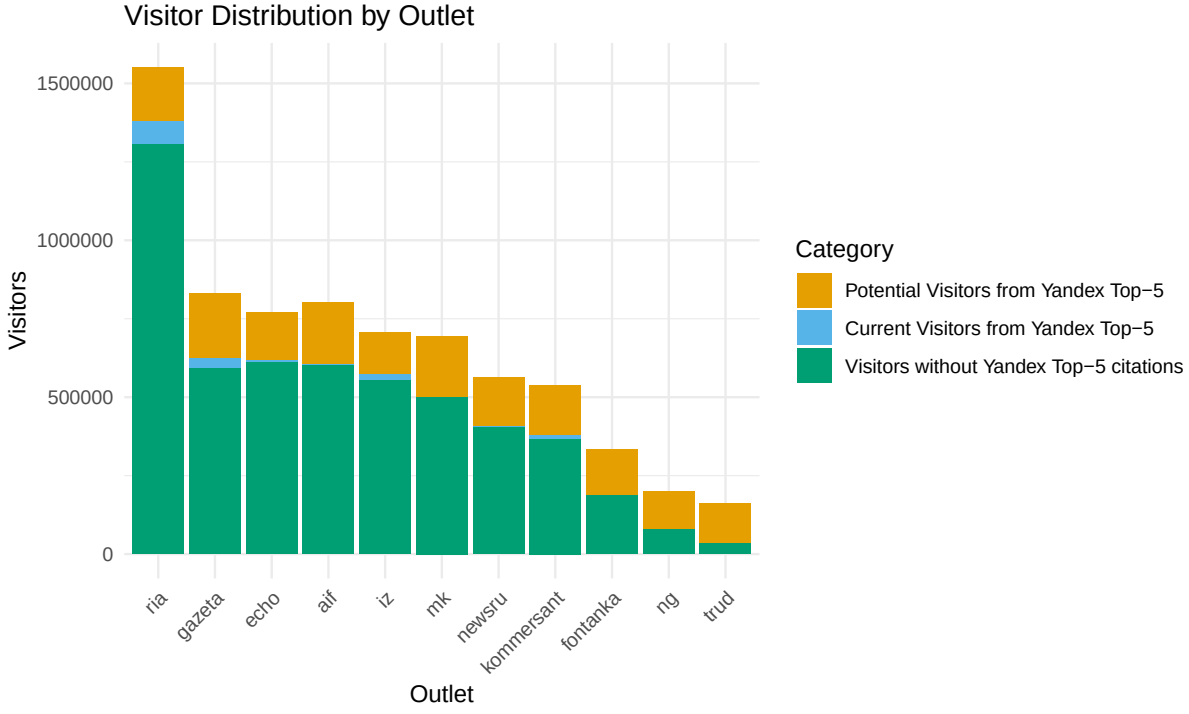
To assess the scenarios where Yandex fully commits to the first step of the news aggregation algorithm, we re-simulate the traffic outcomes with Yandex fully preferencing news outlets in the second step but not being able to reference these outlets if they did not cover the cited news event. Figure 9 presents the number of visitors to 11 outlets in our sample. Compared to the unrestricted case, the ability of Yandex to direct traffic to news outlets decreases. Yandex can increase the market share of the outlets from 13% (ria.ru) to 330% (trud.ru). The power of Yandex to direct traffic is more restricted for more specialized outlets (e.g. `echo.msk.ru`) since such outlets are less likely to cover the events cited in the top-5 news section.

Figure 8: Counterfactual Traffic to News Outlets: Yandex is Unrestricted



Visitors are predicted using the estimates of regression 4 with the traffic level as an outcome variable.

Figure 9: Counterfactual Traffic to News Outlets: Yandex Cites the Same Events as Now



Visitors are predicted using the estimates of regression 4 with the traffic level as an outcome variable.

6 Effects on News Production

Given the volume of traffic that outlets get from Yandex references, it is reasonable to expect that they adjust their production to increase their chances of getting references in the top-5 section. Anecdotally, we confirm this in our interviews with media professionals who worked in news outlets at the time of our study. The interviewees correctly argued that Yandex takes into account the structure of the title, the number of unique facts in the article, and the degree to which a news article copies the original source of the news event, and explain how they have adjusted their news to increase the changes of getting referenced by Yandex.

Such attention of news outlets to Yandex's algorithm and the traffic they receive implies that news outlets can change their production decisions in response to a change in the algorithm. We test for such effects empirically by leveraging the change in Yandex's algorithm that happened in August 2016 in response to the law on news aggregators. As we show in Section 4, Yandex has fully stopped referencing 9 out of 24 outlets after the change, including some of the outlets that were frequently referenced in the top-5 section before the adjustment (e.g. Forbes, Novaya Gazeta). To measure the effect of the algorithm's change, we compare the news production decisions of the removed news outlets to the rest of news outlets.

