Measuring the diversity of news content online

Economics Discussion Paper Series Issue 15

Published 5 February 2025

Ofcom's discussion paper series

Ofcom's discussion paper series

Ofcom is committed to encouraging debate on all aspects of media and communications regulation and to creating rigorous evidence to support its decision-making. One of the ways we do this is through publishing a series of discussion papers, extending across economics and other disciplines. The research aims to make substantial contributions to our knowledge and to generate a wider debate on the themes covered.

Disclaimer

Discussion papers contribute to the work of Ofcom by providing rigorous research and encouraging debate in areas of Ofcom's remit. Discussion papers are one source that Ofcom may refer to, and use to inform its views, in discharging its statutory functions. However, they do not necessarily represent the concluded position of Ofcom on particular matters.

Contents

Section

1. Overview

1.1 This paper presents an analysis of the diversity of news content online and investigates the relationship between the way in which individuals access news online and the diversity of news content to which they are exposed. In particular, we analyse the online news diets of individuals who rely more on Online Intermediaries (OIs) for news and those who rely more on Public Service Broadcasters (PSBs). This research is an extension of work previously conducted as part of our programme of work on media plurality and online news[1] and has been carried out to support our ongoing review of Public Service Media (PSM).[2]

1.2 While television broadcast news, provided by PSBs and other licensed broadcasters, has traditionally been among the most important sources of news for UK citizens, in 2024 OIs, which comprise search, social media and news aggregators, overtook television as the most used platform for news in the UK.[3] OIs, and in particular social media platforms, have incentives to attract and retain audience attention, and have the ability to personalise the news that they show their users. These features have raised concerns that news delivered via OIs may be narrowly focused on individuals' existing views and preferences and consequently could lead to news diets that lack a diversity of viewpoints.

1.3 The empirical literature has mostly demonstrated that news consumption accessed through OIs is more diverse in the sense that it covers a larger number of news outlets. In this research, and in contrast to most of the literature, we focus on the diversity of news *topics* consumed by individuals. While we are not the first to analyse topic diversity in relation to OIs, to our knowledge there is only one other study which does this using the content of people's browsing data. Our work allows us to assess online news diets more directly than previous approaches that are based on the number or range of outlets people use. Further, in this discussion paper we expand on our previous analysis in 2024 to analyse the topic diversity of news that people access online through PSBs.

1.4 To measure topic diversity, we collected the news headlines viewed in an internet browser by a sample of approximately 8,500 internet users based in the UK over a one-month period in autumn 2021 and used natural language methods to group similar news headlines into topics. We then computed the topic diversity viewed by each person and related this measure to the share of news that they consumed through different OIs and PSBs.

What we have found – in brief

In line with the literature, we find that greater use of OIs to access news correlates with exposure to a higher number of news outlets. However, for topic diversity we find the opposite: more reliance on OIs (in particular social media and search engines) is associated with lower topic diversity. This evidence is consistent with concerns around the impact of OIs on the diversity of users' news diets.

We also find that people that get a larger proportion of their online news from a PSB have a higher diversity of topics in their news diet and that people that make little or no use of PSBs online have a lower diversity of news topics.

[1] Ofcom, 2024, Online news: research update.
[2] Review of Public Service Media (2019 – 23): Challenges and opportunities for Public Service Media.
[3] Ofcom, 2024, News consumption in the UK: 2024.

These findings come with some limitations. While our analysis is based on a sample which is representative of the UK population based on several demographic markers, the population of people willing to have their browsing and app usage tracked could be different from the general population in ways we cannot measure. Like most of the literature, we were not able to access data about users' offline news consumption. Further, we were not able to observe what news articles are presented to and viewed by users within social media platforms and news aggregators; we could only infer – with some uncertainty – whether a person arrived at a news article through an OI or by directly accessing the news provider's web site. We also stress that our results only document associations between diversity and how news is being accessed. No causal conclusions can be drawn from the data and our research design.

2. Introduction

2.1 A vibrant media landscape, with a variety of news providers across a range of platforms, helps to ensure that citizens are well-informed and able to access a wide range of viewpoints where and how they want to. This access to accurate news and to a plurality of viewpoints, including from the PSBs, is the cornerstone of a well-functioning democratic society.

2.2 The UK has a high quality and richly varied news media landscape. It is anchored by trusted news from our PSBs, news from broadcasters that is duly accurate and duly impartial, and bolstered by a strong tradition of incisive journalism and insightful commentary from news publishers across all forms of media. Enabled by new technology we have also seen new players emerge and well-respected news brands adapt and develop their business models.

2.3 OIs have come to play an important – and for certain populations a leading – role in news consumption. Ofcom reports that in 2024, 70% of UK adults used broadcast TV to consume news, 52% used social media, and 34% used newspapers.[4] Reliance on social media is higher for younger populations, and the evidence suggests that people do not change their news sources as they grow older. Therefore, the importance of OIs in news consumption is likely to increase in the future.

2.4 OIs now play a key role in many stages of the online news supply chain, including discovery, distribution, curation and monetisation. Social media providers offer their users an individualised curation of news items offered by different news producers, and they have considerable power to boost or suppress attention to a news item by prioritising it or simply not showing it in a user feed or among search results. For example, Ofcom has carried out recent research using eye-tracking technology which shows that the ranking of a news article in a social media feed strongly influences the amount of attention it receives and whether it is remembered.[5] Ulloa & Kacperski (2023) also find that ranking affects attention paid to news.[6]

2.5 There is a growing literature on the association between OIs (and social media in particular) and adverse outcomes, which we discuss in Ofcom (2022a) and Ofcom (2024). In research we carried out in 2022, we found that people who consumed news primarily through social media were less likely to correctly identify important factual information, were more polarised, and had lower trust in institutions, than those who consumed news via traditional media. In contrast, in new research which we are publishing alongside this document we find that people who consume PSB news are more likely to correctly identify important factual information, are less polarised and have higher levels of trust in institutions, than respondents who did not use PSBs for news.

