Platform Power in News Aggregation*

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

March 2026

Abstract

News aggregators are among the primary gatekeepers of online news, shaping both consumption and production through the algorithms that rank stories and sources. We provide the first large-scale estimates of this algorithmic power using Yandex News, Russia's largest aggregator. Combining snapshots of its front-page rankings with 12.3 million articles from 50 outlets and daily traffic data, we estimate that a front-page reference generates roughly 380,000 additional daily visitors for an outlet—an effect that is stable across news topics, event types, and consumer demographics. Exploiting a 2016 Russian law that made aggregators liable for cited content as a natural experiment, we show that Yandex halved its references to independent outlets, redistributing traffic by up to 27% across outlets. Consumers did not penalize this manipulation: referred outlets continued to capture the same share of front-page visitors. Outlets removed from the rankings subsequently shifted toward longer articles and stopped optimizing headlines for the algorithm, showing that aggregator algorithms affect not only consumption but also production incentives.

*Authors contributed equally and names are listed in reverse alphabetical order. 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 Scott Gehlbach, Quan Le, Alexey Makarin, Karine Van Der Straeten, 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.

†Columbia University and CEPR. Email: as5443@gsb.columbia.edu.

‡Stanford University. Email: daniilmikhailov11@gmail.com

§ICREA-Universitat Pompeu Fabra, IPEG and CEPR. Email: Ruben.Enikolopov@upf.edu

¶National University of Singapore, ICREA-UPF, CESifo, IZA, and CEPR. Email: rubendurante@gmail.com

1 Introduction

Digital platforms that mediate access to information—such as search engines, social media, and news aggregators—have become the primary gatekeepers of online news consumption, responsible for the majority of news readership (Reuters, 2024). These platforms do not produce news but design algorithms that decide which stories are surfaced, in what order, and from which sources. The most vivid example is news aggregators, whose only product is rankings of news events and sources, but the resulting product impacts how much and which news are consumed.

How much can news aggregators impact the news ecosystem through their control of the algorithm that ranks news content? On the one hand, platforms can adjust the algorithm and manipulate news content shown to consumers. If consumers are swayed by these rankings, the platform can impact which news events are read and which outlets get traffic, impacting news consumption and changing incentives for news production. This implies that news aggregators have market and media power (Prat, 2018) through the distribution of the media content produced by others, which is different from the media power of traditional media that make production decisions.[1] On the other hand, consumers might see through deviations from truthful aggregation of news events—that ranks events and outlets by their importance and relevance—and choose alternative channels for news consumption. In that way, the power of platforms can be constrained by the demand factors in the same way it constrains the media power of traditional media (Durante and Knight, 2012; Knight and Tribin, 2019).

We study the ability of news aggregators to influence online news consumption and production using the example of the major Russian news aggregator, Yandex News. We focus on the period between 2013 and 2017, which combines two useful features for our analysis. First, during this period, the front page of Yandex search engine prominently featured the top-5 news stories from Yandex News, exposing tens of millions of daily users to the selected stories and referring these users to selected outlets. This allows us to collect historical data on Yandex's recommendations and gives statistical power to measure the impact of Yandex News's ranking on upstream news outlets' traffic. Second, during this period, Yandex made major changes to its news aggregation algorithm in response to changes in Russian legislation and pressure from the Russian government(Wijermars, 2021); we use this variation to examine whether discretion in changing the algorithm is constrained by demand factors and identify the impact of algorithms on news production. In addition to providing a convenient empirical context to study general effects of news aggregators, the case study of Yandex News in Russia is important in itself, since it helps to assess the role Yandex played in the gradual government capture of the Russian online news market (Simonov and Rao, 2022).

To understand the impact of Yandex's aggregation algorithm on news consumption and pro-

[1] Similarly, social media affects news consumption primarily through its effect on the personalized distribution of news content created by other online outlets (Braghieri et al., 2025)

duction, we combine three new datasets. First, we use the Wayback Machine to scrape 29,841 snapshots of top-5 news stories on Yandex's homepage in 2013-2017. This allows us to track Yandex's news citations and outlet referrals over time. We augment these data with the universe of publication records (12.3 million articles) covering 50 major Russian online news websites, helping us to understand the underlying distribution of news articles and events in Russian media and measure changes in news production over time. Finally, for a subset of news outlets, we collect data from an internet tracker on daily traffic broken down by types of visitors, traffic sources, and other features. This data allows us to measure the effect of Yandex referrals on news consumption and outlets' market shares.

We start by estimating the impact of Yandex's top-5 referrals on the upstream news outlets' traffic. Our identification strategy relies on the sharp difference in visibility between articles that appear in the top-5 section on Yandex's front page and those that do not. We exploit day-to-day variation in the exact share of top-5 slots that an outlet occupies over the course of a day, after absorbing outlet-by-week, outlet-by-day-of-the-week, and day fixed effects. As an additional check, we control for whether the outlet appears in the less visible positions just below the top-5, confirming that our estimates capture the effect of prominent front-page placement rather than the general newsworthiness of the outlet's coverage on that day.

Our results indicate that Yandex referrals have a large impact on traffic. A one standard deviation increase in how often an outlet's link appears in at least one top-5 slot during a day increases outlet traffic by around 27,600 visitors. If all slots in Yandex's top-5 positions were allocated to a single outlet for an entire day, its traffic would grow by nearly 380,000 visitors. Because the effect is pinned down in levels by the number of visitors to Yandex's front page rather than by outlet size, smaller outlets benefit disproportionately: the same referral traffic represents a 26% increase for the largest outlet in our sample but a 1,000% increase for the smallest one. The effect is larger for higher positions, ranging from 58,000 additional visitors for the fifth slot to 97,000 for the first, and becomes small and imprecise for positions 6–10, which do not appear on Yandex's front page.

We validate the causal nature of our estimates in a series of robustness checks. First, we show that referrals in Yandex's top-5 slots impact only outlets' traffic from Yandex but not direct traffic or traffic from other sources (e.g., social media and other aggregators), indicating that the effect is not driven by the general newsworthiness of the articles featured in Yandex's top-5. Second, appearing in Yandex's top-5 slots only affects an outlet's traffic on the day it is referred to but has no detectable effect on future traffic, ruling out strong intertemporal spillovers. On that day, however, incremental referrals generate 1.8 page views per visit, implying that referred users also browse articles beyond the one directly linked by Yandex. Third, Yandex's referrals disproportionately affect casual consumers—who do not regularly visit the outlet and make fewer page visits—rather than regular ones.

Having measured the impact of Yandex's algorithm on upstream outlets' traffic, we check whether Yandex's ability to direct traffic to outlets is constrained by the types of cited topics or the composition of visitors. We find little evidence for such constraints. First, the effects of Yandex's referrals on an outlet's traffic are similar across major topic areas (e.g., politics, sports), for government-sensitive coverage (e.g., related to political opposition or government corruption), and for both high- and low-prominence events, suggesting that Yandex's capacity to drive traffic is not meaningfully constrained by the news events it cites. Second, we compare the effectiveness of Yandex across user characteristics. We find stronger effects for visitors identified by the traffic tracker as older and male, but no difference across operating systems (e.g., iOS, Windows) and device types (e.g., mobile, desktop).

To understand whether Yandex can change its algorithm at will and if it is constrained by the demand factors, we turn to the analysis of the effect of the adoption of the major law "on news aggregators" in August 2016. The law made news aggregators liable for the content they cite, with some exceptions based on the source of news (Wijermars, 2021).[2]

We observe a major change in Yandex's algorithm shortly after the signing of the law. 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 but two business-focused independent outlets entirely excluded from the top-5 slots. At the same time, Yandex maintained broadly the same share of citations of sensitive events, indicating that the algorithm change operated through outlet removal rather than direct topic suppression.

Despite this drastic change, the ability of Yandex to refer traffic to news outlets is maintained after the passing of the news aggregators law. We see no evidence of users losing confidence in the aggregator and discounting its recommendations. We also see no detectable change in Yandex's market share. Taken together, we find no evidence that the ability of the news aggregator to change its algorithm under political pressure is constrained by the demand factors.

To understand the mechanism through which Yandex could respond to the law and can exercise its control over traffic allocation more broadly, we reverse-engineer its algorithm, guided by Yandex's public description of its design principles. The algorithm proceeds in two steps: first, it ranks news events—clusters of recent news articles—by prominence; second, it selects a specific article from a particular outlet for each event. We estimate both steps using our data on news events, constructed via a topic detection and tracking algorithm (Allan et al., 1998; Cagé et al., 2020), and Yandex's observed citations.

We find that in the first step, events covered by more outlets, with more articles, and devel-

[2]Before the law, news aggregators were not held liable for the information they aggregated and distributed, with the legal responsibility carried by news publishers. The law made aggregators liable for spreading "fake news" unless the information came from the media registered with the government.

oping faster are more likely to be cited. Notably, Yandex's algorithm under-reports events that government-controlled outlets systematically under-cover, even before the law.

In the second step, Yandex favors outlets that publish more articles about the event, do so faster, mention more facts, and have shorter titles that do not copy earlier coverage. However, much of the variation in referrals is explained by outlet weights, captured as fixed effects, which can be adjusted seamlessly over time. We show that the removal of independent outlets after the 2016 law was driven by a direct change in these weights, rather than adjustments to how article features are scored.

Using these estimates, we evaluate counterfactual scenarios that quantify Yandex's potential power over traffic allocation. If Yandex could direct all referrals to a single outlet with no restrictions on which events to cite, it could increase that outlet's traffic by between 26% and 1,000%, depending on outlet size. Even when constrained to cite the same events, Yandex could increase outlets' market shares by 15%–260% by adjusting only outlet weights. Evaluating the actual 2016 algorithm change, we find traffic losses of up to 20% for removed outlets and gains of up to 27% for favored ones, with aggregate traffic shifting from independent to government-controlled and potentially influenced outlets.

We also document upstream production effects of the algorithm change. Comparing the 9 outlets removed from Yandex's top-5 slots to the remaining 15, we show that removed outlets shifted toward longer articles, reduced their coverage of aggregator-cited events, and stopped optimizing headline features for the algorithm, with adjustments taking roughly 10 months to materialize.

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 provides background information. Section 3 describes our data sources. Section 4 explains how we classify news articles into events, as well as describes the events. Section 5 estimates the effect of referencing outlets in Yandex's top-5 section on the outlets' traffic. Section 6 looks at the changes after the adoption of the law regulating news aggregators. Section 7 concludes.

2 Background Information

2.1 Yandex News

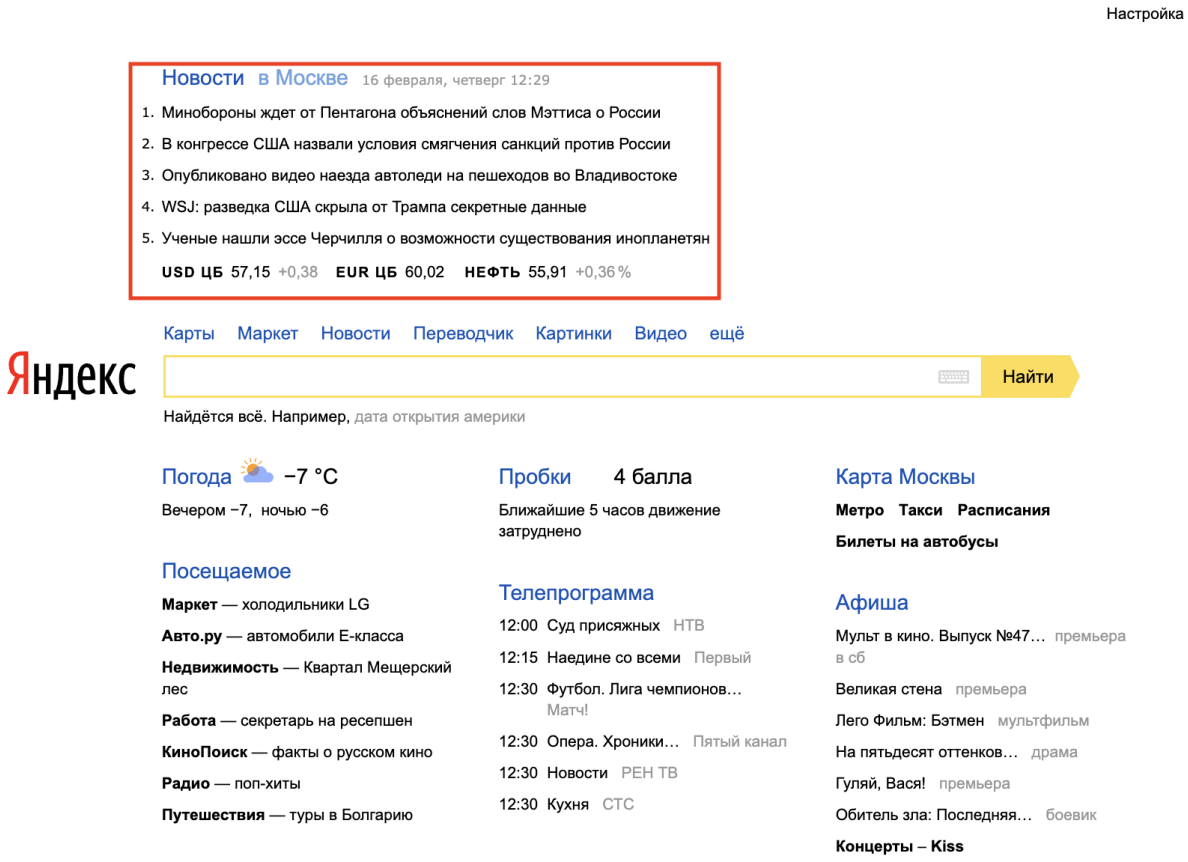
Yandex News (*Yandex.Novosti*) is an automated news aggregation and recommendation service operated by Yandex, a major Russian search engine. Like other aggregators, it does not produce original reporting; instead, it continuously ingests articles from a large pool of news outlets, detects underlying news events, ranks news events by prominence, and displays the most prominent events. The prominence of news events is determined by an algorithm, which takes into account available data on news production (e.g., volume of articles, recency) and consumption (e.g., clicks, shares on social media) (e.g. see Chiou and Tucker, 2017; Athey et al., 2021).

Our analysis focuses on Yandex News in 2013–2017, when Yandex was one of the two dominant search engines in Russia and Yandex News was the country's largest news aggregator.[3] The webpage of Yandex News `news.yandex.ru` displayed the five most important stories across all categories and by category (politics, international, sports, etc.). A defining feature of this period is that Yandex prominently embedded the list of the five most important stories across categories

[3]Yandex's market share in search was similar to Google. 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`.

on the Yandex homepage `yandex.ru`, immediately above the search box (see Figure 1). The top-5 news stories were global (similar for all visitors) and were frequently updated. A click on a news title directed a user to the referenced news outlet.[4] This design choice exposed a very large daily audience to the ranked headlines and generated substantial referral traffic to the selected outlets, making Yandex News a plausible "gatekeeper" for both news consumption and upstream production incentives.

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 the 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).

2.2 The Law "On News Aggregators"

The 2013–2018 period is especially important because it spans a major regulatory and political inflection point for Russian news intermediaries. On 23 June 2016, President Vladimir Putin

[4]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.

signed a new federal law targeting news aggregators with large daily audiences (Wijermars, 2021), which became effective on 1 January 2017. Before the law, news aggregators, such as Yandex News and Google News, were not held liable for the information they aggregated and distributed, with the legal responsibility carried by news publishers.[5] The law changed this, making aggregators liable for spreading "fake news" unless the information came from the media registered with the government.

We use this discrete policy-driven shift in the legal environment to study how it affected algorithmic rankings and how, in turn, this change in algorithmic rankings affected both downstream consumption (traffic to outlets) and upstream news production.

3 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.

3.1 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`.

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 29,841 snapshots of the front page covering the period between June 2013 and December 2017. The median time between two snapshots is 32 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, which, in addition to the top-5 news titles, also contains the articles that were close to entering the top-5 by being among the top news stories in different categories, but did not quite make it to appear on the front page of Yandex.