Figures 11 and 12 in Appendix A.1 present the event study plots with estimates of the effect of Yandex's algorithm change on news production. We examine both consequential decisions of outlets – e.g., whether to produce shorter "spot news" that have a chance to be cited by aggregators or differentiate towards "feature" news – and inconsequential decisions (e.g., the length of articles' titles). Compared to the outlets still cited by Yandex, the outlets that were downweighted by the algorithm were 30% less likely to write news articles classified into events and write 50% longer news articles (Figure 11), suggesting that these outlets differentiated towards longer "feature" news stories. The removed outlets also stopped optimizing their titles, with the length of the titles increasing by 7-8 symbols and having 0.5-1 more punctuation symbols (Figure 12). It took around 10 months for the news outlets to adjust their production decisions to a new level.[13]

As a result, we see some evidence of the upstream outlets' adjustments of news production in response to Yandex's change of the algorithm. This change is interesting since it affects all the readers of the adjusting news outlets. This implies that Yandex's algorithm changes can affect not only marginal consumers to visit outlets from Yandex, but also inframarginal consumers of news outlets, who now get different news coverage from their preferred media.

[13]We note that one of the outlets removed by Yandex, `forbes.ru`, also changed its owner around a year before the algorithm change, in September 2015. This new owner had a preference for a new editorial strategy that focused less on political news (e.g. `https://www.rbc.ru/interview/technology_and_media/16/10/2015/561fd53e9a794751f652b1d5`), which Forbes eventually adopted. Thus, for Forbes, it is hard to disentangle the extent to which the new editorial strategy was due to lower references by Yandex or the preferences of a new owner (or both factors).

7 Conclusion

In this paper, we use a case study of Yandex News in Russia to show that news aggregators hold a lot of market power through the control of their algorithm. The ability to adjust which news events they cite and which news outlets they reference allows aggregators to substantially change the market shares of news outlets. This ability is enabled by a limited response of consumers to changes in the algorithm, allowing Yandex to direct traffic through links on different events and to different outlets.

Beyond the direct effect on incoming traffic, news aggregators can impact news outlet production decisions, as news outlets optimize their news production to increase the chance of being referenced by an aggregator. As a result, the design of news aggregator algorithms affects not only marginal consumers from aggregators, but also inframarginal readers of the outlets who are now served a different news product.

There are several extensions to the current analysis that are of interest. The first is the degree and implications of competition between news aggregators; while Yandex News is the largest news aggregator in Russia in terms of referral traffic to outlets, it competes with Google News, Rambler and Mail.ru in the market. It is unclear to what extent news aggregators differentiate and adjust their algorithms in response to each other.

The second extension is the ability of news aggregators to expose readers to new factual information. In the context of Russia, one of the most important dimensions of product differentiation in the online news market at that time was whether outlets censor news topics that are sensitive to the Russian government (Simonov and Rao, 2022). In this context, a news aggregator has the ability to increase the exposure of news readers to censored information. In a more general context with strong ideological polarization in news consumption, a news aggregator can decrease the ideological segregation by referencing outlets from both sides of the political spectrum. And, on the flip side, the preferencing of outlets with a particular political position can potentially influence the beliefs of the public, providing an example of how a news aggregator can exercise media power.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., et al. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, volume 1998, pages 194–218. Citeseer.

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., and Lalmas, M. (2020). Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the web conference 2020*, pages 2155–2165.

Aridor, G., Gonçalves, D., Kluver, D., Kong, R., and Konstan, J. (2022). The economics of recommender systems: Evidence from a field experiment on movielens. *arXiv preprint arXiv:2211.14219*.

Athey, S., Mobius, M., and Pal, J. (2021). The impact of aggregators on internet news consumption. Technical report, National Bureau of Economic Research.

Boxell, L. and Conway, J. (2022). Journalist ideology and the production of news: Evidence from movers. *Available at SSRN 4295587*.

Cagé, J., Hervé, N., and Mazoyer, B. (2022). Social media and newsroom production decisions.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The production of information in an online world. *The Review of economic studies*, 87(5):2126–2164.

Calzada, J., Duch-Brown, N., and Gil, R. (2023). Do search engines increase concentration in media markets?

Calzada, J. and Gil, R. (2020). What do news aggregators do? evidence from google news in spain and germany. *Marketing Science*, 39(1):134–167.

Chiou, L. and Tucker, C. (2017). Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy*, 26(4):782–805.

Dellarocas, C., Katona, Z., and Rand, W. (2013). Media, aggregators, and the link economy: Strategic hyperlink formation in content networks. *Management science*, 59(10):2360–2379.

digiday.com (2022). News aggregators aren't growing like they used to, but publishers still see their value. `https://digiday.com/media/news-aggregators-arent-growing-like-they-used-to-but-publishers-still-see-their-value` [Acccessed: March 14, 2023].

Dujeancourt, E. and Garz, M. (2023). The effects of algorithmic content selection on user engagement with news on twitter. *The Information Society*, 39(5):263–281.