2.6 In this paper, we focus on the diversity of news as one mechanism by which accessing news via OIs or from PSBs can potentially influence these societal outcomes. For example, an OI may or may not present news on a variety of topics, which may affect how well-informed a

[4] Ofcom, 2024, *News consumption in the UK*.
[5] Ofcom, 2024, Online news: research update.
[6] Ulloa & Kacperski, 2023, *Search engine effects on news consumption: Ranking and representativeness outweigh familiarity in news selection*. New Media & Society.
Ofcom, 2023c, *Media Plurality Online: Attention to News on Social Media*. Retrieved June 12, 2024, from ofcom.org.uk.

user is. Similarly, an OI may or may not present news that is balanced, which can affect polarisation and trust.

2.7 To investigate this issue, we collected the news headlines viewed by a sample of approximately 8,500 internet users over a one-month period in autumn 2021 and used natural language methods to group similar news headlines into topics. We then computed the topic diversity viewed by each person and related this measure to the share of news that they consumed through different OIs and from PSBs.

2.8 The rest of this paper is structured as follows:

- We summarise the relevant literature relating to the diversity of news consumption, and explain how our analytical approach builds on existing research;
- We describe the data sets we have used in our analysis;
- We explain the methodology we have used in our analysis, including our approach to natural language processing and topic modelling and our econometric approach to measuring diversity;
- We present out results, including additional analysis we have carried out to test the robustness of our findings; and
- We provide some conclusions and set out some potential areas for future analysis.

3. Literature

3.1 In this section we review the relevant literature on online news diversity. We first discuss previous approaches to measuring news diversity online and then consider the existing evidence on the extent to which social media platforms increase the risk of echo chambers. Finally, we set out how the present research adds to this evidence base.

Measuring news diversity

3.2 The literature on news and media diversity distinguishes between different types of diversity.[7] A traditional focus has been on diversity at the market level, e.g., the number and variety of news producing organisations, and the risk of the news media being dominated by one owner or voice. However, the digitisation of news and the increasingly important role played by OIs have shifted attention towards exposure diversity, defined as the extent to which audiences are exposed to a diverse array of news content and sources.

3.3 We focus our review on the diversity of news consumption, as it is the algorithmic personalisation of news online which presents challenges to our established understanding of media plurality. While the news media market and certain individual news outlets might be diverse in both variety and balance, the news that a user is exposed to on an OI can be narrow and skewed. Indeed, news diversity at the market level (e.g., the number of news outlets) can conceivably result in less diverse individual news consumption since the OI can draw on a larger pool of news to build a news feed specifically tailored to an individual.[8]

3.4 Most papers on exposure diversity have looked at ideological or political diversity, reflecting the American context of a two-party system (and thus a more straightforward definition of diversity and related concepts).

3.5 Flaxman, Goel & Rao (2016) analyse the browsing histories of American internet users and found that articles accessed via social media or web-search engines are associated with higher ideological segregation than those an individual reads by directly visiting news sites.[9] However, they also found, somewhat counterintuitively, that these channels are associated with greater exposure to opposing perspectives.[10] Fletcher, Kalogeropoulos, & Nielsen (2023) replicate this finding for a British panel of internet users: news accessed directly from the news outlet is more centrist, and at the same time less likely to include counter-attitudinal content.[11] The study also finds that the diversity of news outlets increases with the users' reliance on social media and search engines compared to directly accessed news.

---

[7] Voakes et al., 1996, *Diversity in the news: a conceptual and methodological framework*. Journalism & Mass Communication Quarterly; and Loecherbach et al., 2020, *The unified framework of media diversity: A systematic literature review*. Digital Journalism.
Napoli, 1999, *Deconstructing the diversity principle*. Journal of Communication.

[8] Levy, 2021, *Social media, news consumption, and polarization: evidence from a field experiment*. American Economic Review, p. 851.

[9] Flaxman, Goel & Rao, 2016, *Filter bubbles, echo chambers, and online news consumption*. Public Opinion Quarterly.

[10] Hereafter we define 'counter-attitudinal news' as news that challenges or opposes the position of a reader, and 'like-minded news' as news that conforms with the position of the reader.

[11] Fletcher, Kalogeropoulos & Nielsen, 2023, *More diverse, more politically varied: How social media, search engines, and aggregators shape news repertoires in the United Kingdom*. New Media & Society.

3.6 Cardenal et al. (2019) analyse a Spanish panel of internet users.[12] They find that news accessed directly from news outlets and news accessed through Facebook exhibit similar levels of counter-attitudinal exposure while news accessed through Google's search engine increases the probability of counter-attitudinal exposure. Similarly, Wojcieszak et al. (2022) conclude for a panel of American internet users that search engines and social media are significantly more likely to expose people to counter-attitudinal news than direct access.[13] Fletcher & Nielsen (2018) use survey data from the UK, the USA, Spain and Germany to demonstrate that people who use search engines for news discovery use more news sources and are more likely to use news sources from both ends of the political spectrum.[14]

3.7 These and other studies consider the number of distinct news outlets to which individuals are exposed as the outcome of interest.[15] The emerging consensus among these articles is that outlet diversity for news accessed through OIs is at least as high as news accessed directly.[16]

3.8 Very few papers have analysed topic diversity – the type of diversity that is the focus of this paper. Haim, Graefe & Brosius (2018) create artificial Google accounts with different preferences and browsing histories to compare the topic distribution on Google News across these accounts.[17] They find that these artificial accounts were presented with news articles aligned with their preferences in their news feed, but that a news search containing the same search words produced a nearly identical selection and ranking of news articles across the different accounts. Möller et al. (2018) compare different recommender systems applied to news articles from a Dutch broadsheet newspaper and conclude that recommendation algorithms present a more diverse range of topics than human editors.[18] Both these studies consider topic diversity in a stylised setting (e.g., using artificial Google accounts, and simulating article recommendations) rather than in the context of actual news consumption, leaving open the question of how diversity in consumed news differs across different access and discovery modes.

3.9 Closer to our own research, Jürgens & Stark (2022) use content analysis to classify news articles into topics and analyse how OI use relates to news topic diversity.[19] They look at a panel of German news consumers and find mixed results. On the one hand, if an individual increases their use of OIs (i.e. a comparison over time), their news diet becomes more

---

[12] Cardenal et al., 2019, Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure. The International Journal of Press/Politics

[13] Wojcieszak et al., 2022, *Avenues to news and diverse news exposure online: comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks*. The International Journal of Press/Politics.