In particular, we construct the measure that reflects the share of Yandex top-5 slots occupied by a particular outlet throughout the entire day (it equals 1 if the outlet held all top-5 slots for the whole day). We also construct an indicator of whether the outlet appears in the "top-6" slot on a particular day. While the top-6 slot does not exist on the main page of Yandex, we can infer which news events are almost in the top-5 slots by looking at the top news on Yandex News

[5]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.

webpage by news topic subcategories [6]. We also 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. These measures are constructed based on the data scraped from the Yandex News webpage `news.yandex.ru`, which has more sparse records on the Wayback Machine than the main Yandex webpage `yandex.ru` due to a lower traffic volume to this page. As a result, we have fewer observations for these variables.

3.2 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 12,336,080 news articles published during the period from 23 June 2013 to 4 December 2017.

Both government-controlled and independent news outlets coexisted in the Russian media space at the time. We follow the approach of Simonov and Rao (2022) and use outlets' ownership structure, location, and information on the independence of their editorial boards to classify news outlets into four groups: government-controlled, which are directly owned and controlled by the government; independent but potentially-influenced, which are 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.

We match the publication records to the news that Yandex cites 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. [7] 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. See section 4 for more details

[6] The number of sub subcategories varied somewhat during this period, but the main categories included Politics, Society, Economics, World News, Sports, Crime, Culture, Science, Technology, Auto.

[7] Yandex stopped directing users to source outlets in April 2018. Four months earlier, in December 2017, Yandex redesigned its main page to display a "dynamically updated" top-5 ranking that changed while users remained on the page. Although the HTML code contains entries for all articles that appeared in the top-5 positions, we restrict our analysis to the period prior to this redesign to ensure that each ranked position was actually observed by users.

3.3 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), visitor demographics (age, gender, region), total pageviews, desktop and mobile visitors, among other metrics. For a subset of these websites, we observe more detailed statistics, such as daily visits to particular webpages, the location of visitors, their browser, and their operating 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 and abroad. In particular, we use `liveinternet.ru`, `rambler.ru`, and `semrush.com`; while these domains do not allow us to pull historical data from 2013-2017, we examine the correlation in the traffic records from these trackers and `top.mail.ru` in the more modern data. Appendix A.1 provides details of this validation exercise. We also use traffic data for all 50 news outlets since 2017, provided by `semrush.com` to estimate the effect of changes in the aggregator's algorithm on overall news consumption.

We supplement the traffic data with data on political preferences, beliefs, and demographic characteristics from the MegaFOM opinion poll conducted by the Public Opinion Foundation (Fond Obschestvennogo Mneniya, or FOM) in October–November 2011. This is a regionally representative survey of 56,900 respondents from 79 regions. We use these data to characterize the effectiveness of Yandex in directing consumers of different political beliefs to news articles and topics. See Appendix A1 for details about the dataset.

4 Classification of News Events

Because Yandex's algorithm operates on news events rather than individual articles (see Section 6.1), we need to detect and characterize events in the data both to characterize the effect of Yandex's referrals on outlet traffic and the algorithm itself.

We classify articles into news events using the topic detection and tracking algorithm (TDT) developed by Allan et al. (1998) and used in previous work (e.g. Cagé et al., 2020). The method represents articles as TF-IDF vectors and clusters those published within overlapping two-day windows based on cosine similarity; clusters that share articles across windows are merged into a single event. We select hyperparameters—including the vectorization method, clustering algorithm, and radius—using the fact that our dataset includes 19 Ukrainian Russian-language outlets

alongside the 50 Russian ones. Since Ukrainian outlets should cover largely distinct events from the Russian ones, we select the specification that best separates Russian and Ukrainian articles while maintaining adequate clustering coverage. Our preferred specification uses Agglomerative Clustering with a radius of 0.6; we provide details and verify robustness to alternative settings in Appendix A2.

The resulting classification produces 1.88 million news events (defined as clusters of at least two articles), covering 71% of the 12.3 million articles in our data. Panel A of Table 1 presents summary statistics: a median event has three articles from two outlets and lasts 5.5 hours. The distribution is heavily right-skewed, with the largest event containing 1,709 articles from 49 outlets. We match these events to Yandex's citations using the article URLs observed on Yandex's front page. Events cited by Yandex (Panel B) are substantially larger—a median of 14 articles from 12 outlets lasting 12.3 hours—as expected given that the algorithm prioritizes prominent events.

Table 1: Summary Statistics of News Events

	Min	25%	Median	Mean	75%	Max
Panel A: All Events						
articles per topic	2	2	3	4.69	5	1709
outlets per topic	1	2	2	3.57	4	49
topic length (hours)	0	1.68	5.52	15.08	20.52	359.95
Panel B: Yandex Top-5 Events						
articles per topic	2	7	14	30.77	28	1709
outlets per topic	1	6	12	14.34	21	49
topic length (hours)	0	4.32	12.30	33.90	40.47	357.90
Panel C: Sensitive Events						
articles per topic	10	14	19	34.06	31	1543
outlets per topic	10	12	15	18.46	22	49
topic length (hours)	0.37	10.84	22.68	40.24	50.76	356.93

Panel A reports statistics for all 1.88 million detected news events.
Panel B reports statistics for 34 thousand events cited in Yandex Top-5 news at least once.
Panel C reports statistics for 8734 events censored by government media.

We also classify events as government-sensitive based on whether they are systematically under-reported by government-controlled outlets relative to independent ones, following the procedure of Simonov and Rao (2022) (see Appendix A3 for details). Panel C of Table 1 summarizes these events. They are larger and longer-lasting than typical events, and predominantly involve political opposition and government corruption—for example, the activities of Alexey Navalny and changes in the editorial leadership of independent media.

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 30 different topics. We

grouped these topics into 16 broader categories.[8]

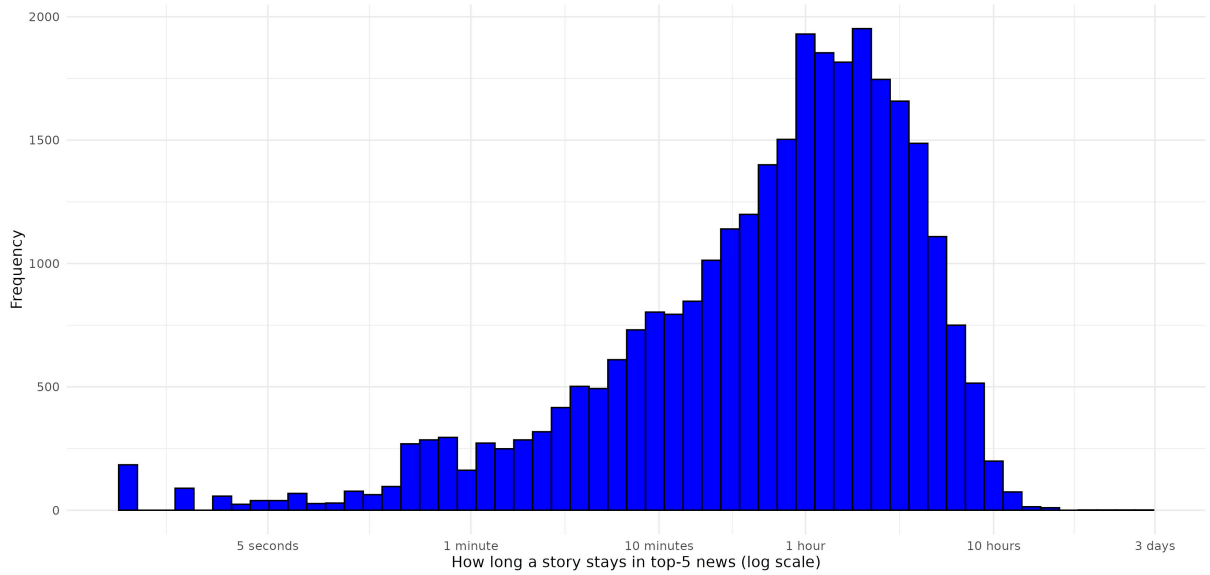
5 Effects on News Consumption

5.1 Descriptive Evidence

Before turning to the main estimates, we document two features of the Yandex citation process that are important for our analysis.

First, news articles rotate rapidly through the top-5 section: on a typical day, 60–70 different articles appear in the top-5, and a median story remains there for only 56 minutes (Figure 2).[9] Almost no articles are observed in the top-5 for more than 10 hours. This high turnover generates substantial day-to-day variation in outlets' top-5 exposure, which we exploit for identification.

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.

Second, the same outlet can occupy multiple top-5 slots simultaneously: the most cited outlet of the moment holds an average of 1.84 slots. In section 6 we examine how this concentration increased after the 2016 law on news aggregators.

[8] A set of 16 categories closely corresponds to the topic groups employed by Yandex News. These categories include Foreign Politics, Military, Economy, Culture, Incidents, Local, Society, Education, Sports, Domestic Politics, Science, Crime, Weather, Health, Technology, and Environment.

[9] This is based on articles observed in at least two snapshots, implying that actual rotation is even faster.

5.2 Effects on Outlet Traffic

To 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 variation in news traffic to different outlets to estimate the impact of being featured in Yandex News' top-5 section. In particular, we estimate the following model:

$$y_{jd} = \alpha X(j \text{ in top-5})_d + [\beta I(j \text{ in top-6})_d] + \rho_{jw} + \delta_{j,\text{wd}} + \theta_d + \xi_{jd} \quad (1)$$

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 day d. As mentioned in section 2, we have fewer observations for this variable, so we have fewer observations once 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 realizations of the share of references an outlet gets on a given day in the top-5 section, as well as the exact position of these references within the top-5 section. 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 ρ_{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 2 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 56.8% (column 2), or 374 thousand visitors (column 4). The effect is robust for controlling for appearing in the "top-6", which indicates that the effect on traffic is not driven by more general "newsworthiness" of the stories. 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 3 presents the estimates. As expected, higher positions bring more traffic to an outlet than lower positions. For instance, occupying position 1 for an

Table 2: Effect of Outlet References in Yandex's Top-5: Overall Traffic

	Dependent variable:			
	$log(y_{jd})$		y_{jd}	
	(1)	(2)	(3)	(4)
$X(j$ in top-5$)_d$	0.46***	0.45***	341.52***	374.03***
	(0.03)	(0.04)	(32.53)	(38.13)
$I(j$ in top-6$)_d$		0.04***		10.64***
		(0.01)		(3.43)
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	22,971	20,084	22,971	20,084
Adjusted R^2	0.98	0.98	0.95	0.95
Note:				*p<0.1; **p<0.05; ***p<0.01

entire day brings an outlet 97 thousand additional 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.

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 1 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.

Separating the new and regular visitors indicates thata full-day reference in all of the top-5 links of Yandex brings in 157 thousand new visitors, but no regular visitors (see Columns (2)-(3) in Table 4). This supports the result that the referral traffic from Yandex's top-5 is incremental.

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 outlets, as none of the outlets used a subscription model at the time and instead relied on display advertising for monetization.[10] The 374 thousand new visitors bring in 681 thousand more page views (See Column 4 of Table 4).

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

Table 3: Effect of Outlet References in Yandex's Top-5: By Position

	Dependent variable:			
	y_{jd}			
	(1)	(2)	(3)	(4)
$X(j$ in position $1)_d$	92.00***	95.05***	97.47***	97.26***
	(15.33)	(18.90)	(18.77)	(18.75)
$X(j$ in position $2)_d$	83.92***	95.34***	96.86***	96.82***
	(15.61)	(18.85)	(18.75)	(18.79)
$X(j$ in position $3)_d$	58.33***	67.20***	68.60***	68.61***
	(13.94)	(15.67)	(15.61)	(15.54)
$X(j$ in position $4)_d$	55.28***	59.29***	61.29***	61.65***
	(13.55)	(15.72)	(15.68)	(15.70)
$X(j$ in position $5)_d$	51.50***	57.09***	58.63***	58.44***
	(13.53)	(15.70)	(15.69)	(15.72)
$I(j$ in top-$6)_d$		10.57***		
		(3.43)		
$\sum_{p=6}^{10} X(j$ in position $p)_d$			17.69*	
			(9.33)	
$X(j$ in position $6)_d$				28.76
				(24.28)
$X(j$ in position $7)_d$				9.55
				(22.70)
$X(j$ in position $8)_d$				18.05
				(26.39)
$X(j$ in position $9)_d$				4.98
				(25.46)
$X(j$ in position $10)_d$				28.10
				(25.85)
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	22,971	20,084	20,084	20,084
Adjusted R^2	0.95	0.95	0.95	0.95
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: 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)
$X(j \text{ in top-5})_d$	374.03***	157.06***	3.41	681.30***	-0.76***
	(38.13)	(15.42)	(2.79)	(88.10)	(0.09)
$I(j \text{ in top-6})_d$	10.64***	7.40***	-0.21	15.51	-0.06***
	(3.43)	(1.73)	(0.23)	(10.24)	(0.01)
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	0.95	0.88	1.00	0.95	0.96

Note: *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.

The daily average number of page views per visitor on the outlet falls when there is more traffic from Yandex (column 5 of Table 4). Taking into account estimates from Tables 2 and 4, the back-of-the-envelope calculation shows that marginal visitors attracted by references in the top-5 section have fewer page views (1.8) compared to inframarginal visitors (around 4 pageviews for an average website).[11]

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 (1) with only the traffic referred by Yandex (y_{jd}^{Y}), all traffic net of Yandex referrals (y_{jd}^{-Y}), and all external traffic net of Yandex referrals ($y_{jd}^{\text{ext}-Y}$). We use specifications with traffic in logs to improve the precision of our estimates (since we are relying on a subset of outlets).

Table 5 presents the results. First, in Column (1), we replicate the estimates presented in Column 2 in Table 2 using the subsample of four websites for which we have information on referral traffic. While the estimate is less precise – it has a standard error of 0.10 instead of 0.04 – 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 5, we split the total traffic to the outlets by their referral source.

[11]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 4 and the result from Table 2, which showed that 374 thousand consumers correspond to 56.2% of visitors for an average outlet.

Table 5: Effect of Yandex-top5 on Traffic 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)
$X(j \text{ in top-5})_d$	0.68***	1.16***	0.08	0.07
	(0.10)	(0.18)	(0.09)	(0.09)
$I(j \text{ in top-6})_d$	0.06***	0.16***	0.02**	0.02**
	(0.01)	(0.02)	(0.01)	(0.01)
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	0.98	0.95	0.98	0.97
Note:				*p<0.1; **p<0.05; ***p<0.01

A full-day reference in some of the top-5 links more than triples the traffic coming to outlets from Yandex (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.

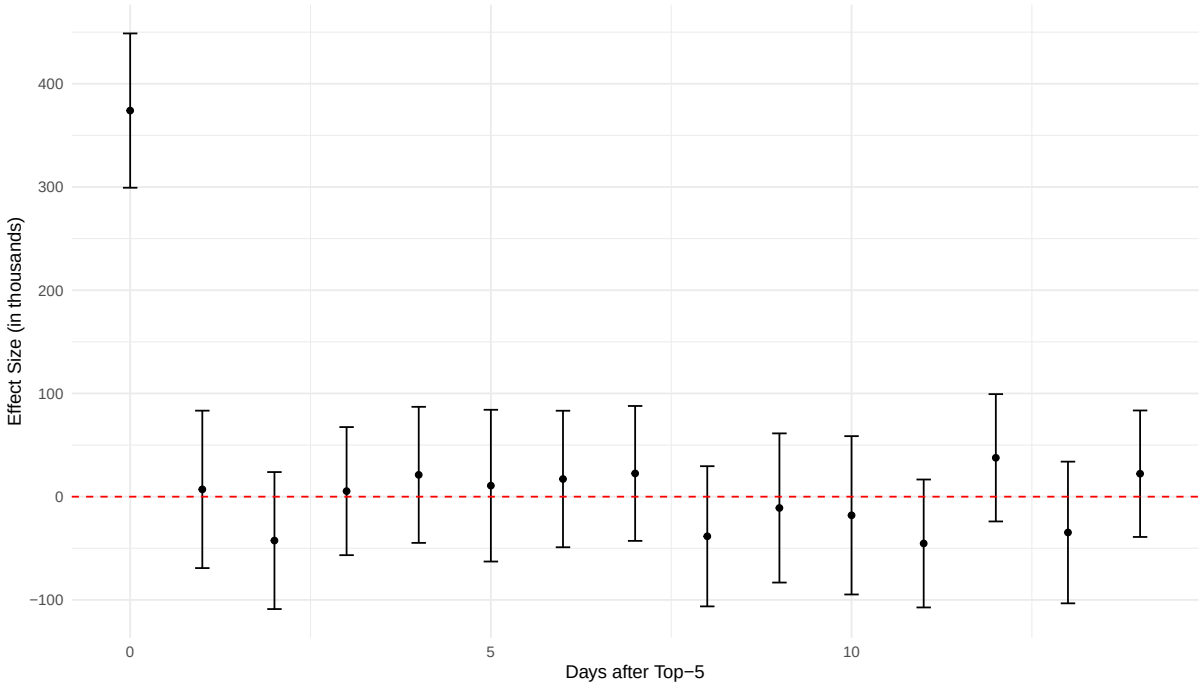
We further show that the incremental visitors coming to outlets because of the Yandex's references are not likely to become loyal readers and return to the same outlets later. For this, we re-run the regression 1 using daily total traffic to outlets from days $d+1, d+2, ...d+14$ instead of d. Figure 3 visualizes the estimates. For all the following days, the effect is not statistically different from zero and 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.