Enikolopov, R. and Petrova, M. (2015). Media capture: empirical evidence. In *Handbook of media Economics*, volume 1, pages 687–700. Elsevier.

Federal Trade Commission (2010). Potential policy recommendations to support the reinvention of journalism. *Washington, DC, Federal Trade Commission*, pages 1–47.

Garz, M. and Szucs, F. (2023). Algorithmic selection and supply of political news on facebook. *Information Economics and Policy*, 62:101020.

George, L. M. and Hogendorn, C. (2020). Local news online: Aggregators, geo-targeting and the market for local news. *The Journal of Industrial Economics*, 68(4):780–818.

Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., and Aral, S. (2020). The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 75–76.

Jeon, D.-S. and Nasr, N. (2016). News aggregators and competition among newspapers on the internet. *American Economic Journal: Microeconomics*, 8(4):91–114.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Moehring, A. (2024). Personalization, engagement, and content quality on social media: An evaluation of reddit's news feed.

Peukert, C., Sen, A., and Claussen, J. (2023). The editor and the algorithm: Recommendation technology in online news. *Management science*.

Prat, A. (2018). Media power. *Journal of Political Economy*, 126(4):1747–1783.

Prat, A. and Valletti, T. (2022). Attention oligopoly. *American Economic Journal: Microeconomics*, 14(3):530–57.

Reimers, I. and Waldfogel, J. (2023). A framework for detection, measurement, and welfare analysis of platform bias. Technical report, National Bureau of Economic Research.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Sen, A., Grad, T., Ferreira, P., and Claussen, J. (2023). (how) does user-generated content impact content generated by professionals? evidence from local news. *Management Science*.

Sen, A. and Yildirim, P. (2016). Clicks bias in editorial decisions: How does popularity shape online news coverage? *Working paper*.

Simonov, A. and Rao, J. (2022). Demand for online news under government control: Evidence from russia. *Journal of Political Economy*, 130(2):259–309.

Sismeiro, C. and Mahmood, A. (2018). Competitive vs. complementary effects in online social networks and news consumption: A natural experiment. *Management Science*, 64(11):5014–5037.

Statista (2023). From which search engines do you usually read news, information reports? https://www.statista.com/statistics/1102121/russia-most-popular-search-engines-for-news/ [Acccessed: March 14, 2023].

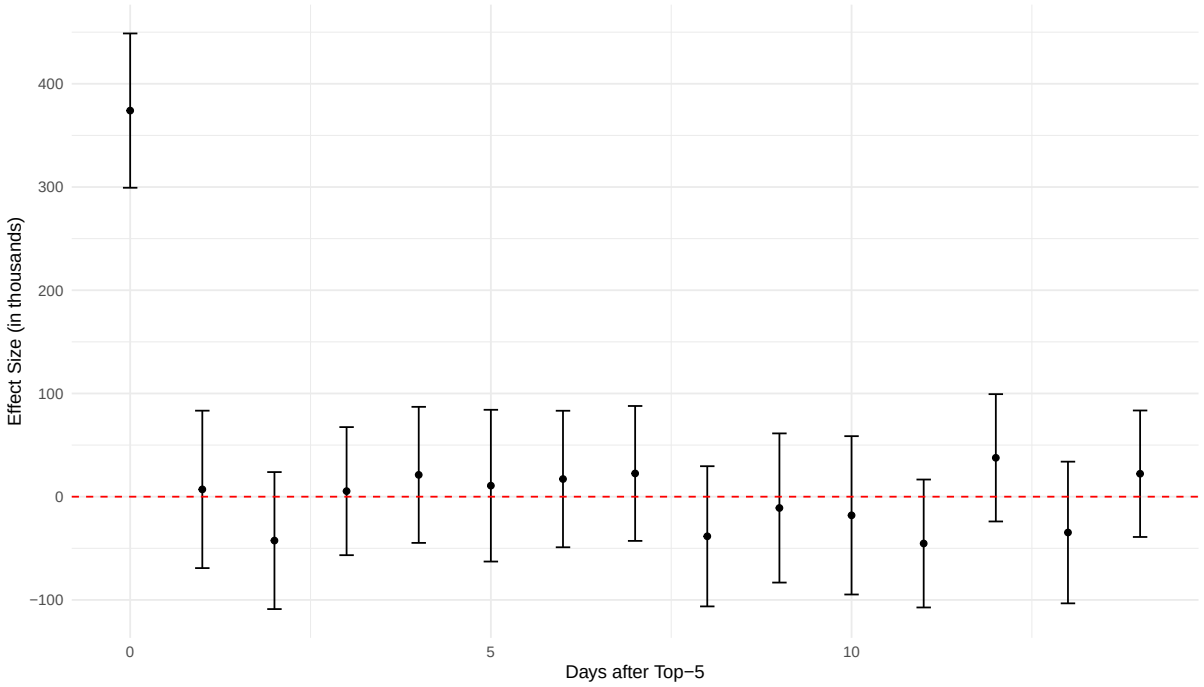
Wijermars, M. (2021). Russia's law 'on news aggregators': Control the news feed, control the news? *Journalism*, 22(12):2938–2954.