[14] Fletcher, R., & Nielsen, R. K. (2018). Are people incidentally exposed to news on social media? A comparative analysis. New Media & Society.

[15] See for instance: Fletcher, Kalogeropoulos & Nielsen (2023); Scharkow et al, 2020, *How social network sites and other online intermediaries increase exposure to news*. PNAS; Stier et al, 2022, *Post post-broadcast democracy? News exposure in the age of online intermediaries*. American Political Science Review; and Ulloa & Kacperski, 2023.

[16] Ross Arguedas et al., 2022, *Echo chambers, filter bubbles, and polarisation: a literature review*. Oxford: Reuters Institute for the Study of Journalism, p. 17.

[17] Haim, Graefe & Brosius, 2018, *Burst of the filter bubble? Effects of personalization on the diversity of Google News*. Digital Journalism.

[18] Möller et al., 2018, *Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity*. Information, Communication, and Society.

[19] Jürgens & Stark, 2022, *Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption*. Journal of Communication.

diverse. On the other hand, people who use search engines and certain social media more have less diverse news consumption when compared to people who use them less (i.e. a comparison between individuals at a point in time).

Social media and echo chambers

3.10 A range of research in recent years has investigated concerns that social media platforms have contributed to harmful societal outcomes through the ways in which news is presented to users on social media news feeds, and the ways in which users are encouraged to interact with and share news on these platforms. For example, some research suggests that user attention on social media is often drawn to news content that is like-minded[20], emotionally charged[21] or false.[22] As a result, an algorithm designed to maximise user engagement may end up promoting such forms of content.

3.11 Of particular interest to the present research are the concerns around echo chambers on social media platforms. An echo chamber, broadly defined, is a place where the ideas seen reflect the ideas that a user already holds. There is a risk that if people see a disproportionate amount of information which reflects opinions they already hold, it can give them a one-sided view of events. There is also the potential for those in echo chambers to become more polarised as a result, and potentially more likely to believe and circulate misinformation.[23]

3.12 Echo chambers could arise for a number of different reasons including:

- Filter bubbles: algorithms may filter the news users receive based on their previous online behaviour or the behaviour of people like them, in order to drive engagement.
- High segregation: in a highly segregated news environment people with different viewpoints are unlikely to read the same news articles.

3.13 In relation to filter bubbles, a news feed algorithm might predict that a user is more likely to engage with news consumed by a user's network, and it might predict high engagement with a news item if engagement was high with similar news items in the past. If a user has read news on a particular topic or presenting a particular viewpoint, and if the user is connected to people with similar interests and opinions, then conceivably this user will find themselves in a feedback loop: their news feed will feature news on a certain topic, resulting in engagement with this news, resulting in more news on the topic in the news feed, and so forth. The literature has referred to this phenomenon as 'filter bubbles' as the algorithm and the user mutually reinforce the filtering out of non-engaging news. There is some evidence in the literature suggesting that filter bubbles could occur on social media. One study found that Facebook's algorithm is more likely to show users content from news outlets that shared their political views, compared to news outlets of a different political slant.[24]

---

[20] Bryanov et al., 2020, *Effect of partisan personalization in a news portal experiment*. Public Opinion Quarterly.
[21] Rathje, Van Bavel & Van Der Linden, 2021, *Out-group animosity drives engagement on social media*. Proceedings of the National Academy of Sciences; and
Robertson et al., 2023, *Negativity drives online news consumption*. Nature human behaviour.
[22] Vosoughi, Roy & Aral, 2018, *The spread of true and false news online*. Science.
[23] Acemoglu et al. (2022) use a theoretical model to show that users are more likely to share misinformation within their social network if this is made up of those who have similar views to themselves, because users expect positive feedback from sharing articles that align with their network's views.
[24] Levy (2021), Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.

3.14 On the other hand, others have argued that social media could in principle facilitate the discovery of news which the news consumer would otherwise not view. For example, social media sometimes features a degree of 'automated serendipity' of news articles to prevent monotony, boredom, and eventually a loss of interest in a platform.[25] Social media users can also 'stumble' upon news browsing through their social media feed without having intended to look for news, for example through news items recommended by weak ties.[26] This mechanism is sometimes referred to as 'incidental exposure'.[27] These features of OIs can potentially increase the diversity of a person's news consumption.

3.15 In relation to segregation, González-Bailón et al. (2023) and Levy (2021) both report that news articles visited on Facebook are more segregated than news sites accessed directly.[28] We note that their segregation measures capture the audience diversity of a news article, while our research focuses on the diversity of news articles viewed by individuals. Levy also finds that Facebook seems to promote more articles from like-minded than counter-attitudinal sources, even if the user follows both.

3.16 Bakshy, Messing, & Adamic (2015) and González-Bailón et al. (2023) both find that counter-attitudinal news content on Facebook goes through a 'funnel': a randomly picked news article has a good chance of being counter-attitudinal for a user, but a news article shared by the user's connection is less likely to be counter-attitudinal.[29] The likelihood of being exposed to a counter-attitudinal news article in Facebook's news feed and engaging with such a news article is lower still. Nyhan (2023) also finds considerable exposure of Facebook users to like-minded sources: 50.4% of a user's Facebook content comes from like-minded sources as opposed to 14.7% from counter-attitudinal sources.[30]

Contribution of this research to the evidence base

3.17 As discussed above, the empirical literature has mostly considered diversity in terms of the number of news sources, the ideological range of news sources, or both. These studies tend to find that, in comparison to news accessed directly through the homepages of news outlets, news viewed through OIs tend to come from a larger number of outlets and are more balanced across the left-right spectrum. However, a diversity of news outlets does not necessarily mean that users are getting a diversity of viewpoints. It is possible, for example, that users are exposed to the same, or a limited range of viewpoints from many different outlets.