5.3 Heterogeneity of Effects

In the above estimates, we examine the volume of traffic that Yandex is able to direct to news outlets, given its current algorithm. However, we are interested in assessing the ability of Yandex to redirect traffic to 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 start by

Figure 3: 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.

estimating and comparing its effectiveness in directing traffic to news outlets for different types of 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 number of people using Yandex search that day (y_{jd}^{Yandex}).[12] 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 (1).

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 16 broader categories. 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).

[12] We use the number of unique people navigating from Yandex to websites tracked by `top.mail.ru` to measure the number of Yandex users on a given day. This measure counts more active users of Yandex, but does not capture users who came to Yandex and did not search. We assume that these inactive users were also not interested in the news section on the front page. On average, there are 1.85 more visitors to the front page compared to the number of users who search on Yandex. As a result, all effectiveness measures need to be divided by 1.85 to account for these less active users. See Appendix A0.1 for details.

Table A11 in Appendix A4 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 were censored by government-controlled outlets. Table A12 in Appendix A4 shows that there are no significant differences in Yandex's ability to direct its front page visitors to both stories that are under-reported by government-controlled media, implying that Yandex is able to increase or decrease citizen's exposure to censored news events.

We further show that the newsworthiness of articles does not matter for the share of visitors who navigate to news outlets. For this, we classify news titles by their predicted probability to be cited by Yandex in the top-5 news from regression (2). We then check if articles with a higher predicted probability generate more traffic to news outlets. Table A14 in Appendix A4 presents the estimates, splitting news events into the highest and lowest quartiles by newsworthiness. There are no significant differences between traffic generated by the lowest and the highest newsworthiness quartiles.

While visitors of Yandex seem indifferent between various news topics and news types, it is not because they do not have a preference for 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 (1) 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 A13 in Appendix A4 shows that consumers have very strong tastes for news about their region – the share of visitors navigating to those news stories increases from 2.7 percentage points for a typical news story to 17.5 percentage points, a 650% increase.

Overall, the results indicate that the effect of being cited in Yandex top-5 news does not depend much on the topic of the news, whether the news is censored by the government-owned outlets, or the newsworthiness of the events. However, if the news explicitly mentions a particular region of Russia, the effect is significantly stronger for users coming from this region.

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 A15 in Appendix A4 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 demographics. For this, we re-estimate the effectiveness of Yandex's references on traffic split by the

age and gender of visitors. Tables A16 and A17 in Appendix A4 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.

Table 6: Factors Correlated with Yandex Top-5 Effectiveness by Political Views

	Dependent variable:							
	Yandex Top-5 Effectiveness							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gov. party supporters	−0.036***							−0.026***
	(0.002)							(0.002)
Low income		−0.014***						−0.011***
		(0.002)						(0.002)
Likely voters			0.002					0.001
			(0.002)					(0.002)
Likely protesters				0.010***				−0.001
				(0.002)				(0.002)
See corruption as an issue					0.025***			0.011***
					(0.002)			(0.001)
Support democracy						0.023***		0.016***
						(0.007)		(0.005)
Low internet usage in demo							0.024***	0.020***
							(0.001)	(0.001)
Constant	0.040***	0.028***	0.023***	0.022***	0.013***	0.022***	0.015***	0.024***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.0005)	(0.001)
Weights:	# Respondents	# Respondents	# Respondents	# Respondents	# Respondents	# Respondents	# Respondents	# Respondents
Observations	1,229	1,229	1,229	1,229	1,229	1,229	1,229	1,229
R^2	0.271	0.033	0.001	0.014	0.137	0.010	0.285	0.526
Adjusted R^2	0.270	0.032	−0.00003	0.013	0.136	0.009	0.284	0.524

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

Regression unit is one demographic group: Region x Age x Gender. Yandex Top-5 Effectiveness was estimated as a coefficient in the regression for corresponding demographics. Each X variable represents the share of answers. Only respondents with often internet usage are taken for computing all X's, except for "Low internet usage in demo".

Finally, we examine differences in Yandex's effectiveness in directing news readers of different political views and demographics to referred news outlets. For this, first, we re-estimate the effectiveness of Yandex for each age, gender, and region combination. This results in 1,229 estimates for various combinations of demographics. We then merge these estimates with measures of the incumbent government support, income, voting and protest activity, and attitudes towards corruption and democracy from the opinion poll described in Section 3 for each of 1,229 demographic cells. [13]

The estimates are presented in Table 6. We find that Yandex News is more effective in directing traffic to news outlets for news readers who tend not to support the government, have higher incomes, are more likely to protest, are more likely to see corruption as a major issue for

[13]We use only respondents who say they use the internet frequently to make sure both survey estimates and traffic estimates are based on respondents who do spend time online. We get qualitatively similar estimates if we use all respondents instead.

the country, and are more likely to support democracy.

Overall, the estimates suggest that Yandex is more effective in directing to referred news outlets the readers who are more politically active and oppose the government. This makes Yandex's ability to control the aggregation algorithm even more attractive from the perspective of the incumbent government that wants to exercise media capture.

6 Effects of the Law

We next examine whether Yandex adjusted its algorithm in response to the external pressure of the 2016 law "on news aggregators" and measure the impact of this adjustment on news consumption and production. In this section, we document how Yandex changed the set of outlets and events it referenced in response to the law, estimate the underlying aggregation algorithm before and after the reform, and use these estimates to assess how alternative algorithmic choices would have reshaped outlets' traffic.

6.1 Descriptive Evidence

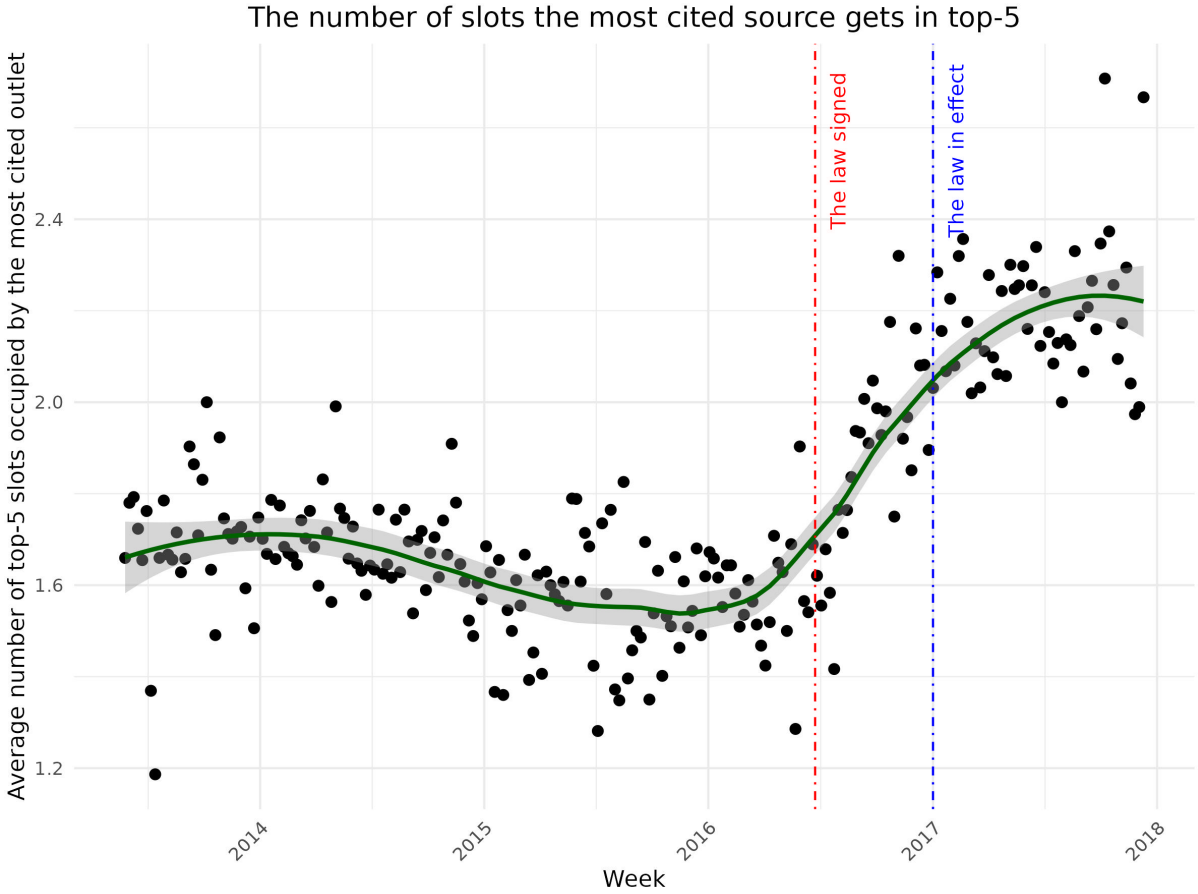
As was mentioned in section 2, in 2016, Russia adopted a new federal law that made news aggregators liable for the information they aggregated and distributed. Shortly after the law was signed, Yandex sharply changed its citation algorithm. Figure 5 makes this change visible by plotting the number of 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 dropped to 10 and stayed around 10 for nearly a year. We see a slight increase in the number of outlets cited by Yandex by the end of 2017, although it still stayed lower than in the period before the signing of the law. [14].

This drop in the number of unique outlets cited by Yandex increased the probability that the same outlet is referenced multiple times for different news in the top-5. Figure 4 presents the evolution of the average number of slots occupied by the most cited outlet. There is a visible increase in the probability that the same outlet occupies multiple slots after the law on news aggregators was adopted in mid-2016, with an average of 2.13 slots after that.

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

[14]The number of outlets appearing in the is lower than 15, which was mentioned in a journalistic investigation as the number of outlets that were pre-approved by the administration of President Putin `https://meduza.io/feature/2022/05/05/my-zamuchilis-borotsya`

Figure 4: The Number of Top-5 Slots Occupies by the Most Cited Outlet



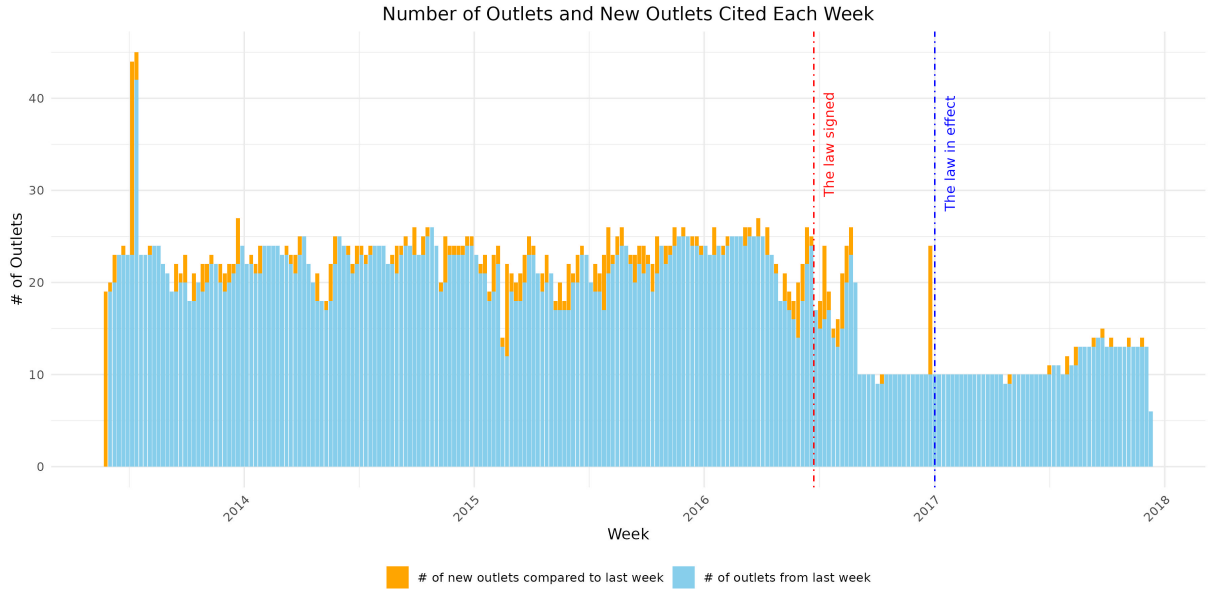
Dots represent a weekly average of the number of top-5 slots occupied by the most cited outlet at that moment. For instance, the value of two means that in an average snapshot, two out of five slots are occupied by the same outlet (e.g., ria.ru).

independent, both with a specialization in business news (RBC and Vedomosti)[15]. 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 6).

Did the removal of independent outlets also change the composition of news events that Yandex cited? Figure 7 plots the share of sensitive events—those systematically under-reported by government-controlled outlets—both in the overall news and in Yandex's top-5 citations. Throughout the sample period, including after the law was passed, Yandex's share of sensitive events remains broadly in line with the market average. Thus, despite the sharp reduction in citations of independent outlets documented in Figure 6, the algorithm change did not directly

[15]Note that our classification is based on the situation in the beginning of the sample period. Coincidentally, RBC has changed the main editors who were considered to be too independent in May 2016 and changed ownership in 2017; Vedomosti changed ownership and the editorial team in 2020.

Figure 5: 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 was signed. The blue vertical line corresponds to January 1, 2017, when the law on news aggregators came into effect.

suppress sensitive topics — it operated primarily through the removal of outlets rather than the selection of events.

6.2 Changes in the Algorithm

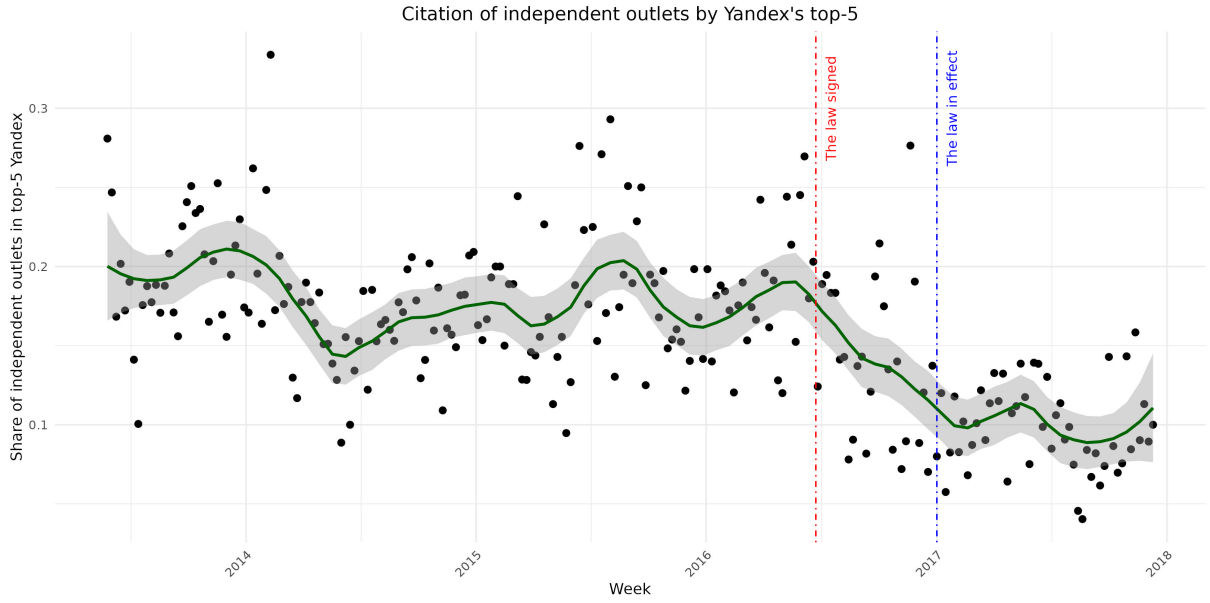
As we saw above, Yandex can exercise a lot of discretion over which news outlets it references in the links in the top-5 section. A natural implication of such discretion is that Yandex is potentially able to refer traffic to news outlets of its choosing. To measure more precisely the potential implications of changes in the citation algorithm and quantify scenarios, we now proceed to reverse-engineer the news aggregation and citation algorithm employed by Yandex and how it changed after the adoption of the law.