Yandex (2022). Yandex will have a new main page. what will change [in russian]. https://yandex.ru/blog/company/u-yandeksa-budet-novaya-glavnaya-chto-izmenitsya [Accessed: March 16, 2023].

A Appendix

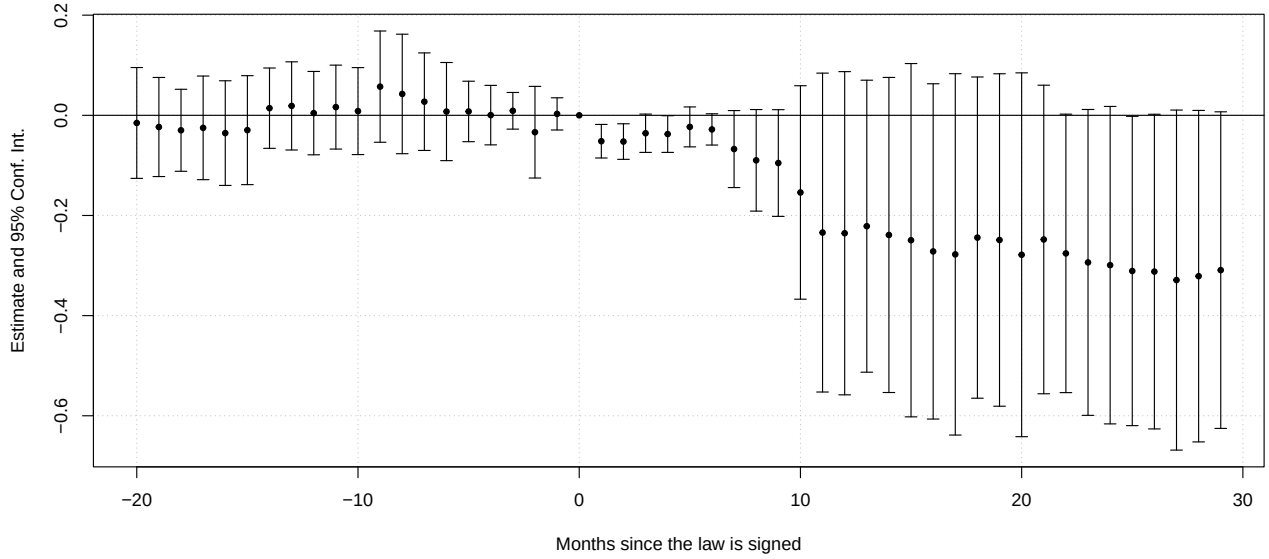
A.1 Additional Tables and Figures

Figure 10: Effect of Yandex's Top-5 Citations: Next Two Weeks

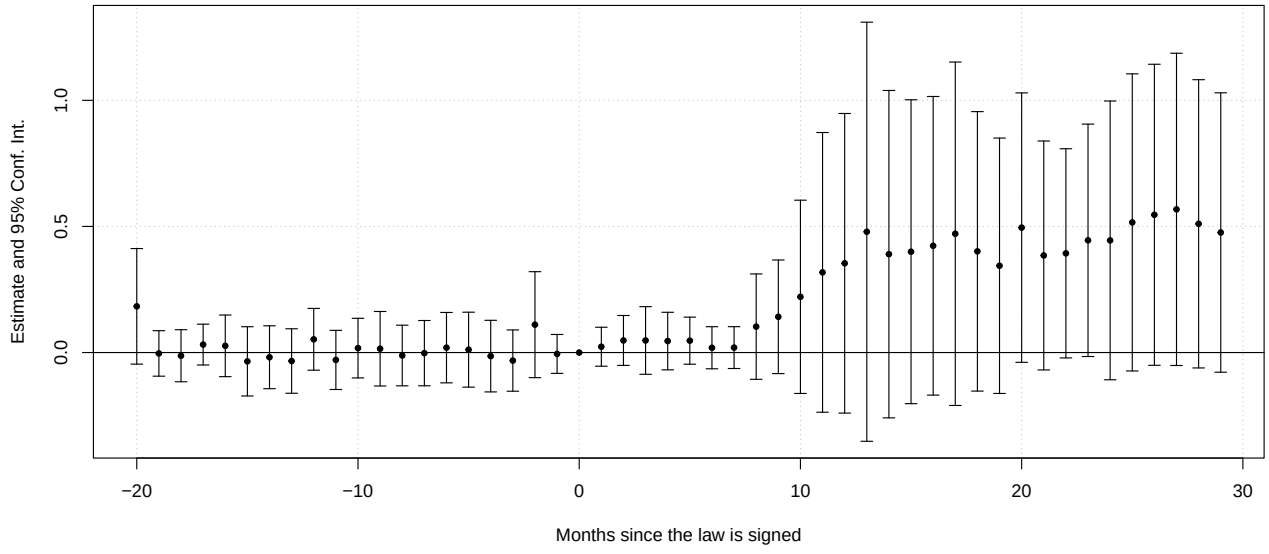


Points correspond to a estimates from regression (1) with $y_{id'}$, $d' \in \{d, d+1, ..., d+14\}$ as an outcome variable. Error bars correspond to the 95% confidence intervals.