3.18 In this research, and in contrast to most of the literature, we focus on the diversity of news *topics* consumed by individuals. We use Shannon entropy as a measure of diversity[31] measure: it encompasses both variety (number of topics) and balance (the dominance of some topics) of news consumption. A person's news consumption is thus more diverse if it

[25] Möller et al., 2020, *Explaining online news engagement based on browsing behaviour: Creatures of Habit?* Social Science Computer Review.
[26] Barberá, 2014, *How social media reduces mass political polarization. Evidence from Germany, Spain, and the US*. Unpublished manuscript.
[27] Cardenal et al., 2019, and Fletcher & Nielsen, 2018, *Automated serendipity: The effect of using search engines on news repertoire balance and diversity*. Digital Journalism.
[28] González-Bailón et al., 2023, *Asymmetric ideological segregation in exposure to political news on Facebook*. Science.
[29] Bakshy, Messing & Adamic, 2015, *Exposure to ideologically diverse news and opinion on Facebook*. Science.
[30] Nyhan et al., 2023, *Like-minded sources on Facebook are prevalent but not polarizing*. Nature.
[31] McDonald & Dimmick, 2003, *The conceptualization and measurement of diversity*. Communication Research.

covers a wider range of topics, and if one or few topics do not dominate the total news consumption.

3.19 While this is not the first study to analyse topic diversity in relation to OIs, it is to our knowledge only the second one to do so using people's actual browsing data.[32] This allows us to assess news diets more directly than previous approaches based on outlets. For example, a person might read news from different outlets but with very similar content, in which case their news diet might be considered diverse if measured in terms of numbers of outlets, but it would be narrow in terms of news topics.

3.20 The present research is also novel in investigating how news consumption on PSBs relates to news diversity. PSBs have a statutory requirement to include news programming of high quality and covering national and international matters and might therefore expose their audiences to a wide range of news topics.[33] We therefore also analyse people's topic diversity in relation to how much of their news consumption comes from the BBC and other PSBs.

[32] Other studies which have looked at topic diversity are:
Haim, Graefe & Brosius, 2018, and Möller et al., 2018. These studies use simulations instead of real browsing data. Jürgens & Stark, 2022, also look at topic diversity and use browsing data from a German panel.
[33] Communications Act 2003, section 279. Note that, unlike the BBC, the commercial PSBs do not have statutory requirements relating to the provision of news online. As discussed in the results section, the BBC is by far the largest provider of online news among the PSBs in our sample.

4. Data

4.1 Our main source of data is the Ipsos Iris online audience measurement panel, which tracks the web and app activity of a representative sample of UK adults (15+) over time.[34] Ofcom purchased one month of web tracking data covering the period between 15 September and 15 October 2021. The dataset comes as a table with one row for each visit to a website by an individual on desktop or a mobile device. The dataset does not record any content viewed on a social media feed or on an app. Thus, we do not observe news consumed directly on social media or on any app.

4.2 We filtered the dataset to only include visits to a pre-defined list of web domains that correspond to 23 news outlets in the UK. These are the same outlets as those included in Fletcher, Kalogeropoulos & Nielsen (2023), with the addition of iNews and CNN. The BBC, Channel 4, and ITV are PSBs, and all other outlets are non-PSBs. The final sample contains close to 58,000 article headlines, and close to 230,000 article views (as some articles are read by several people).

4.3 Each visit to a news article on a provider website is categorised according to the route an individual took to get to that article (access mode). We distinguish between the following access modes for a news article: direct; social media; search engine; news aggregator; and other. We infer the access mode for an article from the user's browsing and app usage history using the following algorithm:

- If a user accesses a homepage of a news outlet and afterwards opens a news article on that outlet's website, then we consider the access mode for this article to be direct.
- If the access mode cannot be classified as direct using the above approach, we proceed to assess whether it can be classified as an OI. If a user visits an OI website or uses an OI app and subsequently opens a news article on their browser (e.g., through clicking a hyperlink), then the access mode for this news article is OI (which we categorise as either social media, search engine or news aggregator).[35]
- To allow for the possibility that the user does not access the news article immediately after accessing the news outlet homepage or an OI – for example by opening a new tab on their browser before opening the news article – we also use the relevant classification if the news article access is at most five steps after the visit to the news outlet homepage or OI.[36] If more than one access mode is detected within these five steps, then we use the most recent (least distant in terms of steps) access. For

[34] Ofcom, 2022, Media Plurality and Online News Annex 5: Ipsos Iris passive monitoring data analysis.

[35] To classify the access mode as social media after using a social media app (rather than visiting a social media website) we also require the news article visit to be within five minutes of using this social media app. This is because accessing social media via using an app – unlike using a web browser to visit a website – does not allow for the possibility of leaving a tab open and coming back to it at a later stage to continue browsing; therefore, the delay reduces our confidence that the news visit originates from the social media app.

[36] Example: A person opens the website of news outlet on their browser tab X. They then open a new browser tab Y to look for holiday destinations. Then they go back to tab X and click on a link to a news article. If the person has spent up to five steps (websites) on browser Y, then this article will be classified as 'direct access'. Otherwise it will be 'not attributed' (see below).

example, if a user opens Google's search engine, then the BBC homepage, and two steps later an article on the BBC, then the access to this article is classified as direct.

4.4 If a first visit to a news article on an outlet's website is followed by a chain of other news article visits within the same outlet's website, then the access mode for the subsequent visits can be indeterminate. For example, if a user exhibits a browsing history of (social media -> news 1 on outlet A -> news 2 on outlet A) then news 1 has social media as access mode, but it is unclear whether news 2 should be classified as direct access, or as social media access. We thus follow Fletcher, Kalogeropoulos & Nielsen (2023) in that we only classify the access mode for the first news visit according to the rules above, but not the following news visits in a chain of news visits within the same outlet and within one hour (the access mode for these news visits is recorded as 'indeterminate'). We refer to such a chain within an outlet as a *news session*.

4.5 We also classify the access mode as 'indeterminate' if an article cannot be attributed through the steps outlined in 4.3, and if the user accessed another article from the same outlet within the past 24 hours. This is because we cannot confidently interpret the news visit as a continuation of the past news session or as a new news session. Finally, we also do not classify a news visit's access mode if the same user has visited the same article in the past. All remaining article views are classified as 'other': These are news sessions which started without the user accessing a news outlet or OI intermediary recently (e.g., through links from other websites or emails).