6.2.1 First Step: Ranking of News Events

Although the details of Yandex's news aggregation and citation algorithm are not public, Yandex disclosed the general framework and main features used as input to the algorithm [16]. The

[16]The description can be found on the blog of Yandex `https://yandex.ru/blog/company/76641`. In particular, we can use the same parser that is used to analyse texts by Yandex `https://github.com/yandex/tomita-parser/`.

Figure 6: 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 was signed. The blue vertical line corresponds to January 1, 2017, when the law on news aggregators came into effect.

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.

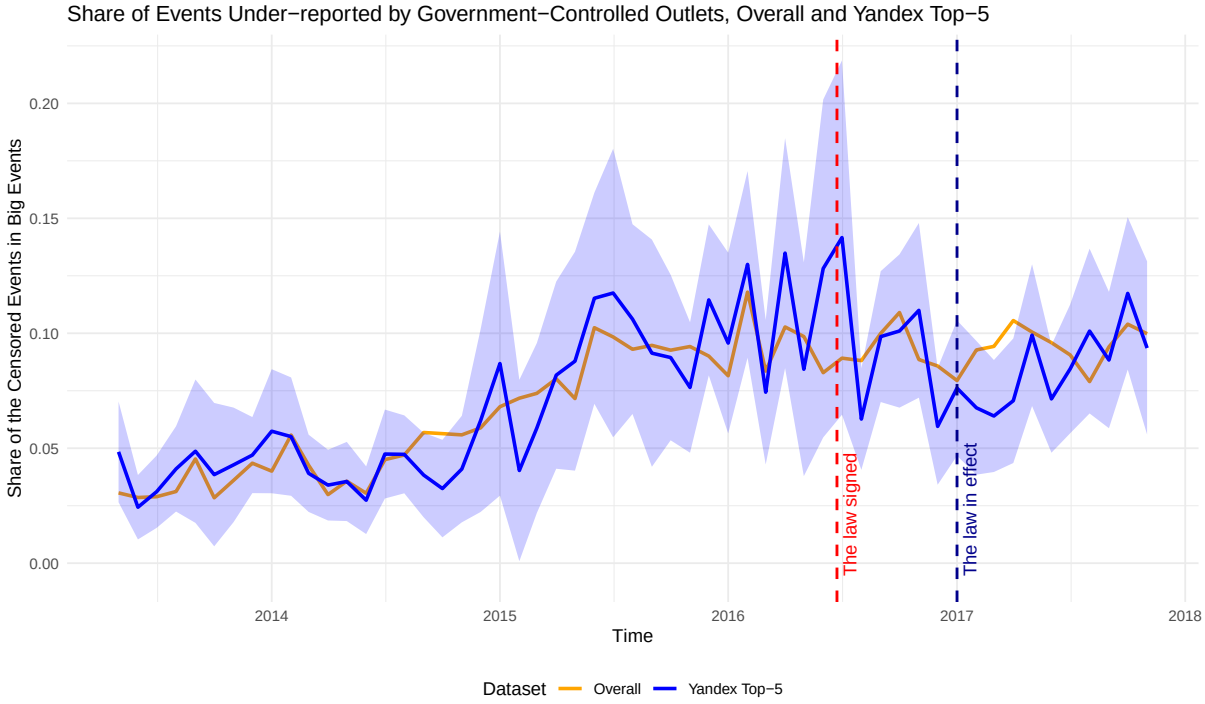
We estimate Yandex's algorithm using the data on news events and their citations. Every time period t, there is a set of news events E_t that are realized. A news event $e_i t$ is characterized by its newsworthiness, n_e, which is not observed by the researcher. A news outlet j chooses the events to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

$$s_{et} = \boldsymbol{X'_{ejt}}\boldsymbol{\beta} + \epsilon_{et} \tag{2}$$

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.

Figure 7: Share of Events Under-Reported by Government-Controlled Outlets



Each point corresponds to a monthly average share of events that are under-reported by the government-controlled outlets (compared to independent outlets). Shaded areas are 95% confidence intervals, computed based on the variation of daily shares in that month. 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 was signed. The blue vertical line corresponds to January 1, 2017, when the law on news aggregators came into effect.

We use the realized rankings of news events on Yandex and the underlying set of events to estimate parameters in the above scoring model. Table 7 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. Note that the regression fit is relatively high, with three variables explaining 16.8% of the variation in Yandex's citation decisions. In contrast, day (column 2) and news topic fixed effects (column 3) add little explanatory power.

Next, we formally test whether Yandex's algorithm treats sensitive events differently (see Columns (4) and (5) of Table 7). We find that, even before the law, sensitive events are 5.8 percentage points less likely to be cited in Yandex's top-5, keeping everything else constant. After the law, this probability falls by an additional 2.7 percentage points (column 4). However, this post-law drop is entirely explained by outlet fixed effects (column 5), confirming the pattern in Figure 7: the reduction in sensitive coverage operates through the removal of independent outlets, not through direct suppression of sensitive topics.

Overall, the results show that Yandex's algorithm already under-reported sensitive events

Table 7: Factors Correlated with Event Citations by Yandex

Dependent Variable:	I(Event in Yandex Top-5)				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Constant	-0.0404***				
	(0.0008)				
# of outlets	0.0134***	0.0133***	0.0133***	0.0136***	
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	
# of articles	0.0013***	0.0013***	0.0013***	0.0013***	0.0011***
	(7.33×10^{-5})	(7.33×10^{-5})	(7.32×10^{-5})	(7.33×10^{-5})	(6.49×10^{-5})
Share of coverage in first hour	0.0148***	0.0146***	0.0127***	0.0129***	0.0138***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Censored by controlled outlets				-0.0584***	-0.0691***
				(0.0057)	(0.0056)
Censored by controlled outlets x After law				-0.0267***	-0.0035
				(0.0097)	(0.0088)
Fixed-effects					
Day		Yes	Yes	Yes	Yes
Topic			Yes	Yes	Yes
Outlet					Yes
Fit statistics					
Observations	1,794,010	1,794,010	1,794,010	1,794,010	1,794,010
R^2	0.16835	0.17719	0.17861	0.17965	0.20354
Within R^2		0.16850	0.16479	0.16586	0.19015

*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.

before the law, and the further decline after the law was driven by changes in outlet weights rather than topic-level filtering.

6.2.2 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.

Formally, 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 the outlet. Based on the public information disclosed by Yandex and our interviews with media professionals who worked at news outlets at the time, in choosing the article 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.

Table 8: 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	-0.1611*	-0.1606*	0.0895***
	(0.0890)	(0.0889)	(0.0323)
log(hours published after first event article)	-0.0165***	-0.0160***	-0.0039**
	(0.0040)	(0.0040)	(0.0015)
# of words in event articles by j	-2.43×10^{-5}**	-2.45×10^{-5}**	-2.02×10^{-5}***
	(1.07×10^{-5})	(1.07×10^{-5})	(5.61×10^{-6})
# of named entities in event articles by j	0.0019***	0.0019***	0.0004***
	(0.0006)	(0.0006)	(0.0001)
Title length (symbols)	-0.0006**	-0.0006**	-0.0007***
	(0.0003)	(0.0003)	(0.0002)
# of punctuation marks in the title	-0.0049**	-0.0049**	-0.0025**
	(0.0019)	(0.0019)	(0.0010)
Title-to-text distance	-0.0964***	-0.0944***	-0.0247***
	(0.0260)	(0.0261)	(0.0078)
I(j published the earliest article)	0.0828***	0.0776***	0.0424***
	(0.0295)	(0.0272)	(0.0085)
Difference from the first publication	0.0724***		
	(0.0162)		
Difference from the first publication: 0-20%		-0.0265***	-0.0090*
		(0.0068)	(0.0046)
Difference from the first publication: 20-40%		-0.0120***	-0.0034***
		(0.0032)	(0.0011)
Difference from the first publication: 60-80%		0.0103***	0.0025*
		(0.0032)	(0.0013)
Difference from the first publication: 80-100%		0.0140***	0.0035
		(0.0048)	(0.0021)
Fixed-effects			
Event	Yes	Yes	Yes
Outlet			Yes
Fit statistics			
Observations	778,348	778,348	778,348
R^2	0.08466	0.08452	0.16423
Within R^2	0.01682	0.01667	0.00396

*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. Standard errors clustered on the outlet level.

To estimate factors that affect the choice of particular articles withing a given news event, we estimate the following model:

$$c_{je} = \tilde{\boldsymbol{X}}'_{\boldsymbol{ej}}\tilde{\boldsymbol{\beta}} + \gamma_e + \alpha_j + \tilde{\epsilon}_{ej} \tag{3}$$

where c_{je} is a dummy variable for whether the article on event e by outlet j made it into top-5, 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.

The estimates of from equation 3 are presented in Table 8. 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),[17] 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.[18] 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.

Finally, in Column (3) of Table 8 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 8.5% to 16.5%. This suggests that while the characteristics of coverage are important in Yandex's citation decisions, outlet weights also play an equally 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:

$$Pr(\text{Outlet } j \text{ in Yandex Top-5}_e | j \in J_e) = \frac{exp(\alpha_j + \alpha'_j I(\text{post-law})_e + \tilde{X}'_{ej}\tilde{\beta}}{\sum_{j' \in J_e} exp(\alpha_{j'} + \alpha'_{j'} I(\text{post-law})_e + \tilde{X}'_{ej'}\tilde{\beta}} \tag{4}$$

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 3.

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

[18] 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 9: 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	10.20	0.04 (0.03)	0.01 (0.00)	-0.03 (0.03)
Independent	26.53	28.89 (0.43)	15.90 (0.36)	-12.99 (0.57)
Potentially-Influenced	42.86	30.85 (0.37)	37.11 (0.47)	6.26 (0.65)
Government-Controlled	20.41	40.23 (0.35)	46.99 (0.50)	6.76 (0.64)

Implied probabilities of references based on outlet weights α_j and α'_j from equation 4. 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 9 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 A10 in Appendix A4 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 9 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 43%.

The outlets' weights change dramatically after the law on news aggregators is signed. Independent outlets, which were not penalized at all before the passing of the law, experience the greatest change, with their probability of references decreasing from 28.9% to 15.9%. 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.[19]

The estimates of outlet weights confirm the suggestive results in Figures 5 and 6: 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 its algorithm and decide to promote or penalize a particular set of outlets. We next explore the implications of such control of the algorithm.

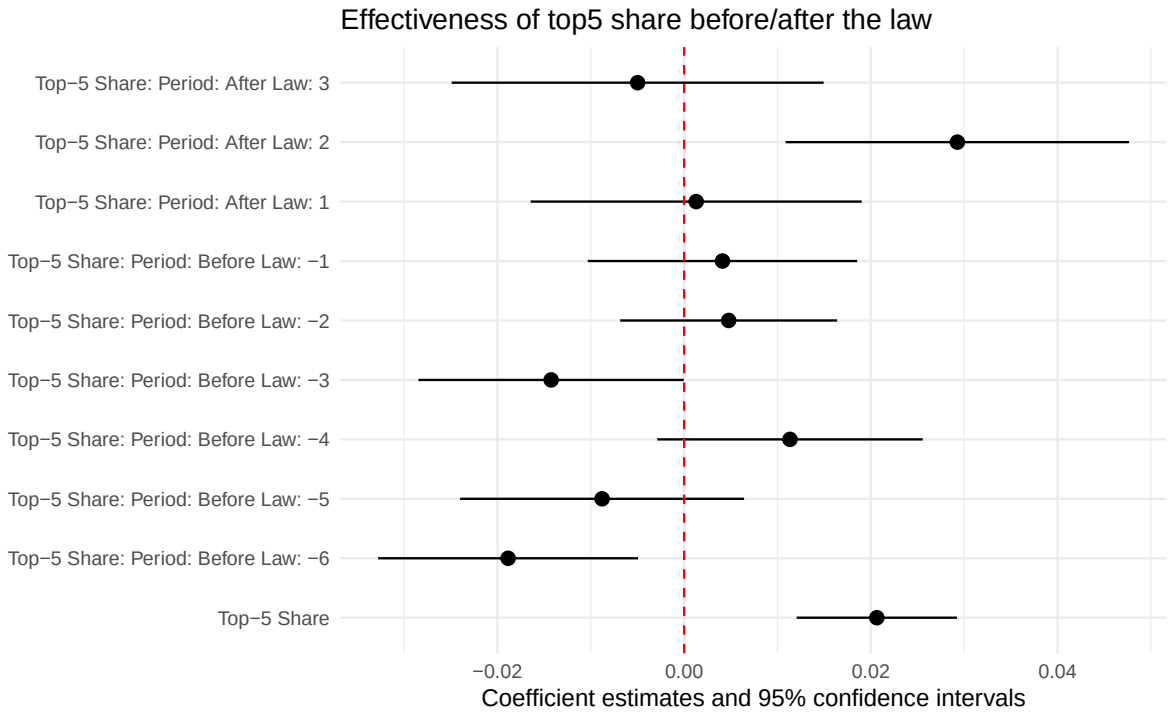
6.3 Changes in Effectiveness

The ability of a news aggregator to affect news consumption by changing its algorithm may be constrained by demand factors. If the users are sensitive to the deviations from the optimal

[19]See Table A10 in Appendix A4 for outlet-level results.

suggestion algorithm, they should react to changes and become less responsive to the news aggregator's suggestions. 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 in whether the effectiveness of the top-5 references has decreased because Yandex restricted the outlet sample.

Figure 8: 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 for six months periods. Error bars correspond to the 95% confidence intervals. .

Figure 8 presents the estimates of Yandex's effectiveness in referring traffic based on semi-annual periods before and after the law on news aggregators. Overall, referencing a particular outlet in all top-5 links for a full-day allows Yandex to send around 2% of its front page visitors to the outlet (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.

Thus, the results show that reducing the number of outlets that Yandex references did not decrease its ability to direct traffic to news outlets, indicating that the ability of the news aggregator to channel traffic to particular outlets was not constrained by demand factors.

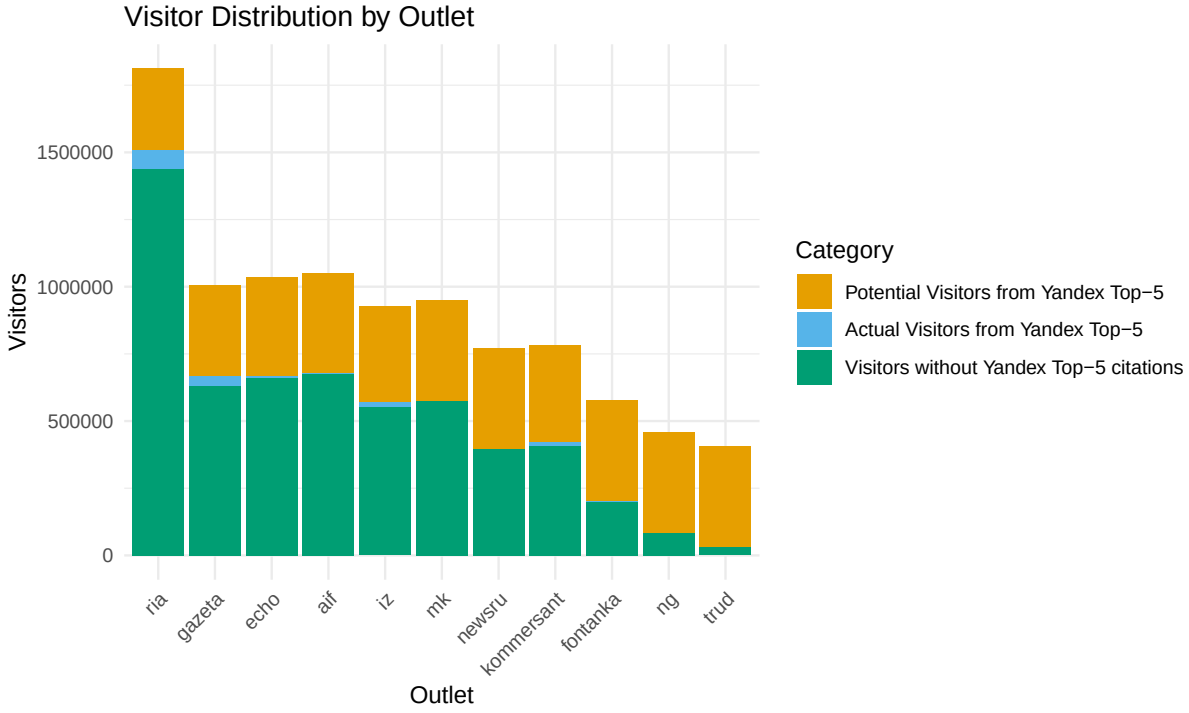
6.4 Counterfactual References of Yandex

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 quantify theimpact 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 9 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 have received if Yandex had referenced it all the time. Under the current algorithm, news outlets get from close to zero (e.g., fontanka.ru) to 5.89% (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 26% (for ria.ru, the largest outlet in this sample) to 1132% (for trud.ru, the smallest outlet).