Figure 11: Impact of Yandex's Algorithm Change on News Production: Style of News



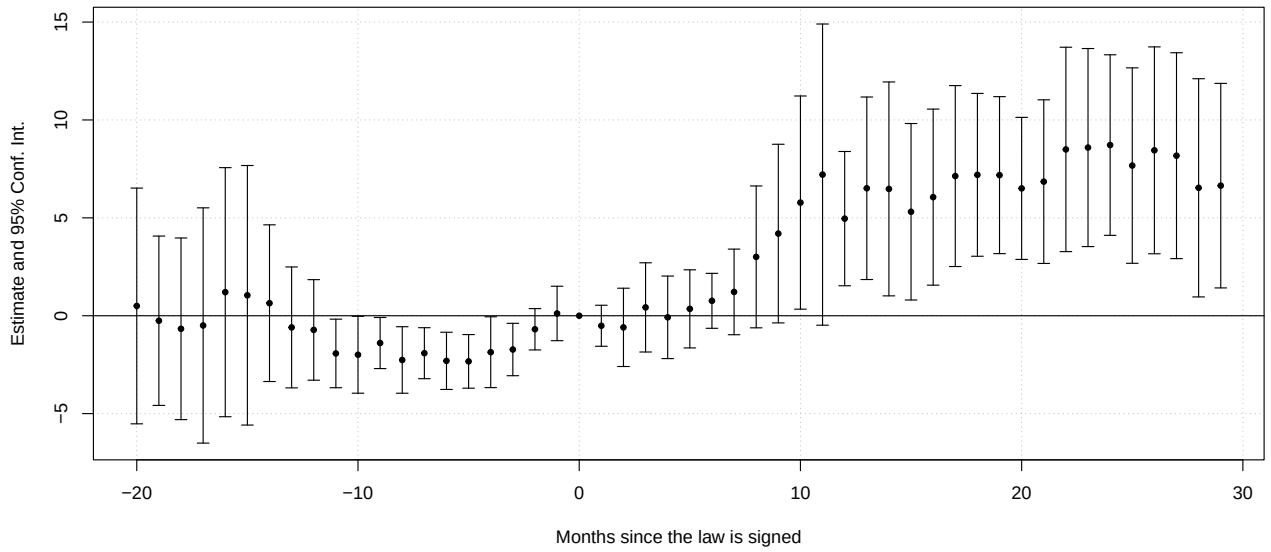
(a) Share of News Classified into Events