4.6 Importantly, even for news visits for which we cannot determine the access mode, we still classify the topic and the outlet of the news article. This information enters our computations for the news diversity measures for users. The distribution of access modes for the articles that can be attributed to an access mode is shown in Table 1. In addition, on average 30% of news articles in people's news diets are composed of PSB articles (though this masks a largely bimodal distribution – see results section below).

Table 1: Distribution of news sessions across access modes

Access mode	Share
Direct	43.3%
Social	5.5%
Search	6.7%
Aggregator	0.5%
Other	44.0%

Source: Ofcom analysis of Ipsos Iris panel-only data, 15 September – 15 October 2021.

4.7 The final dataset used for this report therefore consists of all unique visits to the domains of major news outlets by members of the Ipsos Iris panel, tagged with the most likely mode of access. In total we identified 57,648 unique news articles, and 322,660 visits to news outlet domains. Of the 322,660 total visits, we were able to determine the access mode for 230,000.

4.8 Within this context, the analysis is focused on article headlines. This choice was made for both methodological and conceptual reasons. Firstly, we could more easily collect the article

headlines than the article bodies due to access restrictions. Furthermore, article lengths vary considerably across and within news outlets and we are concerned that topic classifications might vary systematically by article length – especially if the article length is a feature of data truncation (e.g., where a paywall restricts access to some of the text of an article).

4.9 Secondly, article headlines are more likely to contain words and expressions which capture the gist of the news content since they are selected by publishers to do so, and less likely to contain expressions which might make it more difficult for a topic model to determine clusters of similar articles such as filler words. We therefore think that for our topic analysis, headlines are better suited than full article texts.[37]

[37] Headlines are of course not fully immune against a confounding of topics. In an exploratory stage of this research, we observed that articles referring to a boxing *fight* match and articles referring to a certain court *fight* were often categorised under the same topic.

5. Methodology

5.1 The methodology for this analysis can be split into three components: natural language processing, topic modelling, and analysis of diversity. We first need to turn the news article headlines into a useable format for quantitative analysis. For this we use state of the art tools for extracting numeric features from text language models. Once we have numeric representations of each of the headlines in the dataset, we then proceed to using statistical techniques to identify clusters of similar headlines which we use as topics. We then identify the distribution of topics for each individual and construct a diversity measure described further below. This makes it possible to relate the diversity of topics in everyone's news diet to the share of OIs and PSBs in their news browsing sessions.

Natural Language Processing

5.2 The first point of departure between this work and other work in the area is the use of natural language processing (NLP) to understand differences in the content of news. Information about the outlet that produced a news article only provides limited information about the diversity of views to which users have access.

5.3 Traditionally, NLP has tended to involve analysing raw word counts or word counts weighted by their frequency in each document relative to the whole corpus of documents (we refer to this technique as 'tf-idf'). These methods can be successful for simple tasks, but they do not consider word order or context. Sentences with similar meanings that share few words in common will have very different representations and vice versa. Additionally, if the number of words in the corpus is very large then the word counts for individual sentences with 10 or so words may have many zeros. When we compare the similarity of sentences later, this can introduce measurement error by distorting the measured distance between sentences.[38]

5.4 Consider the following two sentences:

1. She likes biscuits.
2. He enjoys cookies.

5.5 NLP methods using word counts alone will fail to capture the similarity between these two sentences, because they do not share any words in common. On the opposite extreme, sentences that contain the same words but have different meanings will mistakenly be seen as similar:

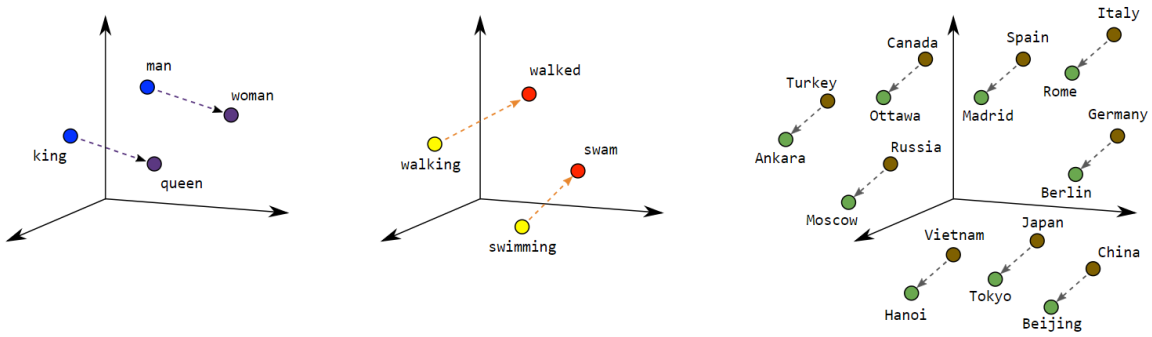
1. I sat on the sand by the bank.
2. I sat in the waiting room at the bank.

5.6 These issues can be partly addressed by making use of word embeddings. Word embedding models represent every word as a vector of numbers. The vectors are learned by deleting words from a sentence and then training a neural network to predict the missing word using the surrounding words. The model will learn that words that often appear together should be represented by vectors that are close together. This makes it possible to easily identify

[38] For example, if (dis)similarity is measured using the Euclidean distance between the tf-idf vectors, then having many entries equal to 0 will make a pair of sentences appear to be more similar than they actually are.

analogies like king is to queen as man is to woman or Paris is to France as London is to the UK. Additionally, words that have similar meanings are likely to be used in similar contexts and will therefore also be close together. Figure 1 below illustrates how this looks if words are represented by a 3-dimensional vector.

Figure 1: Illustration of word embeddings



Male-Female Verb Tense Country-Capital

Source: Google.[39]

5.7 Returning to the pairs of sentences above, word embeddings solve the problem with the first pair, but not the second. This is because the word embeddings for 'he' and 'she' are likely to be close to each other, as will the embeddings for 'likes' and 'enjoys' and 'cookies' and 'biscuits'. On the other hand, the word embedding for 'bank' is always the same regardless of whether it has a different meaning in context. This is where NLP models that make use of the transformer architecture come in.[40] They combine word embeddings within each sentence together by computing a weighted average that takes the meanings of the other words in the sentence into account. Transformers can pick up, for example, that the use of the word 'sand' in the same sentence as the word 'bank' suggests we are talking about a riverbank rather than a financial institution. This distinguishes transformer models from more commonly used language models such as LDA.