Figure 9: Counterfactual Traffic to News Outlets: The Choice of Events is Unrestricted

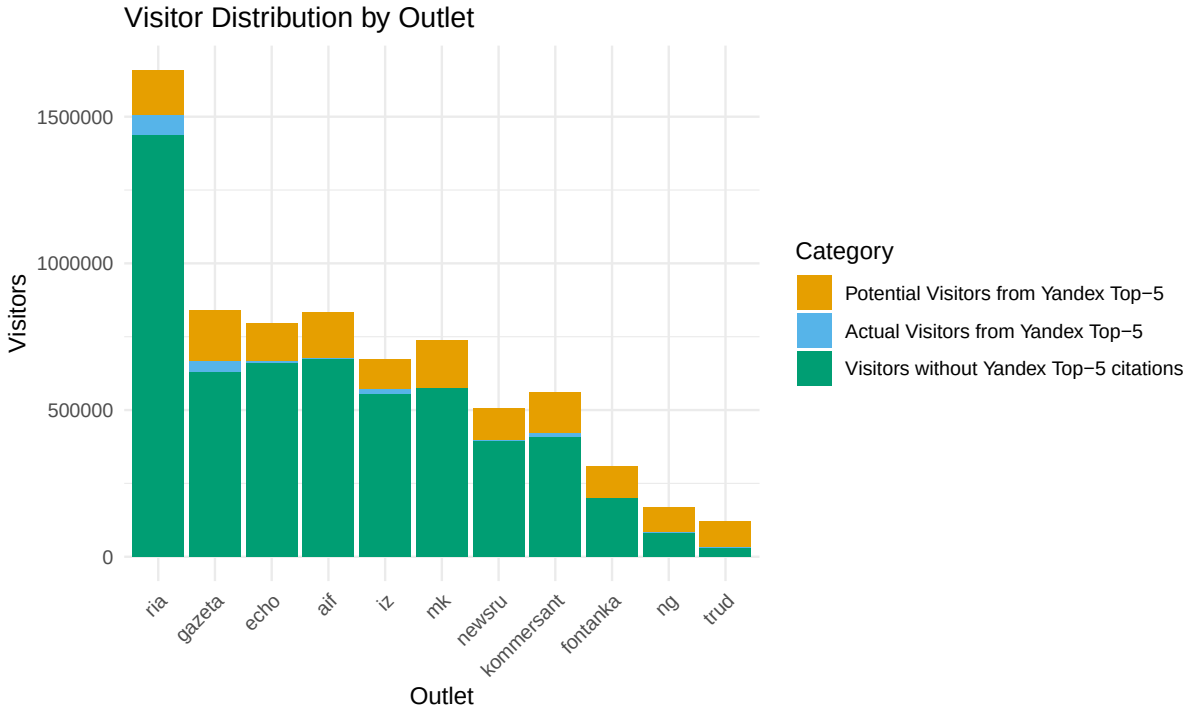


Visitors are predicted using the estimates of regression 1 with the traffic level as an outcome variable.

The scenario above assumes that Yandex has a full discretion 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.

Figure 10: Counterfactual Traffic to News Outlets: The Choice of Events is Fixed



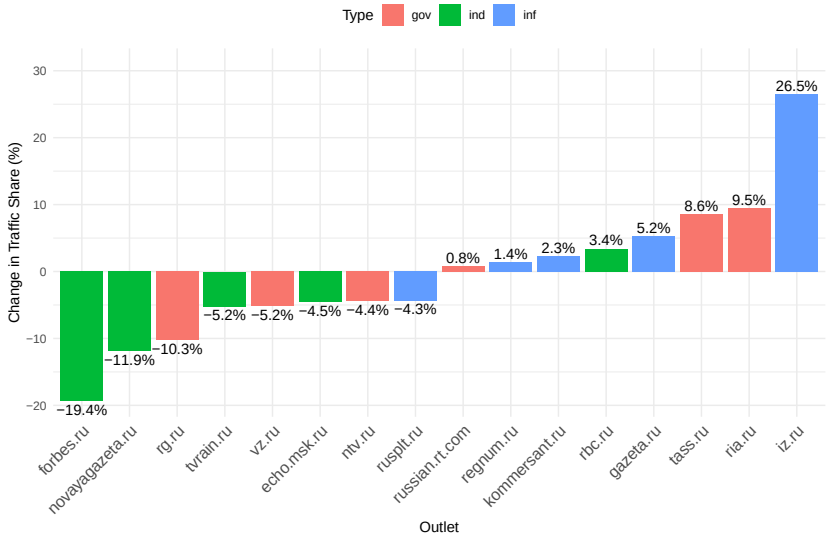
Visitors are predicted using the estimates of regression 1 with the traffic level as an outcome variable.

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 10 presents the results. 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 15% (ria.ru) to 270% (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.

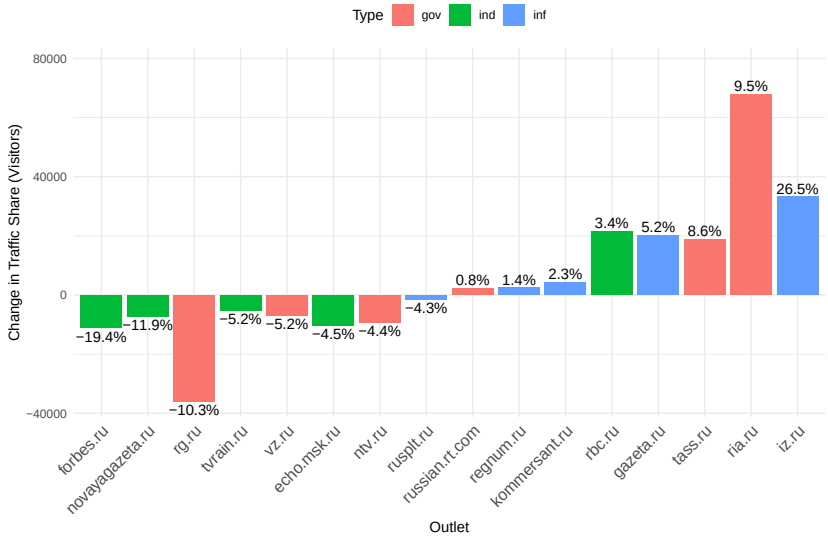
Finally, we evaluate the effect of the change in Yandex's algorithm after the 2016 law on news aggregators described in Section 6. For this, we use a separate dataset covering the traffic to all 50 news outlets in our sample in 2017, collected from SEMRush.[20] Figure 11 presents changes in traffic to top-8 "winners" and "losers" due to algorithm changes. The independent outlet Forbes.ru loses the highest share of its traffic – around 20% – while a potentially-influenced

[20] As mentioned in Section 3, this dataset does not contain traffic measures before 2017. We use our treatment estimates from section 5.2 to measure the effects of referrals on traffic.

Figure 11: Changes in Traffic to News Outlets due to Algorithm Change



Subfigure (a): Traffic in Shares (%)



Subfigure (b): Traffic in Levels (%)

Visitors are predicted using the estimates of regression 1 with the traffic level as an outcome variable. Baseline traffic is taken from SEMRush dataset in 2017.

outlets iz.ru gains around 27%. In terms of the levels of traffic (Subfigure b), the outlet that gains the most is ria.ru, the largest news agency that is government-owned, which gains around 62 thousand daily visitors. Aggregating the effects across all 50 outlets, total traffic to independent outlets dropped by 0.9%, traffic to government-controlled outlets increased by 0.5%, and traffic to potentially influenced outlets increased by around 1.5% due to the change in the algorithm.

6.5 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 explained how they have adjusted their news to increase the chances of getting referenced by Yandex.

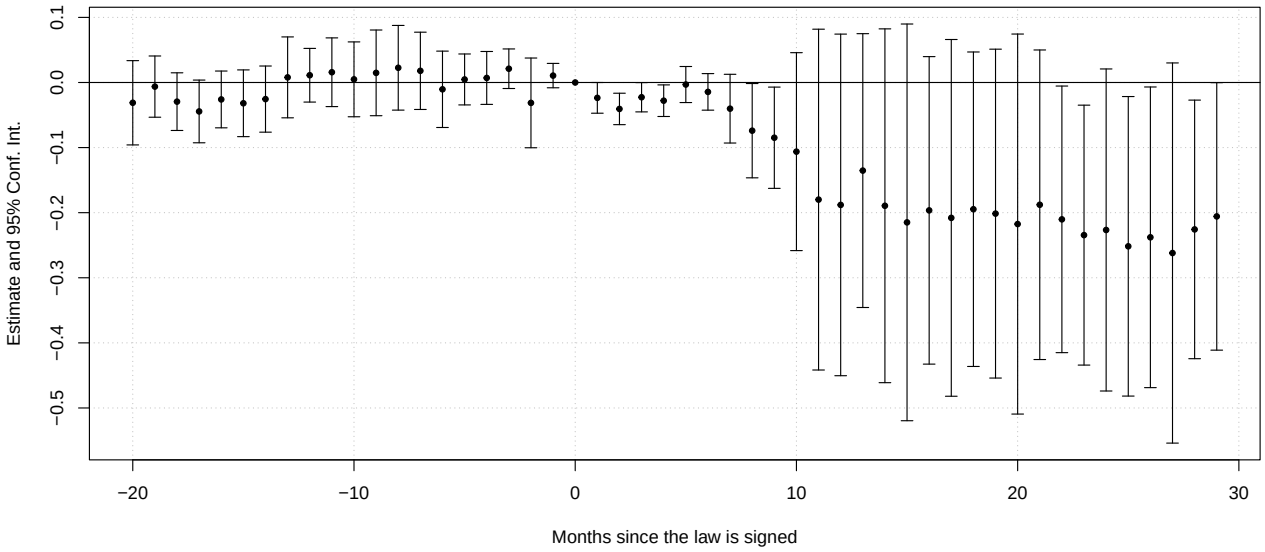
Such attention of news outlets to Yandex's algorithm and the traffic they receive implies that news outlets can change their production decisions in response to a change in the algorithm. We test for such effects empirically by leveraging the change in Yandex's algorithm that happened in August 2016 in response to the law on news aggregators. As we show in Subsection 6.1, 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 the news outlets.

We examine several potential decisions of outlets – a consequential decision on whether to produce shorter "spot news" on events that are covered by other outlets to improve their chance of being cited by aggregators or differentiate towards longer "feature" articles on more unique topics – and inconsequential decisions on the length of articles' titles and the use of punctuation symbols in the titles. Figures 12 and 13 present the event study plots with estimates of the effect of Yandex's algorithm change on these two dimensions of news production. Compared to the outlets still cited by Yandex, the outlets that were down-weighted by the algorithm were 20% less likely to write news articles classified into events and write 50% longer news articles (Figure 12), 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 13). It took around 10 months for the news outlets to adjust their production decisions to a new level.[21]

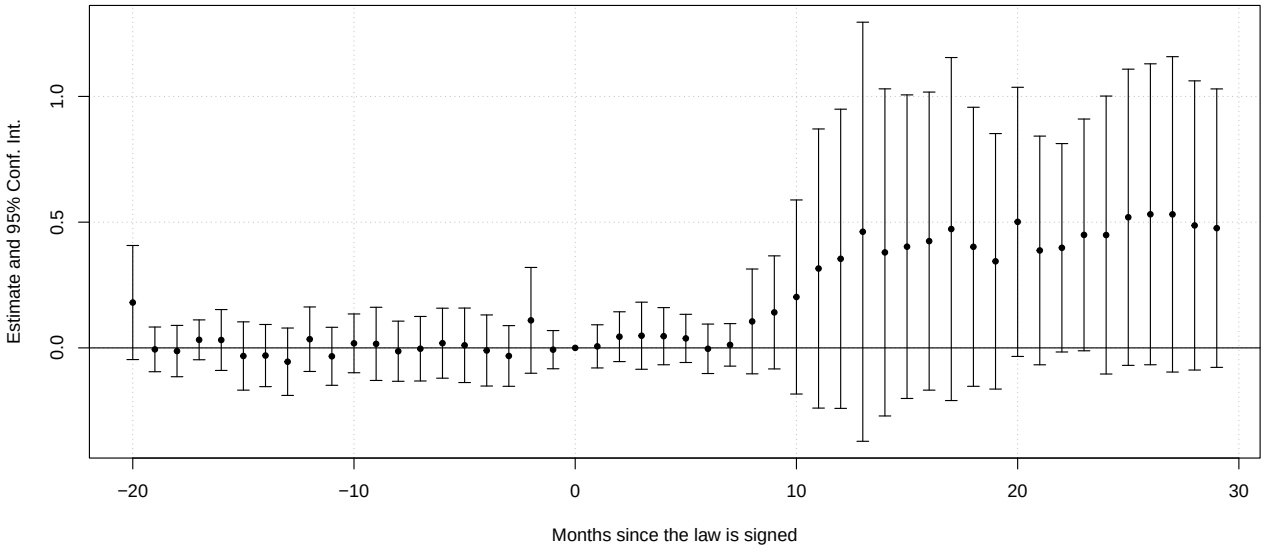
As a result, we see some evidence of the upstream outlets' adjustments of news production in response to Yandex's change of the algorithm. This change is interesting since it affects all the readers of the adjusting news outlets. This implies that Yandex's algorithm changes can affect not only marginal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

[21] 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).