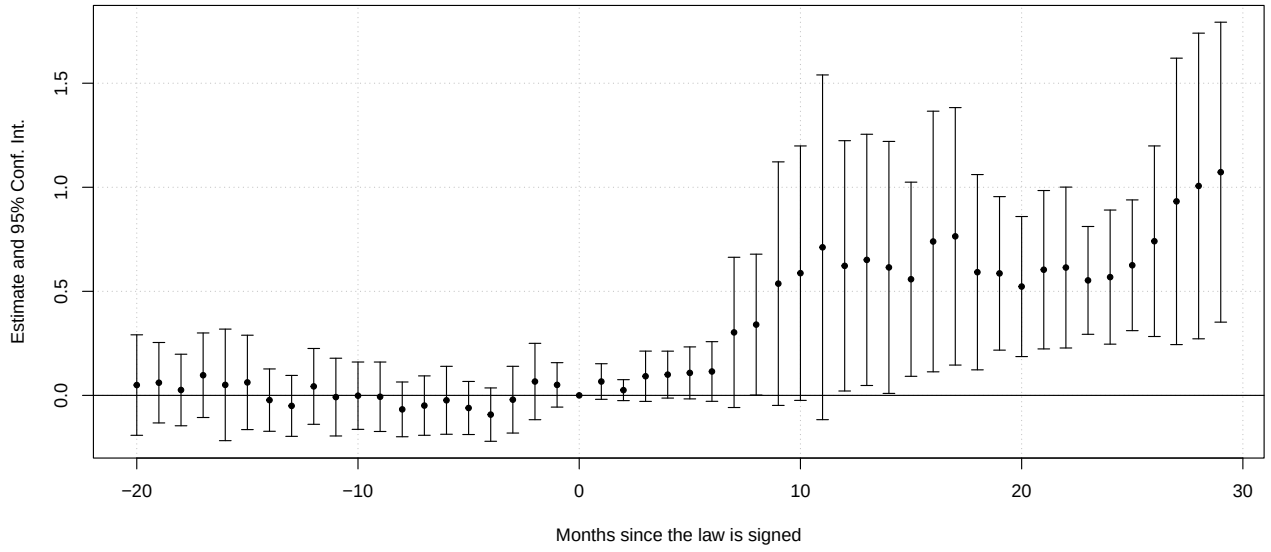
(b) Average Number of Words

Estimates from an event study, where a change in algorithm serves as an event, news outlets that are removed from the top-5 news serve as a treatment group, and the rest of outlets serve as a control group. We control for outlet-by-day-of-the-week and day fixed effects. Outlets are weighted by the pre-treatment probability of being referenced by Yandex in the top-5 section. Error bars correspond to the 95% confidence intervals. Standard errors are clusters at the outlet level.

Figure 12: Impact of Yandex's Algorithm Change on News Production: Features of Titles



(a) # of Symbols in a Title



(b) # of Punctuation Symbols in a Title

Estimates from an event study, where a change in algorithm serves as an event, news outlets that are removed from the top-5 news serve as a treatment group, and the rest of outlets serve as a control group. We control for outlet-by-day-of-the-week and day fixed effects. Outlets are weighted by the pre-treatment probability of being referenced by Yandex in the top-5 section. Error bars correspond to the 95% confidence intervals. Standard errors are clusters at the outlet level.

Table 10: Five news events which were overreported by the independent outlets.

Index	Examples of Titles (Original and Translated) of News Articles in this Event
1	Осетинская: экономических причин для увольнения руководства РБК не было *Osetinskaya: there were no economic reasons to fire the management of RBC*
2	Спецслужбы проводят обыски в Библиотеке украинской литературы в Москве *Security forces are conducting a search in the Ukrainian Literature Library in Moscow*
3	Навальный подал жалобу из-за блокировки интернет-ресурсов с расследованием о Дерипаске *Navalny filed a complaint after the investigation about Deripaska was blocked online*
4	Прокурор попросит наказание для Алексея Навального по делу "Ив Роше" *A prosecutor asks for a penalty for Navalny in the "Yves Rocher" case*
5	На "Эхе Москвы" установят правила поведения сотрудников в соцсетях *"Echo of Moscow" will set rules for the behavior of its employees on the internet*

We exclude events which were reported by less than 10 outlets.

Table 11: Five news events which were underreported by the independent outlets.

Index	Examples of Titles (Original and Translated) of News Articles in this Event
1	В Кузбассе спасателям удалось добраться до заблокированных в шахте горняков *Rescuers reached the miners trapped in the mine*
2	Россия направит в Донбасс очередную колонну с гуманитарной помощью 27 января *Russia will send humanitarian aid to Donbass on January, 27*
3	На Байконуре на стартовый стол установили ракету с пилотируемым кораблем *The rocket with a manned spacecraft was set in Baikonur*
4	Установлен второй арочный пролет моста через Керченский пролив *The second span of the bridge across the Strait of Kerch was set*
5	Прощание с погибшим в Сирии офицером Прохоренко будет открытым *Farewell to deceased in Syria officer Prohorenko will be opened*

We exclude events which were reported by less than 10 outlets.

Table 12: Implied Outlet Weights Before and After the Law

		Probability of Citation (in %)		
Outlet	Type	Before the Law	After the Law	Difference
fontanka.ru		0.00	0.00	-0.00
meduza.io	int	0.00	0.00	-0.00
dni.ru	gov	0.00	0.00	0.00
tjournal.ru	ind	0.00	0.00	-0.00
utro.ru		0.00	0.00	-0.00
life.ru		0.00	0.00	-0.00
zona.media	ind	0.00	0.00	-0.00
polit.ru		0.00	0.00	-0.00
dw.com/ru	int	0.01	0.00	-0.01
znak.com		0.01	0.00	-0.01
ridus.ru		0.01	0.00	-0.01
svpressa.ru		0.01	0.00	-0.01
trud.ru		0.01	0.00	-0.01
sobesednik.ru		0.01	0.01	-0.01
snob.ru	ind	0.01	0.03	0.02
the-village.ru	ind	0.02	0.03	0.01
rosbalt.ru		0.02	0.00	-0.02
kp.ru		0.02	0.02	-0.00
golos-ameriki.ru	int	0.02	0.00	-0.02
bbc.com	int	0.02	0.00	-0.02
fb.ru		0.03	0.00	-0.02
svoboda.org	int	0.03	0.00	-0.02
newtimes.ru	ind	0.09	0.00	-0.08
mk.ru		0.14	0.00	-0.13
thebell.io	ind	0.38	0.00	-0.38
newsru.com	ind	0.55	0.00	-0.55
ng.ru		0.60	0.03	-0.57
rusplt.ru		0.60	0.00	-0.60
kommersant.ru		0.64	0.57	-0.08
bfm.ru		1.24	0.13	-1.11
1tv.ru	gov	1.43	0.07	-1.36
aif.ru	gov	1.47	0.09	-1.38
russian.rt.com	gov	1.51	1.65	0.15
republic.ru	ind	2.07	0.11	-1.96
ntv.ru	gov	2.11	0.37	-1.74
vz.ru	gov	2.69	0.09	-2.60
vedomosti.ru	ind	2.71	3.37	0.66
tvrain.ru	ind	2.71	0.05	-2.66
echo.msk.ru	ind	2.80	0.31	-2.49
regnum.ru		3.06	5.02	1.96
novayagazeta.ru	ind	3.89	0.02	-3.88
tass.ru	gov	4.16	9.79	5.63
vesti.ru	gov	4.60	0.71	-3.88
iz.ru		6.53	6.00	-0.52
forbes.ru	ind	6.88	0.10	-6.78
lenta.ru		7.30	7.33	0.03
rbc.ru	ind	7.41	12.17	4.76
gazeta.ru		8.06	13.67	5.61
rg.ru	gov	10.61	15.66	5.05
ria.ru	gov	13.51	22.56	9.05