5.8 Since they address our theoretical concerns about using count-based models and have shown state of the art performance in identifying similar sentences, we have chosen to use sentence transformers for our analysis.[41] We use the SentenceTransformer package developed for Python to encode news headlines. In particular, we chose the pre-trained all-MiniLM-L12-v2 model given its strong performance and small size. This model was trained on a large and diverse corpus of online text. For example, the training data includes academic paper titles and abstracts as well as comments from social media platforms such as Reddit.[42] The input to the transformer is a news headline. The output is a 384-dimensional numerical representation (a sentence embedding) of the input headline. We do this for our entire set of headlines.

---

[39] Google Developers, *Embeddings: Translating to a lower-dimensional space*.
[40] Devlin et al., 2018, *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805.
[41] Reimers & Gurevych, 2019, *Sentence-bert: Sentence embeddings using siamese bert-networks*. arXiv:1908.10084.
[42] See the online documentation.

5.9 We illustrate the usefulness of sentence embeddings for content analysis of news in Table 2. It shows the similarity in headlines based on the Euclidean distance between each pair of sentence embeddings for 5 fictitious headlines. Scores are scaled to range from 0 to 1. A value of 1 indicates that two sentences are exactly the same and a value of 0 indicates that they are completely unrelated.

Table 2: Example of distances between news headlines

	Rising fuel prices are causing households hardship	Anger at expansion of low-traffic neighbourhoods	Russian army advancing on Kharkiv	Two soldiers killed in explosion in Kabul
Rising fuel prices are causing households hardship	1.00	0.21	0.02	0.02
Anger at expansion of low-traffic neighbourhoods	0.21	1.00	0.03	0.00
Russian army advancing on Kharkiv	0.02	0.03	1.00	0.14
Two soldiers killed in explosion in Kabul	0.02	0.00	0.14	1.00

5.10 Naturally, every sentence achieves a perfect similarity score with itself, but there are a couple of other patterns that emerge. The two headlines 'Rising fuel prices are causing households hardship' and 'Anger at expansion of low-traffic neighbourhoods' have the highest similarity score (0.21), presumably because they are both related to traffic and transport. Similarly, the two headlines 'Russian army advancing on Kharkiv' and 'Two soldiers killed in explosion in Kabul' have a relatively high similarity scores of 0.14 because they both relate to events around armed conflict. However, the headlines relating to traffic show little similarity to the headlines relating to armed conflict. If we were to crudely partition this set of five headlines into topics using their similarity scores, we would therefore end up with two topics, 'traffic' and 'armed conflict'. There are however much more sophisticated ways of doing this called topic modelling.

Topic modelling

5.11 Topic modelling broadly consists of three steps. First, raw text must be turned into a vector representation. We do this by using sentence transformers as described above. The next step is to reduce the dimensionality of the resulting embeddings. This is not strictly necessarily, but creating clusters for 58,000 headlines and 384 features (the dimensionality of the sentence embedding) is computationally very intensive. Reducing the dimensionality can

speed up the task while retaining most of the information captured in the sentence embedding as some of the embedding elements will be important for distinguishing individual headlines from each other, but others will not.

5.12 Dimensionality reduction algorithms collapse as much of the important sources of variation between sentence embeddings as possible onto a small number of dimensions (in our research we chose five dimensions). Along with most current work on topic modelling, we use a non-linear method called UMAP that aims to preserve the distances between individual sentences and has achieved state of the art results in identifying clusters in high-dimensional data.[43] This is standard practice among embedding-based topic modelling methods.[44]

5.13 The value of dimensionality reduction for topic modelling is that it substantially reduces the computational burden to identify clusters. Instead of computing the similarity between two 384-dimensional vectors, we simply do it with 5-dimensional ones without losing much of the original information.

5.14 Once we have applied dimensionality reduction to the sentence embeddings, we then use cluster analysis to identify groups of headlines that are most like each other. We used a hierarchical clustering algorithm called *hdbscan* for this.[45] It has two advantages over alternatives for our purposes: it classifies headlines that do not clearly belong in any of the topics as outliers and it does not try to ensure that the clusters it identifies are all the same size. This means that some clusters of similar headlines that cover more popular news stories (such as the coronavirus epidemic) can be larger than ones that cover less popular stories or stories which attract less coverage (such as the volcanic eruption in the Canary Islands).

5.15 The number of clusters and the number of elements in each cluster are determined by the clustering algorithm. The user can specify parameters which govern how strict the algorithm is in considering a group of points to be a cluster, such as the minimum number of points in a cluster, and how close two points would need to be to be considered part of the same cluster. The user can also specify the exact number of clusters. If this number is smaller than the number of initial clusters, then the algorithm starts merging the most similar clusters until the desired number of clusters is achieved.

5.16 In this instance our baseline model requires a cluster to have at least 50 points and uses the default settings for the remaining parameters. We then inspect the resultant topics visually and we examine several different restrictions on the number of topics as part of our robustness checks.

Econometric analysis of diversity

5.17 Once all articles have been tagged with a topic (or as an outlier), we measure the diversity of the news diets of all individuals in the dataset. We have chosen individuals as the unit of analysis instead of all news visits for each mode of access (i.e., direct, social, etc.) for several reasons. Most importantly, OIs might show users a wide range of viewpoints collectively, but this might still result in low diversity at the individual level. Consider a news aggregator with

[43] McInnes, Healy & Melville, 2018, *Umap: uniform manifold approximation and projection for dimension reduction*. arXiv:1802.03426.
[44] Grootendorst, 2022, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794; Angelov. *Top2Vec: Distributed representations of topics*. arXiv:2008.09740, 2020.
[45] Campello, Moulavi & Sander, 2013, *Density-based clustering based on hierarchical density estimates*. In Pei et al., Advances in Knowledge Discovery and Data Mining, pp. 160-172, Springer.

two users, one who likes reading about politics and another who likes reading about sports. Even if the news aggregator only shows each person articles that they are interested in, the overall set of articles it recommends will cover a diverse set of topics. Consequently, we choose to measure the diversity of topics that individuals are exposed to – regardless of access mode – and relate this to the proportion of their news diet that comes from each access mode and to the proportion of PSB articles in their news diet.