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



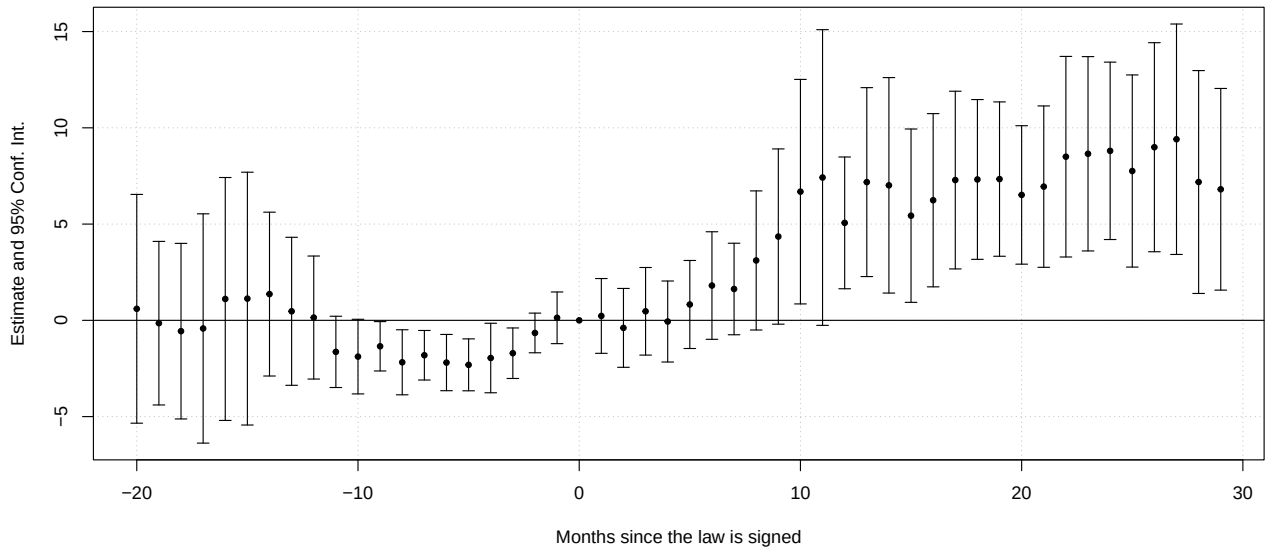
(a) Share of News Classified into Events



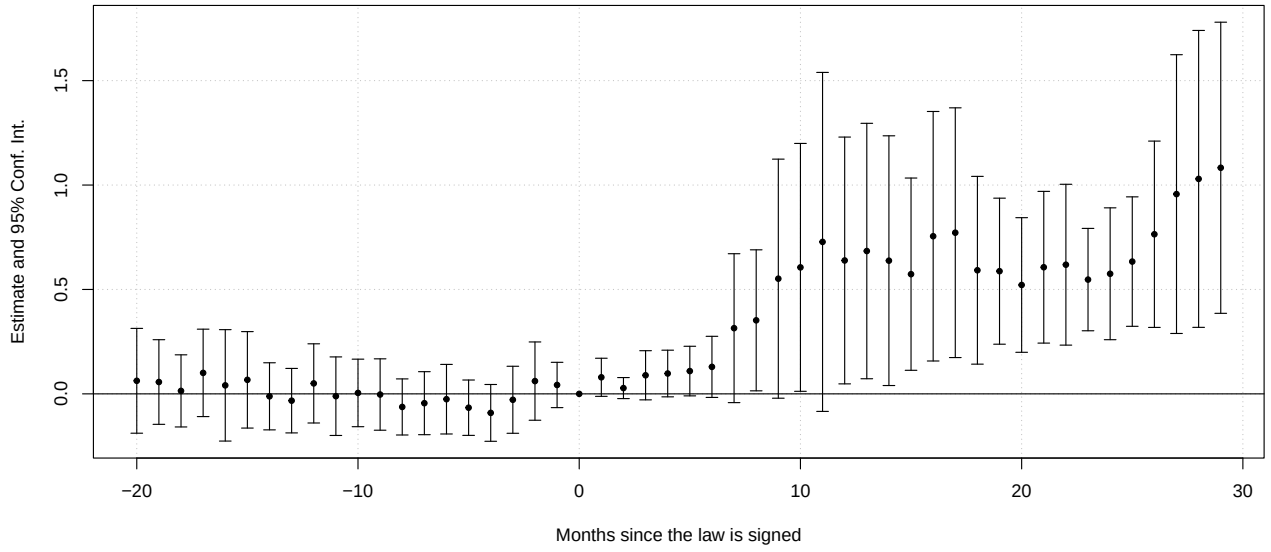
(b) Average Number of Words

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 13: 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.

news outlets, who now get different news coverage from their preferred media.

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*.

Braghieri, L., Eichmeyer, S., Levy, R., Mobius, M., Steinhardt, J., and Zhong, R. (2025). Article level slant and polarization of news consumption on social media. *Working Paper*.

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.

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.

Durante, R. and Knight, B. (2012). Partisan control, media bias, and viewer responses: Evidence from berlusconi's italy. *Journal of the European Economic Association*, 10(3):451–481.

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

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

Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

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.

Knight, B. and Tribin, A. (2019). The limits of propaganda: Evidence from chavez's venezuela. *Journal of the European Economic Association*, 17(2):567–605.

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.

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

Reuters (2024). Overview and key findings of the 2024 digital news report. `(https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/dnr-executive-summary` [Acccessed: December 10, 2025].

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 Traffic Data Validation

There are four sources of daily traffic data of the Russian media market:

- `top.mail.ru` – a traffic counter which provides data for a subset of 13 relevant outlets since 2010. We use it as the main source for our estimates.
- `semrush.com` – a service that approximates websites' daily audience using third-party data and its own algorithms. It provides data on all 50 outlets studied, available since 2017. We use this source to estimate the average number of visitors for all outlets.
- liveinternet.ru – a traffic counter from which we have collected data on a small subset of outlets since 2022.
- `top100.rambler.ru` – another traffic counter from which we have collected data since 2020.

We use `top.mail.ru` as the primary data source for our estimates, because it is the only one that provides data for periods earlier than 2017. In this section we use overlapping time periods and websites across different data sources to validate the quality of the data provided by `top.mail.ru`.

Table A1 presents the number of overlapping websites across different data sources, while Table A2 provides detailed information for each pair of sources.

Table A1: Overlap of Websites Across Traffic Data Sources

	mail	semrush	liveinternet	rambler
mail	34	13	13	20
semrush	13	50	23	28
liveinternet	13	23	84	33
rambler	20	28	33	895

Table A2: Websites Shared Between Pairs of Data Sources

Data Sources	**Overlapping Websites**
mail & semrush (13)	aif.ru, echo.msk.ru, fontanka.ru, gazeta.ru, iz.ru, kommersant.ru, lenta.ru, mk.ru, newsru.com, ng.ru, rg.ru, ria.ru, trud.ru
mail & rambler (20)	9tv.co.il, aif.ru, argumentiru.com, echo.msk.ru, fedpress.ru, fontanka.ru, gazeta.ru, kommersant.ru, lenta.ru, m24.ru, newsru.co.il, newsru.com, ng.ru, pln-pskov.ru, portal-kultura.ru, ren.tv, rg.ru, ria.ru, tvzvezda.ru, ura.ru
mail & liveinternet (13)	aif.ru, echo.msk.ru, fontanka.ru, gazeta.ru, kommersant.ru, lenta.ru, m24.ru, mk.ru, ren.tv, rg.ru, ria.ru, tvzvezda.ru, ura.ru
semrush & rambler (28)	aif.ru, bfm.ru, echo.msk.ru, fontanka.ru, forbes.ru, gazeta.ru, kommersant.ru, kp.ru, lenta.ru, life.ru, newsru.com, newtimes.ru, ng.ru, ntv.ru, rbc.ru, regnum.ru, rg.ru, ria.ru, rosbalt.ru, rusplt.ru, russian.rt.com, snob.ru, sobesednik.ru, svpressa.ru, tass.ru, tvrain.ru, utro.ru, vesti.ru
semrush & liveinternet (23)	aif.ru, bfm.ru, dni.ru, echo.msk.ru, fontanka.ru, gazeta.ru, kommersant.ru, kp.ru, lenta.ru, mk.ru, novayagazeta.ru, rbc.ru, regnum.ru, rg.ru, ria.ru, ridus.ru, rosbalt.ru, svpressa.ru, tass.ru, tvrain.ru, utro.ru, vedomosti.ru, znak.com
liveinternet & rambler (33)	aif.ru, bfm.ru, echo.msk.ru, expert.ru, finmarket.ru, fontanka.ru, gazeta.ru, govoritmoskva.ru, inosmi.ru, kazanfirst.ru, kommersant.ru, kp.ru, lenta.ru, m24.ru, mir24.tv, moslenta.ru, rbc.ru, regnum.ru, ren.tv, rg.ru, ria.ru, rosbalt.ru, russian7.ru, sm-news.ru, smi2.ru, svpressa.ru, tass.ru, tvrain.ru, tvzvezda.ru, ura.ru, utro.ru, vestinn.ru, wek.ru

Thus, for some websites and days, we observe the number of unique visitors reported by more than one source. We remove observations for which the number of visitors is close to zero, because this usually indicates that the data source does not have reliable information on the traffic for that time period. We then build regressions predicting the number of visitors to an

outlet j on day d from data source S, based on the number of visitors reported by `top.mail.ru` and the fixed effects included in all our models.

$$y_{jd}^{S} = a y_{jd}^{top.mail.ru} + \rho_{jw} + \delta_{j,\text{wd}} + \theta_d + \xi_{jd} \tag{A1}$$

where y_{jd}^{S} is the total number of visitors to outlet j on day d reported by data source S. 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 reported by source S.

Table A3 presents the model estimates. The R-squared of the projected model using traffic data from `semrush.com` is relatively low, as `semrush.com` is not a traffic counter that directly observes and reports real visitors numbers, but rather a service that estimates traffic using proprietary algorithms. In other cases – particularly with liveinternet.ru – we observe a high correlation between reported traffic, which supports the reliability of data from `top.mail.ru`.

We nevertheless use data from `semrush.com` to approximate the magnitude of daily traffic for each outlet. Table A4 compares average daily traffic estimates from two sources.

Table A3: Validation of top.mail.ru Data using Other Data Sources

	Dependent variable:		
	$y_{jd}^{semrush}$	$y_{jd}^{liveinternet}$	$y_{jd}^{rambler}$
	(1)	(2)	(3)
$y_{jd}^{top.mail.ru}$	0.336***	0.979***	0.578***
	(0.019)	(0.013)	(0.036)
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	29,002	8,650	16,109
R^2	0.982	0.994	0.985
Adjusted R^2	0.976	0.992	0.981
R^2 (Proj. model):	0.23	0.728	0.458
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A4: Outlets Average Daily Visitors in 2018

Outlet	top.mail.ru		semrush.com	
Outlet	Avg Daily Visitors	Ranking	Avg Daily Visitors	Ranking
ria.ru	3426205	1	3205213	1
mk.ru	1641932	2	1317274	2
rg.ru	1034156	3	740947	4
aif.ru	897314	4	1018081	3
echo.msk.ru	804749	5	715925	5
kommersant.ru	761327	6	474322	7
iz.ru	648264	7	561116	6
newsru.com	294035	8	446458	8
fontanka.ru	281809	9	191924	9
ng.ru	103508	10	76333	10
trud.ru	17655	11	14764	11

We compare the numbers for 2018, as `semrush.com` only began accounting for mobile traffic that year. Outlets "gazeta.ru" and "lenta.ru" aare omitted from the table because they stopped using `top.mail.ru` as a traffic counter before 2018.

A0.1 Total Yandex Visitors

To measure the effectiveness of Yandex recommendations, we use the daily number of Yandex search engine users as a proxy for the number of visitors to Yandex's main page. We have two data sources for these statistics:

- `top.mail.ru` – a source that does not track the total number of search engine users, but aggregates data from many websites that use it as a traffic counter. It provides data since 2011.

- `radar.yandex.ru` – the official Yandex platform, which provides data since 2015.Currently, the service reports only daily visit counts for the Yandex search engine, However, it used to publish statistics on visits to the Yandex main page before 2016. Archived versions of these metrics can still be found online, for example in web articles such as `tokar.ua/read/9412/iandeks-yznutry-statystyka-yly-kak-kom/`.

We use `top.mail.ru` as our primary source, as it allows us to analyze earlier time periods and different audience segments. We assess the robustness of the daily variation in our measure using the `radar.yandex.ru` search engine visits statistics. In addition, we use archived data on visits to the Yandex main page to quantify the extent to which our proxy understates total traffic and, consequently, overstates the effectiveness of the Yandex top-5.

Table A5 presents model estimates that validate the data from `top.mail.ru`. We observe that the values reported by `radar.yandex.ru` are systematically higher. This is mostly explained

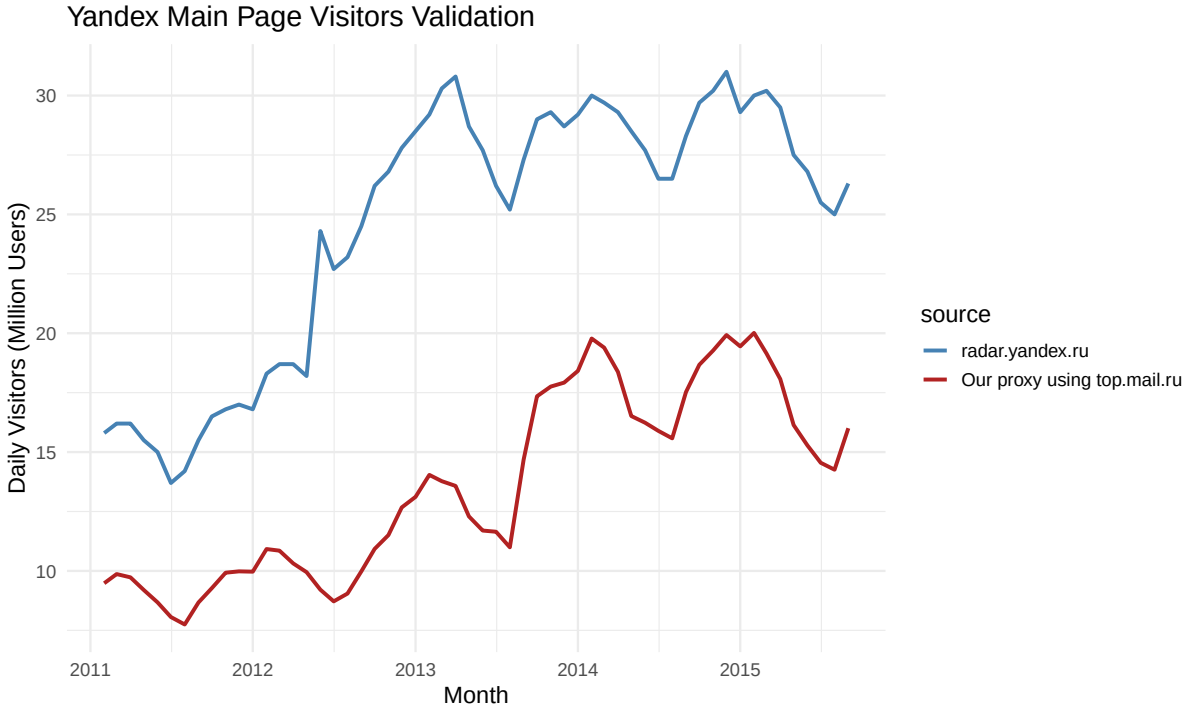
by the fact that `top.mail.ru` captures visitors, while `radar.yandex.ru` provides statistics on the number of visits. Another reason is that `top.mail.ru` doesn't capture all Yandex users – only those visiting websites that have chosen `top.mail.ru` as their traffic tracker. The most important aspect for our analysis is the correlation between the two data sources, and we find that the R-squared values are very high.

Figure A1 compares the daily number of Yandex main page visitors reported by `radar.yandex.ru` with our proxy based on `top.mail.ru`. Two series exhibit a correlation of 0.82, and the numbers from `radar.yandex.ru` are on average 1.85 times higher. This gap arises because the proxy constructed from `top.mail.ru` captures only those Yandex users who visited at least one site tracked by `top.mail.ru` on a given day – that is, a more active subset of users. Using this comparison, we adjust our estimates of the Yandex top-5 effectiveness to account for less active users as well. As shown in Figure 8, the Yandex top-5 section directs approximately 2% of "more active" main-page visitors, or about 1.1% of all visitors.

Table A5: Validation of `top.mail.ru` Yandex Users with `radar.yandex.ru` Data

	Dependent variable:
	$y_{jd}^{radar.yandex.ru}$
$y_{jd}^{top.mail.ru}$	8.287***
	(0.149)
Fixed Effects:	
- Week	Y
- Week Day	Y
Clustered S.E.	Weeks
Observations	15,327
R^2	0.990
Adjusted R^2	0.990
R^2 (Proj. model):	0.897
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure A1: Yandex Main Page Visitors Validation



A1 Foundation of Public Opinion Survey Data

To describe individuals who can be more effectively directed by Yandex Top-5, we use data from the 2011 survey conducted by Foundation of Public Opinion[22], one of leading independent sociological organizations in Russia.

The dataset includes demographic characteristics of 56,900 individuals, along with their responses to questions related to political beliefs. We aggregate these responses by demographic groups – specifically age, gender, and region – to align with our Yandex's Top-5 Effectiveness estimates.

For the regression analysis, we construct the X variables by calculating the share of respondents in each demographic group who chose particular answer options for each survey question. Table A6 presents the variable names, the full survey questions (translated into English), and the corresponding answer options whose shares were extracted.

Only respondents with often internet usage are taken for computing all X's, except for "Low internet usage in demo".