Implied probabilities of citations based on outlet weights α_j and α'_j from equation 3.

Table 13: Effectiveness of Outlet References in Yandex's Top-5: By Event Topic

	Dependent variable:
	$y_{jd}/y_{jd}^{\text{Yandex}}$
$X(j$ in top-5$)_d$	0.029***
	(0.005)
$X(j$ in top-5 with topic Society$)_d$	0.010
	(0.010)
$X(j$ in top-5 with topic Other$)_d$	-0.010
	(0.013)
$X(j$ in top-5 with topic Economics$)_d$	-0.003
	(0.010)
$X(j$ in top-5 with topic Disasters and Terrorism$)_d$	-0.017*
	(0.009)
$X(j$ in top-5 with topic World News$)_d$	0.089*
	(0.049)
$X(j$ in top-5 with topic Domestic Politics$)_d$	-0.013
	(0.011)
$X(j$ in top-5 with topic Sports$)_d$	-0.021**
	(0.008)
$X(j$ in top-5 with topic International Politics$)_d$	-0.009
	(0.009)
$X(j$ in top-5 with topic Defense and Military$)_d$	-0.013
	(0.012)
$X(j$ in top-5 with big event$)_d$	0.005
	(0.006)
$I(j$ in top-10$)_d$	0.001***
	(0.0002)
Fixed Effects:	
- Week x Outlet	Y
- Week Day x Outlet	Y
- Date	Y
Clustered S.E.	Weeks
Observations	20,084
Adjusted R^2	0.957

Note: *p<0.1; **p<0.05; ***p<0.01

Each event is classified by us into one of 10 topics: 9 of them are presented in the table, and the last one is "Unclassified" (when we observe an outlet in top-5, but cannot match its article's title with one of our events).

We also control on the "size" of the event: we call it "big" if more than 10 outlets wrote about it.

Table 14: Effectiveness of Outlet References in Yandex's Top-5: By Event Type

	Dependent variable:
	$y_{jd}/y_{jd}^{\text{Yandex}}$
$X(j$ in top-5$)_d$	0.024***
	(0.004)
$X(j$ in top-5 with big IND event$)_d$	-0.0001
	(0.006)
$X(j$ in top-5 with big GC event$)_d$	0.001
	(0.007)
$X(j$ in top-5 with big event$)_d$	0.002
	(0.005)
$I(j$ in top-10$)_d$	0.001***
	(0.0002)
Fixed Effects:	
- Week x Outlet	Y
- Week Day x Outlet	Y
- Date	Y
Clustered S.E.	Weeks
Observations	20,084
Adjusted R^2	0.957

Note: *p<0.1; **p<0.05; ***p<0.01

We find "big" events (those which were written by more than 10 outlets) and among them we detect IND events (those which were significantly under-reported by the government-controlled outlets) and GC events (those which were significantly under-reported by the independent outlets).

Table 15: Effectiveness of Outlet References in Yandex's Top-5: By Event Type and Regional News

	Dependent variable:
	$y_{jd}/y_{jd}^{\text{Yandex}}$
$X(j$ in top-5$)_d$	0.02381**
	(0.00716)
$X(j$ in top-5 with news about this region$)_d$	0.1659***
	(0.02055)
$X(j$ in top-5 with big IND event$)_d$	−0.00965
	(0.01086)
$X(j$ in top-5 with big event$)_d$	0.00579
	(0.00783)
$I(j$ wrote about this region's news$)_d$	0.000915***
	(0.0000804)
$I(j$ in top-10$)_d$	0.0006347*
	(0.0003038)
Fixed Effects:	
- Week x Outlet x Region	Y
- Week Day x Outlet x Region	Y
- Date x Region	Y
Clustered S.E.	Weeks
Observations:	244,723
Adjusted R^2:	0.9565

Note: *p<0.05; **p<0.01; ***p<0.001.