5.18 Following Fletcher, Kalogeropoulos & Nielsen (2023) we measure diversity using entropy, specifically Shannon's H. Entropy can be regarded as a measure of the unpredictability of the topic of a randomly picked news article. Consider an individual who reads 10 articles. If all the articles they read come from the same topic, then Shannon's H is 0 (we can predict the topic with 100% certainty). If all articles are about different topics, then Shannon's H will be higher to reflect the greater unpredictability.[46] Formally,

$$Entropy = -\sum_{t} p_t \log_2 p_t$$

5.19 where p_t is the proportion of an individual's news diet that comes from topic t. We have also included other measures of diversity, including Simpson's D which is equivalent to the Herfindahl-Hirschman-Index, as part of the robustness checks.

5.20 The entropy calculated over a sample containing few articles can exhibit severe bias. In the extreme, the entropy computed for any person who reads only one article is always 0. We therefore limit our sample to individuals who have accessed at least 10 news articles over the observation period, where we count any article from any of the 23 news outlets listed above as a news article. Out of a total of 8,592 individuals in the original dataset, we calculate entropy for 3,807 of them (the reason for this large drop is that a very large subset of the individuals only read a small number of articles).

5.21 For each of these individuals, we then calculate the share of their news sessions that come from each mode of access as defined in the Data section. We can then use this to relate the topic diversity of each user i's online news diet to the share of their news from each access mode:

(1) $$Entropy_i = \beta_0 + \beta_{So} Social_i + \beta_{Se} Search_i + \beta_{Ag} Aggregator_i + \beta_{Ot} Other_i + \varepsilon_i$$

where $Social_i$ is person i's share of news sessions from social media, and the remaining variables are defined analogously. The share of direct news sessions is the reference category. The estimated coefficients β therefore tell us how much a 1 percentage point increase in the share of an individual's online news sessions coming from each online intermediary at the expense of direct access is associated with a change in the diversity of their overall news consumption in terms of topics.

5.22 Finally, we also inspect the relationship between entropy and the prevalence of PSBs in a person's news diet. This simply replaces the independent variables in equation (1) with the share of PSBs in one's news diet, irrespective of how the PSB was accessed:

(2) $$Entropy_i = \beta_0 + \beta_{PSB} PSB_i + \epsilon_i$$

[46] In this example the index t runs from 1 to 10 (topics are numbered 1 to 10). Each of the ten topics will have $p_t = 0.1$ (10% of the articles are about any particular topic). Applying these numbers to the entropy formula, the resulting entropy is $-10 \times (0.1 \times \log_2 0.1) = 3.3$.

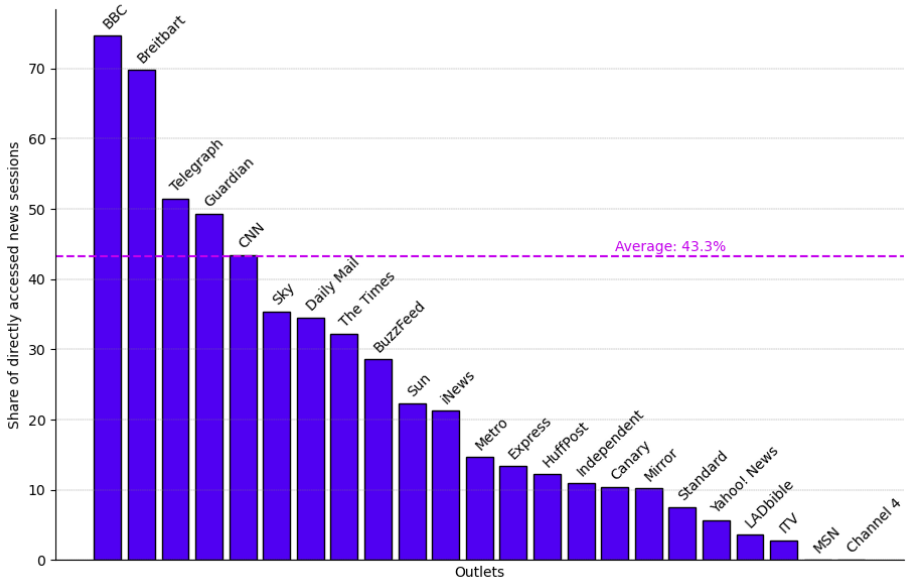
6. Results

6.1 Overall, we find that greater use of OIs to access news correlates with higher outlet diversity. However, for topic diversity we find the opposite; more reliance on OIs (in particular social media and search engines) is associated with lower topic diversity. A larger share of PSBs in the user's news diet correlates positively with topic diversity. The rest of this section sets out the details of our findings.

Mode of access

6.2 Our first observation relates to the reliance of news outlets on OIs. Figure 2 shows that while articles posted on several large media outlets are often accessed directly, there is also a long tail of news outlets who rely on OIs or other access modes for more than half of their article accesses. This aligns with our finding in Ofcom (2022).[47] Importantly, the BBC is least reliant on OIs and other access modes. More than 70% of news sessions[48] on the BBC were accessed through the BBC homepage. This is true only for 10% of news sessions on, for example, the Huffington Post, the Independent, the Mirror, etc. For our analysis, there is therefore high intersection between articles which are accessed directly, and articles published by the BBC which is by far the biggest PSB for online news.

Figure 2: Share of directly accessed news sessions by outlet



Source: Ofcom analysis of Ipsos Iris panel-only data, 15 September – 15 October 2021.

Note: News sessions for an outlet are the sum of all instances that a news session on the outlet has been started by any person in the Ipsos panel.

[47] Ofcom, 2022, Media plurality and online news, Annex 5.
[48] A news session refers to an article or a chain of subsequent articles viewed on the same outlet. See Data section for more details.