Table A6: Survey questions and answer options used to construct regression variables.

Variable Name	Full Question	Answer option
Gov. party supporters	If in December 2011 you participate in the State Duma elections, which party are you most likely to vote for?	"United Russia (Dmitry Medvedev)"
Low income	Which statement best describes the material situation of your family?	"We do not have enough money even for food" OR "We have enough money for food, but cannot buy clothes or shoes"
Likely voters	Please look at the card and indicate: will you personally participate in the State Duma elections in December 2011?	"I will definitely participate in the elections" OR "I will most likely participate in the elections"
Likely protesters	Do you personally allow or exclude the possibility of participating in any protest actions?	"I allow"
See corruption as an issue	In your opinion, which of the tasks listed on the card should the president and the new government focus on the most? Please select no more than five tasks.	"Fight against corruption"
Support democracy	In your opinion, which of the tasks listed on the card should the president and the new government focus on the most? Please select no more than five tasks.	"Protection of human rights and civil liberties" OR "Development of civil society and local self-government" OR "Development of democratic institutions and political competition"
Low internet usage in demo	Have you personally used the Internet? If yes, when did you last use it?	ALL ANSWERS EXCEPT "In the last day" AND "In the last week"

[22] https://fom.ru/

A2 News Events Detection Algorithm

We use the Topic Detection and Tracking (TDT) algorithm developed by Allan et al. (1998) and used in the economic literature (e.g. see Cagé et al., 2020) to cluster articles into news events.

The approach consists of four major steps:

- **Preprocessing** – we remove HTML tags, perform stemming, and tokenize the text.
- **Vectorization** – we combine article titles and body texts and transform them into numeric vectors. Since there are multiple methods for text vectorization, we test several approaches to select the most suitable one and check for the robustness of our results. We focus on several widely used methods: TF-IDF, Word2Vec and Word2Vec weighted with TF-IDF.
- **Clusterization** – in TDT, the clustering process takes into account both the vector representations and the publication time of articles. We use a three-day sliding window, and within each such period we cluster articles based on their cosine similarity. If there is a news article that belongs to two different clusters across the overlapping time periods, we we merge these clusters into a single news event. If no new similar news articles are published the next day, then the news event is finalized.
  There are several hyperparameters that must be selected at this stage. First, we compare different clustering algorithms suitable for our task, including a density-based approach and agglomerative clustering.
  Second, for each of the algorithms, we test different radius values that define clusters size by setting a threshold for cosine similarity.
  We keep the sliding window size (3 days) and the overlap period (1 day) unchanged.
- **Postprocessing** – we remove duplicate articles from the results. An article is considered a duplicate if two or more observations published by one outlet clustered into the same news event with a very small radius[23].

  We also exclude all articles published by Ukrainian outlets (which were initially included to our dataset to help find the best clustering settings). After these adjustments, we say that all clusters with only one article are not news events and change their status to "unclustered".

[23] We apply the DBSCAN algorithm with a radius parameter of 0.001 to cluster TF-IDF vectors of news articles. To validate the robustness of this approach, we compare the results with Agglomerative Clustering applied to the same TF-IDF vectors. The two methods show a high degree of agreement: 99.92% of the articles identified as potential duplicates by DBSCAN are also flagged by Agglomerative Clustering, while 100% of the duplicates detected by the Agglomerative approach are also found by DBSCAN. We rely on TF–IDF vectors instead of word embeddings, since our goal is to identify near-duplicates by exact word overlap rather than by semantic similarity.

Additionally, we "uncluster" 1063 events which are longer then half a month (15 days), because such events usually correspond to recurring reports rather than real events. For example, weather reports, daily war reports, or horoscopes (Table A7).

Table A7: Example of Removed Longest Events

Event ID	Length (days)	Sample Titles
2885230	1311	Гидрометцентр: В Москве 5 августа возможны дожди с грозами *Hydrometcenter: In Moscow on August 5, rain with thunderstorms is possible* В Украине сегодня без осадков, местами жара до +34° (карта) *In Ukraine today no precipitation, in places heat up to +34° (map)* В аэропортах Москвы отменено около 80 авиарейсов и около 60 задержано *About 80 flights canceled and 60 delayed at Moscow airports* МЧС сделало метеопредупреждение на 6 октября *EMERCOM issued a weather warning for October 6* В Липецк ненадолго вернутся ночные заморозки *Night frosts will briefly return to Lipetsk*
4523211	236	К перемирию в Сирии присоединились уже 972 населенных пункта *972 settlements have already joined the ceasefire in Syria* В Дамаске за сутки было зафиксировано пять нарушений перемирия *Five ceasefire violations were recorded in Damascus in 24 hours* Россия поддерживает продление режима перемирия в Сирии на 48 часов *Russia supports extending the ceasefire in Syria by 48 hours* В Сирии за сутки зафиксировано три нарушения режима прекращения огня *Three violations of the ceasefire regime recorded in Syria in 24 hours* В Сирии 207 населенных пунктов присоединились к перемирию *207 settlements in Syria joined the ceasefire*
7291694	232	Гороскоп на 22 октября для всех знаков зодиака *Horoscope for October 22 for all zodiac signs* Гороскоп на 17 июля: Рыбы, не делайте поспешных выводов *Horoscope for July 17: Pisces, do not rush to conclusions* Гороскоп на 10.05.2018 для знаков зодиака *Horoscope for May 10, 2018, for zodiac signs* Гороскоп на 1 мая для всех знаков зодиака *Horoscope for May 1 for all zodiac signs* Гороскоп на сегодня для всех знаков зодиака - Рыбы, Овен, Телец, Весы, Водолей *Horoscope for today for all zodiac signs – Pisces, Aries, Taurus, Libra, Aquarius*
1735093	172	Теплая погода сохранится в Москве и Подмосковье *Warm weather will continue in Moscow and the Moscow region* Погода на завтра: На юге Украины дожди с грозами, температура до +28 *Weather for tomorrow: In southern Ukraine, rain with thunderstorms, temperature up to +28* Главное за день: Богатырева в розыске, Путин рассказал, как помог Януковичу *Main news of the day: Bogatyreva wanted, Putin explained how he helped Yanukovych* До начала декабря в Петербурге не будет снега *No snow in St. Petersburg until the beginning of December* На предстоящей неделе в Красноярске будет дождливо и прохладно *Next week in Krasnoyarsk will be rainy and cool*
2535399	168	В выходные липчанам прогнозируют дождь и град *Rain and hail expected in Lipetsk over the weekend* Воскресенье во Владивостоке будет таким же теплым, как и суббота *Sunday in Vladivostok will be as warm as Saturday* Погода на завтра, 12 февраля *Weather for tomorrow, February 12* В течение недели в Украине дожди и заморозки будут чередоваться с потеплением *During the week in Ukraine, rain and frosts will alternate with warming* Апрельское солнце подогреет столицу *April sun will warm the capital*

Thus, implementing TDT requires selecting several hyperparameters: the vectorization method, the clustering algorithm, and the cluster radius. To find most suitable settings and verify the robustness, we test several commonly used options. For vector representation, we try TF-IDF, Word2Vec (calculating an average vector for each article) and Word2Vec weighted with TF-IDF (using a weighted average). For clustering we compare two approaches: DBSCAN and Agglomerative clustering with complete linkage. We sort through different radius values for each algorithm.

TDT is as unsupervised algorithm, so there is no single clear criterion for selection the optimal settings. However, we suggest some measures to do this choice. First, our initial dataset includes not only the 50 Russian outlets of interest but also 19 Ukrainian outlets that publish articles in

Russian. We consider the clustering to be more accurate if it is effectively separates Russian and Ukrainian news events. Second, we try to avoid excessive fragmentation in the final clusters.

To evaluate how well the clustering separates Russian and Ukrainian news, we calculate the isolation index used as a measure of segregation by Gentzkow and Shapiro (2011).

Let $c \in C$ represent a single news event, i.e., a cluster of articles, and let C indicate the full clustering – that is, the partitioning of all articles in the dataset (we consider all unclustered articles as a separate cluster labeled "unclustered"). Let ru and ukr denote the total number of articles published by Russian and Ukrainian outlets in the dataset. Define $articles = ru + ukr$ as the total number of articles. Similarly, for each cluster c, let ru_c and ukr_c indicate the number of articles by Russian and Ukrainian outlets in that cluster, so $articles_c = ru_c + ukr_c$.

Then, the measure of segregation is calculated as follows:

$$S = \frac{\sum_{c \in C} \left(\frac{ukr_c}{ukr} * \frac{ukr_c}{articles_c}\right) - \left(\frac{ukr}{articles}\right)}{1 - \left(\frac{ukr}{articles}\right)} \tag{A2}$$

Table A8 presents the results for each clustering.

We choose Agglomerative clustering with a radius of 0.6 on TF-IDF vectors as our main specification for the event detection model, because it achieves the highest segregation score.

For robustness checks, we also use two alternative specifications: Agglomerative clustering with a radius of 0.04 on Word2Vec vectors, which is the second-best option according to our metrics; and DBSCAN with a radius of 0.35 on TF-IDF, which has the best performance among the density-based clustering algorithms.

Table A8: Clustering Quality by Vectorization, Algorithm, and Radius

Vectorization	Clustering	Radius	Segregation	Share Clustered	# Clusters
TF-IDF	Agglomerative	0.300	0.341	0.406	3249416
		0.350	0.385	0.482	3539821
		0.400	0.418	0.549	3669509
		0.450	0.440	0.608	3661717
		0.500	0.455	0.659	3554150
		0.550	0.463	0.703	3380271
		0.600	0.466	0.743	3170092
		0.650	0.462	0.780	2948266
		0.700	0.452	0.817	2718771
	DBSCAN	0.150	0.169	0.185	1558906
		0.200	0.230	0.268	1933052
		0.250	0.280	0.362	2133530
		0.300	0.313	0.454	2151632
		0.350	0.329	0.536	2050910
		0.400	0.318	0.605	1904851
		0.450	0.295	0.661	1754202
		0.500	0.260	0.709	1605426
		0.600	0.169	0.786	1246448
		0.800	0.074	0.935	86537
Word2Vec	Agglomerative	0.010	0.137	0.148	1423882
		0.030	0.375	0.512	3497956
		0.040	0.412	0.643	3510413
		0.050	0.405	0.738	3182186
		0.060	0.370	0.805	2705827
		0.070	0.322	0.852	2180347
		0.100	0.196	0.920	943748
		0.150	0.119	0.948	205896
	DBSCAN	0.010	0.141	0.155	1349820
		0.020	0.237	0.372	1843707
		0.030	0.206	0.577	1494763
		0.040	0.125	0.716	966070
		0.050	0.072	0.811	526740
		0.060	0.054	0.874	248623
		0.070	0.059	0.910	108567
		0.100	0.090	0.948	10960
		0.150	0.105	0.956	1234
		0.200	0.108	0.957	241
Word2Vec + TF-IDF	Agglomerative	0.010	0.115	0.126	1223426
		0.030	0.313	0.403	3080790
		0.040	0.375	0.543	3438385
		0.050	0.392	0.652	3349252
		0.060	0.374	0.735	3029465
		0.070	0.337	0.797	2606181
		0.100	0.209	0.898	1317055
		0.150	0.118	0.942	325846
		0.200	0.106	0.952	92688
	DBSCAN	0.010	0.118	0.130	1193450
		0.020	0.210	0.268	1813370
		0.030	0.218	0.455	1725257
		0.040	0.157	0.615	1247854
		0.050	0.093	0.732	780750
		0.100	0.073	0.938	24265
		0.200	0.108	0.957	636

A3 Detecting Big and Sensitive Events

We classify news events obtained after implementing the TDT algorithm into two categories: "big" and "sensitive".

We define "big" events as clusters with articles from 10 or more different outlets. In total, 135,282 events (the top 5% by the number of outlets in the cluster) are labeled as "big" in our dataset.

Among "big" news events, we identify those that are under-reported by government-owned outlets compared to independent ones. We call such events "sensitive".

To detect such clusters, we implement an algorithm developed by Simonov and Rao (2022), but apply it at the event level rather than the word level, as in the original paper. For each event and outlet, we compute the share of articles written about this event. Then, for each event, we order the outlets in descending order by this share and compute the difference between the average rank of government-owned outlets and the average rank of independent ones. This yields a value that indicates how sensitive the event is: a higher value means that the event was more underrepresented by government-owned outlets. Table A9 reports the most "sensitive" events.

Table A9: Five news events which were overreported by the independent outlets.

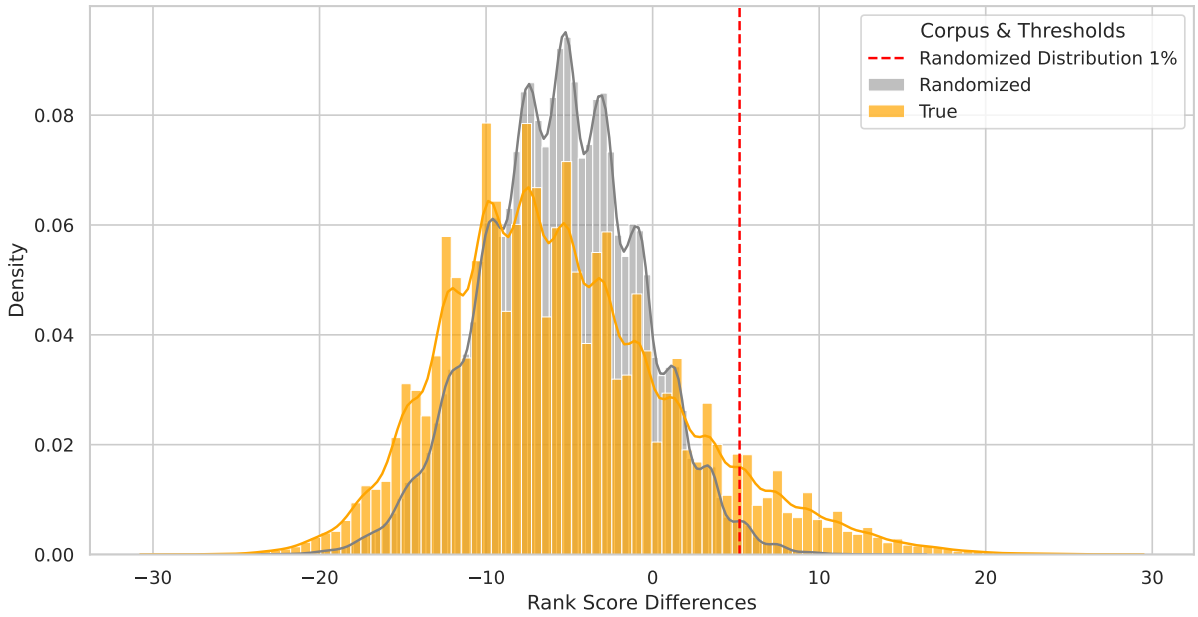
Index	Examples of Titles (Original and Translated) of News Articles in this Event
1	Навальный назвал стоимость расследования о семье генпрокурора
	Navalny revealed the cost of the investigation into the prosecutor general's family
2	Роскомнадзор стал третьей стороной в иске Дерипаски о его фотографиях в соцсетях
	Roskomnadzor joined as a third party in Deripaska's lawsuit over his photos on social media
3	СМИ: Арестованные сотрудники ФСБ сливали информацию ЦРУ
	Media: arrested FSB officers were leaking information to the CIA
4	Навальный удалил с сайта материалы про Дерипаску
	Navalny removed materials about Deripaska from his website
5	Навальный проиграл Усманову
	Navalny lost to Usmanov

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

To choose a threshold for sensitive events, we adopt the strategy described in Simonov and Rao (2022).We shuffle the shares within each outlet, recompute the rankings for each event, and thereby obtain the distribution of possible differences between the average ranks of government-owned and independent outlets.

As in Simonov and Rao (2022), the true distribution has heavier tails than the simulated distribution. We label 10,973 big events as sensitive, as their rank difference exceeds the top 1% of the randomized distribution. Figure A2 shows the true and simulated distributions of event ranks, along with the selected threshold.