Table 16: Effectiveness of Outlet References in Yandex's Top-5: By Event Yandex Ranking

	Dependent variable:
	$y_{jd}/y_{jd}^{\text{Yandex}}$
$X(j$ in top-5$)_d$	0.028^{***}
	(0.003)
$X(j$ in top-5, bottom quartile news value$)_d$	-0.014^{***}
	(0.005)
$X(j$ in top-5, top quartile news value$)_d$	0.001
	(0.005)
$I(j$ in top-10$)_d$	0.001^{***}
	(0.0002)
Fixed Effects:	
Week x Outlet	Y
Week Day x Outlet	Y
Date	Y
Clustered S.E.:	$Weeks$
Observations:	20,084
R^2:	0.966
Adjusted R^2:	0.957
Residual Std. Error:	0.006 (df = 15,545)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 17: Effectiveness of Outlet References in Yandex's Top-5: By Operation System

	Dependent variable:	
	$y_{jd}/y_{jd}^{\text{Yandex}}$	
	Detailed	Grouped
$X(j \text{ in top-5})_d$ x Windows	0.029^{***}	
	(0.005)	
$X(j \text{ in top-5})_d$ x Android	0.042^{***}	
	(0.009)	
$X(j \text{ in top-5})_d$ x iOS	0.037^{***}	
	(0.007)	
$X(j \text{ in top-5})_d$ x MacOS	0.044^{***}	
	(0.009)	
$X(j \text{ in top-5})_d$ x Other	0.029^{***}	
	(0.005)	
$X(j \text{ in top-5})_d$ x Desktop		0.029^{***}
		(0.005)
$X(j \text{ in top-5})_d$ x Mobile		0.040^{***}
		(0.008)
Controls:		
- $I(j \text{ in top-10})_d$ x OS	Y	Y
Fixed Effects:		
- Week x Outlet x OS	Y	Y
- Week Day x Outlet x OS	Y	Y
- Date x OS	Y	Y
Clustered S.E.	Weeks	Weeks
Observations	39,405	15,762
Adjusted R^2	0.966	0.927

Note: *p<0.1; **p<0.05; ***p<0.01

The estimates correspond to the a estimates from regression (1) with y_{id} normalized by the volume of traffic arriving on `yandex.ru` on day d.

Table 18: Effectiveness of Outlet References in Yandex's Top-5: By Age

	Dependent variable:								
	$y_{jd}/y_{jd}^{\text{Yandex}}$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$X(j \text{ in top-5})_d$	0.018***	0.011***	0.014***	0.020***	0.023***	0.025***	0.028***	0.030***	0.034***
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.005)
$I(j \text{ in top-10})_d$	0.0002	0.0002	0.0003*	0.0003	0.0003	0.0003	0.0004	0.001	0.001*
	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0004)
Age (years)	<12	12-18	19-24	25-30	31-35	36-40	41-45	46-50	>50
Fixed Effects:									
- Week x Outlet	Y	Y	Y	Y	Y	Y	Y	Y	Y
- Week Day x Outlet	Y	Y	Y	Y	Y	Y	Y	Y	Y
- Date	Y	Y	Y	Y	Y	Y	Y	Y	Y
Clustered S.E.	Weeks	Weeks	Weeks	Weeks	Weeks	Weeks	Weeks	Weeks	Weeks
Observations	6,920	6,920	6,920	6,920	6,920	6,920	6,920	6,920	6,920
R^2	0.976	0.979	0.976	0.974	0.973	0.975	0.976	0.979	0.982

Note: *p<0.1; **p<0.05; ***p<0.01

The estimates correspond to the a estimates from regression (1) with y_{id} normalized by the volume of traffic arriving on `yandex.ru` on day d.

Table 19: Effectiveness of Outlet References in Yandex's Top-5: By Gender

	Dependent variable:		
	$y_{jd}/y_{jd}^{\text{Yandex}}$		
	(1)	(2)	(3)
$X(j \text{ in top-5})_d$	0.029***	0.018***	0.033***
	(0.004)	(0.003)	(0.005)
$I(j \text{ in top-10})_d$	0.001*	0.0002	0.0004
	(0.0003)	(0.0002)	(0.0003)
Gender	Male	Female	Undetected
Fixed Effects:			
- Week x Outlet	Y	Y	Y
- Week Day x Outlet	Y	Y	Y
- Date	Y	Y	Y
Clustered S.E.	Weeks	Weeks	Weeks
Observations	6,920	6,920	6,920
R^2	0.982	0.970	0.983

Note: *p<0.1; **p<0.05; ***p<0.01

The estimates correspond to the a estimates from regression (1) with y_{id} normalized by the volume of traffic arriving on `yandex.ru` on day d.