Figure A2: True and Randomized Distributions of Ranks Differences for Events



A4 Additional Tables and Figures

Figure A3: Weekly Average Number of Snapshots per Day for Yandex Top-5 on Webarchive

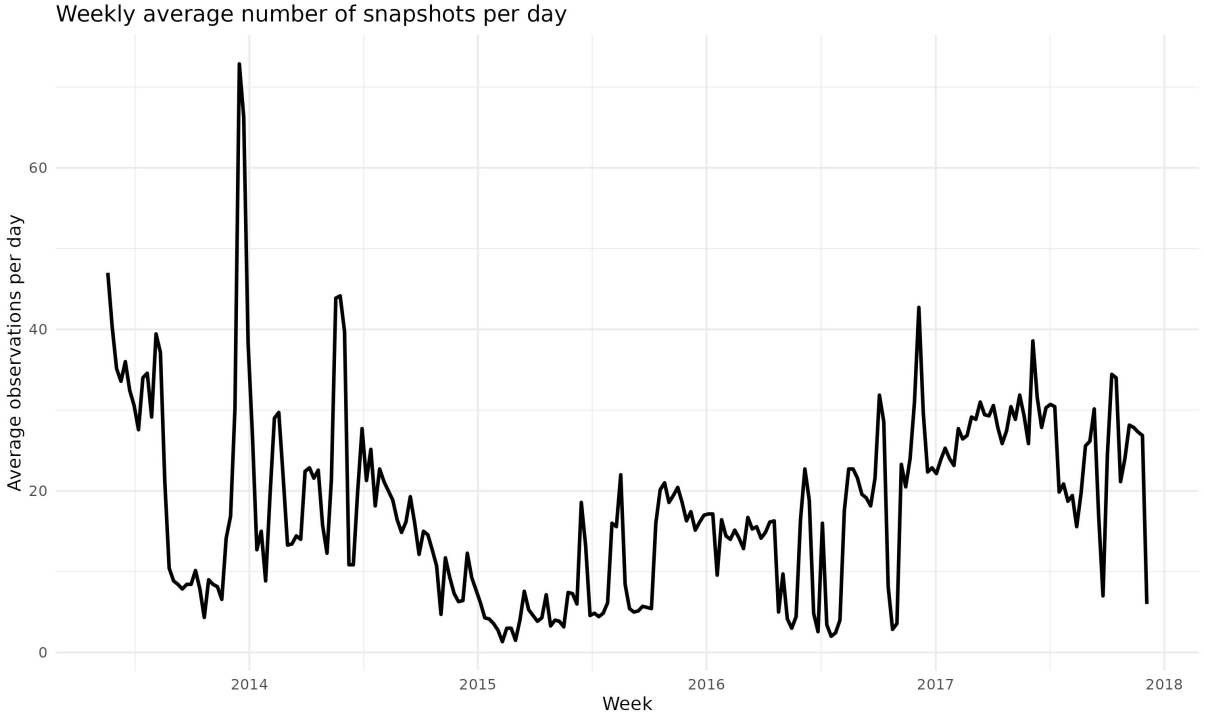


Table A10: Implied Outlet Weights Before and After the Law

		Probability of Citation (in %)				
Outlet	Type	Before the Law	After the Law	Difference	S.E. of the Difference	Tstat
Forbes.ru	ind	6.83	0.18	-6.65	0.25	-26.68
novayagazeta.ru	ind	3.98	0.01	-3.97	0.15	-26.08
vesti.ru	gov	4.34	0.64	-3.70	0.15	-24.14
tvrain.ru	ind	2.72	0.05	-2.68	0.14	-19.18
vz.ru	gov	2.57	0.05	-2.52	0.10	-24.62
Republic.ru	ind	2.27	0.11	-2.16	0.17	-13.06
echo.msk.ru	ind	2.39	0.32	-2.07	0.11	-18.60
iz.ru	inf	7.64	5.75	-1.89	0.45	-4.20
ntv.ru	gov	2.00	0.29	-1.71	0.15	-11.74
1tv.ru	gov	1.60	0.08	-1.52	0.18	-8.38
aif.ru	gov	1.39	0.05	-1.34	0.11	-12.64
rusplt.ru	inf	0.66	0.00	-0.66	0.08	-8.00
NewsRu.com	ind	0.65	0.00	-0.65	0.07	-9.65
ng.ru	inf	0.67	0.02	-0.65	0.07	-9.45
bfm.ru	inf	0.72	0.11	-0.61	0.11	-5.82
mk.ru	inf	0.15	0.00	-0.15	0.03	-5.17
Lenta.Ru	inf	7.24	7.15	-0.09	0.34	-0.27
bbc.com/russian	int	0.03	0.00	-0.03	0.02	-1.51
sobesednik.ru	inf	0.02	0.00	-0.02	0.01	-2.50
trud.ru	inf	0.02	0.00	-0.02	0.01	-1.44
rosbalt.ru	inf	0.01	0.00	-0.01	0.02	-0.52
svoboda.org	int	0.00	0.00	-0.00	0.02	-0.23
dni.ru	gov	0.00	0.00	-0.00	0.01	-0.43
Kp.ru	inf	0.01	0.01	-0.00	0.02	-0.11
russian.rt.com	gov	1.37	1.37	-0.00	0.17	-0.01
tjournal.ru	ind	0.00	0.00	-0.00	0.01	-0.10
newtimes.ru	ind	0.00	0.00	-0.00	0.08	-0.01
ridus.ru	inf	0.00	0.00	-0.00	0.01	-0.07
FB.ru	inf	0.00	0.00	-0.00	0.00	-0.03
zona.media	ind	0.00	0.00	-0.00	0.00	-0.00
golos-ameriki.ru	int	0.00	0.00	-0.00	0.02	-0.00
meduza.io	int	0.00	0.00	0.00	0.00	0.00
Znak.com	inf	0.00	0.00	0.00	0.00	0.06
polit.ru	inf	0.00	0.00	0.00	0.01	0.05
fontanka.ru	inf	0.00	0.00	0.00	0.00	0.10
utro.ru	inf	0.00	0.00	0.00	0.01	0.14
Life.ru	inf	0.00	0.00	0.00	0.01	0.16
dw.com/ru	int	0.00	0.01	0.01	0.00	1.45
svpressa.ru	inf	0.00	0.01	0.01	0.01	0.70
snob.ru	ind	0.01	0.06	0.05	0.03	1.45
the-village.ru	ind	0.00	0.06	0.06	0.04	1.58
vedomosti.ru	ind	2.92	3.28	0.36	0.23	1.53
regnum.ru	inf	2.93	4.21	1.28	0.19	6.56
kommersant.ru	inf	3.69	7.47	3.78	0.31	12.12
rbc.ru	ind	7.10	11.82	4.72	0.34	14.00
rg.ru	gov	10.69	15.42	4.73	0.31	15.40
gazeta.ru	inf	7.08	12.38	5.30	0.34	15.39
tass.ru	gov	3.91	9.36	5.45	0.35	15.73
ria.ru	gov	12.36	19.73	7.37	0.46	16.03

Implied probabilities of citations based on outlet weights α_j and α'_j from equation 4.

Table A11: Effectiveness of Outlet References in Yandex's Top-5: By Event Topic

	Dependent variable:
	full_formula
$X(j \text{ in top-5})_d$	0.036***
	(0.008)
$X(j \text{ in top-5 with topic Society})_d$	0.034*
	(0.020)
$X(j \text{ in top-5 with topic Incidents})_d$	0.002
	(0.010)
$X(j \text{ in top-5 with topic Domestic politics})_d$	0.006
	(0.011)
$X(j \text{ in top-5 with topic Economy})_d$	0.001
	(0.011)
$X(j \text{ in top-5 with topic Crime})_d$	-0.010
	(0.010)
$X(j \text{ in top-5 with topic Military})_d$	0.001
	(0.013)
$X(j \text{ in top-5 with topic Foreign politics})_d$	-0.017*
	(0.009)
$X(j \text{ in top-5 with topic Culture})_d$	0.003
	(0.013)
$X(j \text{ in top-5 with topic Sports})_d$	-0.018**
	(0.008)
$X(j \text{ in top-5 with topic Local})_d$	0.015
	(0.017)
$X(j \text{ in top-5 with topic Health})_d$	-0.011
	(0.011)
$X(j \text{ in top-5 with topic Environment})_d$	0.040
	(0.030)
$X(j \text{ in top-5 with topic Science})_d$	-0.010
	(0.012)
$X(j \text{ in top-5 with topic Technology})_d$	0.017
	(0.029)
$X(j \text{ in top-5 with topic Education})_d$	-0.022
	(0.029)
$X(j \text{ in top-5 with topic Weather})_d$	0.004
	(0.017)
$X(j \text{ in top-5 with big event})_d$	-0.007
	(0.004)
$I(j \text{ in top-6})_d$	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

Table A12: Effectiveness of Outlet References in Yandex's Top-5: By Event Type

	Dependent variable:
	$y_{jd}/y_{jd}^{\text{Yandex}}$
$X(j \text{ in top-5})_d$	0.03^{***}
	(0.004)
$X(j \text{ in top-5 with big IND event})_d$	-0.001
	(0.01)
$X(j \text{ in top-5 with big event})_d$	-0.01
	(0.004)
$I(j \text{ in top-6})_d$	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	0.96
Note:	*p<0.1; **p<0.05; ***p<0.01

Table A13: Effectiveness of Outlet References in Yandex's Top-5: By Regional News

	Dependent variable:	
	$y_{jd}/y_{jd}^{\text{Yandex}}$	
	(1)	(2)
$X(j$ in top-5$)_d$	0.027***	0.027***
	(0.003)	(0.003)
$X(j$ in top-5 with news about this region$)_d$	0.144***	0.148***
	(0.016)	(0.017)
$I(j$ wrote about this region's news$)_d$		0.001***
		(0.0001)
$I(j$ in top-6$)_d$	0.001**	0.0004
	(0.0003)	(0.0003)
Fixed Effects:		
- Week x Outlet x Region	Y	Y
- Week Day x Outlet x Region	Y	Y
- Date x Region	Y	Y
Clustered S.E.	$Weeks$	$Weeks$
Observations	597,888	492,192
Adjusted R^2	0.951	0.953
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A14: 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$)_d$	0.024***
	(0.003)
$X(j$ in top-5, bottom quartile news value$)_d$	0.004
	(0.004)
$X(j$ in top-5, top quartile news value$)_d$	0.003
	(0.005)
$I(j$ in top-6$)_d$	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

Table A15: Effectiveness of Outlet References in Yandex's Top-5: By Operation System

	Dependent variable:	
	$y_{jd}/y_{jd}^{\text{Yandex}}$	
	Detailed	Grouped
$X(j$ in top-5$)_d$ x Windows	0.029^{***}	
	(0.005)	
$X(j$ in top-5$)_d$ x Android	0.042^{***}	
	(0.009)	
$X(j$ in top-5$)_d$ x iOS	0.037^{***}	
	(0.007)	
$X(j$ in top-5$)_d$ x MacOS	0.044^{***}	
	(0.009)	
$X(j$ in top-5$)_d$ x Desktop		0.029^{***}
		(0.005)
$X(j$ in top-5$)_d$ x Mobile		0.040^{***}
		(0.008)
$X(j$ in top-5$)_d$ x Other	0.029^{***}	0.027^{***}
	(0.005)	(0.005)
Controls:		
- $I(j$ in top-6$)_d$ 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	23,643
Adjusted R^2	0.966	0.929
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 A16: Effectiveness of Outlet References in Yandex's Top-5: By Age

	Dependent variable:								
	top5_efficiency								
	(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)
$I(j \text{ in top-6})_d$	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
Adjusted R^2	0.963	0.968	0.964	0.960	0.959	0.962	0.965	0.968	0.973

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 A17: 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)
$I(j \text{ in top-6})_d$	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
Adjusted R^2	0.973	0.955	0.974

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 A18: Impact of Yandex's Algorithm Change on News Productions: Features of Articles

Dependent Variables:	prob_write_topic	av_num_words_norm	av_title_length	av_title_punctuation
Model:	(1)	(2)	(3)	(4)
Variables				
Censored Outlet × I(<year since law)	-0.0692	0.1261	3.622*	0.3370**
	(0.0492)	(0.0972)	(1.732)	(0.1346)
Censored Outlet × I(>year since law)	-0.2096*	0.4389*	8.150***	0.6950***
	(0.1153)	(0.2273)	(1.722)	(0.1862)
Fixed-effects				
Outlet × Weekday	Yes	Yes	Yes	Yes
Day	Yes	Yes	Yes	Yes
Fit statistics				
Observations	28,813	28,813	28,813	28,813
R^2	0.48153	0.30057	0.70417	0.59338
Within R^2	0.14300	0.09572	0.13419	0.18000

Clustered (outlet) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A19: Impact of Yandex's Algorithm Change on News Productions: Features of Articles (Part 1)

Dependent Variables:	prob_write_topic	prob_write_big_topic	av_num_words_norm	av_title_length	av_title_punctuation	av_title_text_dist	av_dist_to_earliest
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables							
Censored Outlet × I(<year since law)	-0.0692	-0.0596	0.1261	3.622*	0.3370**	0.0206	0.0011
	(0.0492)	(0.0415)	(0.0972)	(1.732)	(0.1346)	(0.0151)	(0.0051)
Censored Outlet × I(>year since law)	-0.2096*	-0.1819*	0.4389*	8.150***	0.6950***	0.0400	0.0193
	(0.1153)	(0.0909)	(0.2273)	(1.722)	(0.1862)	(0.0343)	(0.0170)
control_var							-0.4352***
							(0.0108)
Fixed-effects							
Outlet × Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics							
Observations	28,813	28,813	28,813	28,813	28,813	28,813	28,701
R^2	0.48153	0.57124	0.30057	0.70417	0.59338	0.50423	0.77666
Within R^2	0.14300	0.13877	0.09572	0.13419	0.18000	0.05239	0.54974

Clustered (outlet) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A20: Impact of Yandex's Algorithm Change on News Productions: Features of Articles (Part 2)

Dependent Variables:	av_sports	av_society	av_economy	av_incindents	av_education	av_science	av_culture	av_foreign_policy	av_environment
Model:	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Variables									
Censored Outlet × I(<year since law)	0.0121	0.0011	0.0120	-0.0184**	0.0014	0.0002	0.0083***	-0.0052	0.0021
	(0.0078)	(0.0017)	(0.0137)	(0.0072)	(0.0017)	(0.0020)	(0.0023)	(0.0093)	(0.0013)
Censored Outlet × I(>year since law)	0.0128	0.0053	0.0290	-0.0323**	0.0040	0.0093	0.0199**	-0.0427***	0.0065
	(0.0134)	(0.0048)	(0.0297)	(0.0120)	(0.0028)	(0.0065)	(0.0075)	(0.0084)	(0.0049)
Fixed-effects									
Outlet × Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	28,704	28,704	28,704	28,704	28,704	28,704	28,704	28,704	28,704
R^2	0.70376	0.33262	0.71576	0.70610	0.28818	0.29686	0.43878	0.60132	0.20307
Within R^2	0.00437	0.00158	0.00549	0.01902	0.00123	0.00434	0.00556	0.02186	0.00193

Clustered (outlet) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table A21: Impact of Yandex's Algorithm Change on News Productions: Features of Articles (Part 3)

Dependent Variables:	av_local	av_technology	av_military	av_crime	av_weather	av_health	av_ind_events	av_gov_events	av_big_events
Model:	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)
Variables									
Censored Outlet × I(<year since law)	0.0030	0.0008	-0.0084**	-0.0033	0.0085*	-0.0137**	0.0172**	-0.0013	-0.0335
	(0.0040)	(0.0015)	(0.0036)	(0.0078)	(0.0042)	(0.0054)	(0.0069)	(0.0011)	(0.0246)
Censored Outlet × I(>year since law)	-0.0061	-0.0006	-0.0192***	-0.0136	0.0376**	-0.0308**	-0.0002	-0.0071**	-0.1301*
	(0.0048)	(0.0015)	(0.0060)	(0.0202)	(0.0167)	(0.0123)	(0.0200)	(0.0026)	(0.0645)
Fixed-effects									
Outlet × Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics									
Observations	28,704	28,704	28,704	28,704	28,704	28,704	28,704	28,704	28,704
R^2	0.40583	0.39765	0.44888	0.61915	0.36562	0.54170	0.63387	0.57299	0.58304
Within R^2	0.00555	0.00056	0.02806	0.00416	0.03070	0.03079	0.00354	0.00595	0.06799

Clustered (outlet) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*