Modelled Topics

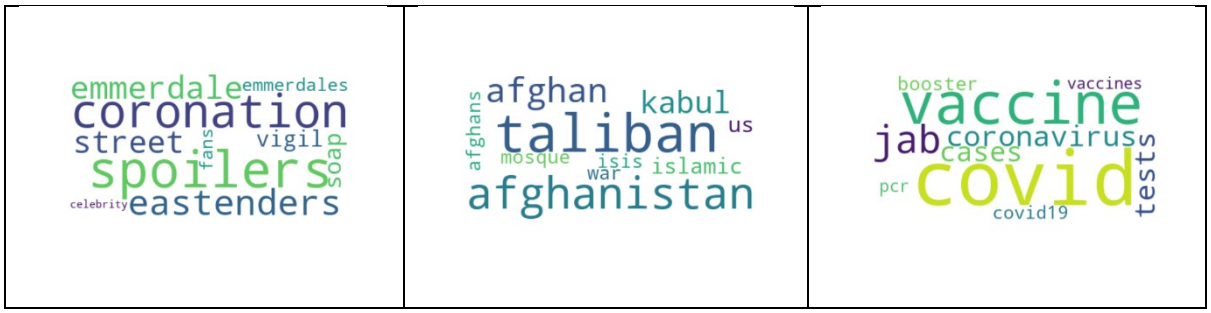
6.3 We next outline some of the major descriptive features of the topic modelling. In total, the baseline version of our model identified 106 topics, the largest of which contained articles about the reality TV show Married at First Sight with 2,606 articles and the smallest of which contained articles about broadband with 50 articles. Table 3 shows a snapshot of the five largest topics, the number of articles belonging to each, representative words identified by the model, and an example headline from the topic. The most popular topics for our sample of news consumers in Autumn 2021 were Married at First Sight, the petrol crisis, the Sarah Everard murder case, Westminster politics/Brexit, and assorted book/TV/film reviews.

Table 3: Top 5 topics and their representations

Topic number/name	Number of articles	Representative words	Example headline
1: Married at First Sight	2,606	Katie, married, she, her, sight, Kardashian, first, Stacey, price, at	Married At First Sight UK: Morag is asked why the 'old Luke' wasn't good enough for her
2: Energy crisis/driver shortage	1,874	Energy, petrol, fuel, crisis, gas, climate, drivers, shortage, bills, driver	Energy crisis UK: Which energy suppliers have gone bust and why?
3: Sarah Everard murder	1,441	Sarah, Couzens, Everard, Wayne, murder, police, jailed, Everard's, killer, man	Police officer Wayne Couzens charged with murder of Sarah Everard appears in court
4: Politics/Brexit	1,365	Brexit, Starmer, Keir, Boris, EU, Labour, Johnson, conference, Ireland, Johnson's	Labour conference 2021: Sir Keir Starmer takes fight to Boris Johnson in deeply personal speech
5: Book/Film/TV reviews	1,355	Review, the, of, books, Netflix, comedy, music, and, best	The week in theatre: A Number; The Visit; Alone in Berlin review

6.4 To give a clearer picture of the performance of the topic modelling, we have also generated word clouds showing the relative importance of the key words for each topic. Figure 3 below shows word clouds for three exemplary topics, one about soap operas, one about the war in Afghanistan, and one about the covid vaccine. The word clouds demonstrate that the characteristic words identified for a topic align with our intuition. For example, the word cloud about the war in Afghanistan groups together the words 'Afghanistan', 'Taliban', 'Kabul', and 'war':

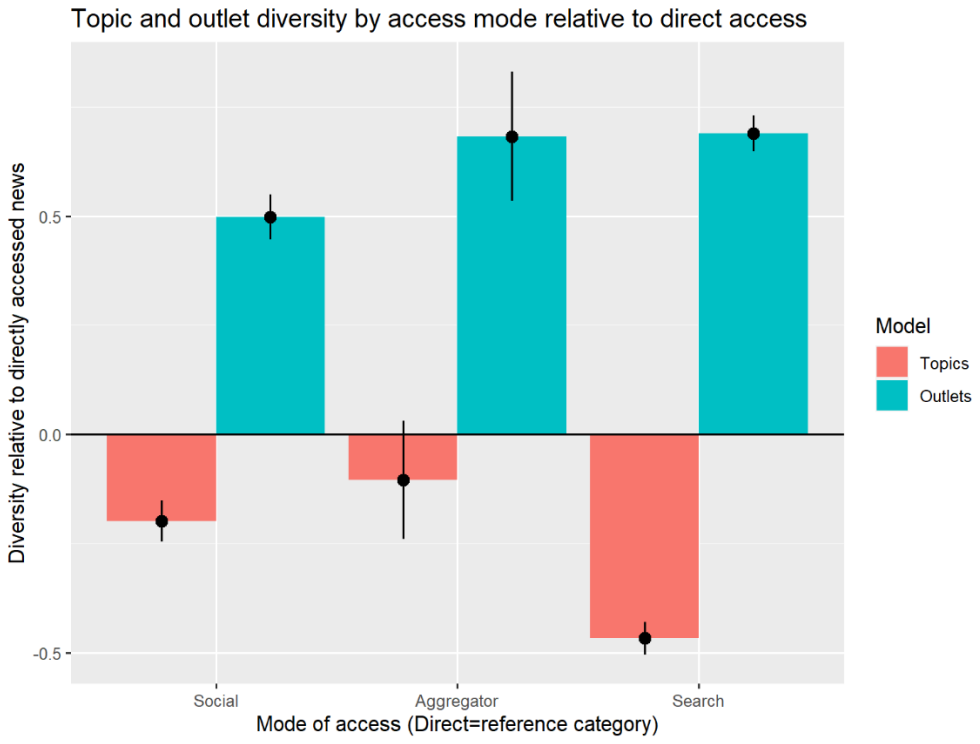
Figure 3: Examples of word clouds from topic models



Online intermediaries and diversity

6.5 Once we have tagged every article with a topic (or as an outlier), we then proceed to calculate the diversity of each individual user's news diet as described above. We do this for both the topics of news articles that they accessed and for the outlets that published those articles. This allows us to compare against the baseline of other research that has focused on the diversity of outlets. Figure 4 and Table 4 report the headline regression results. The y-axis in figure 4 represents the expected value of diversity for someone who gets all their news from that source, compared to someone who gets all their news from direct access. For ease of interpretation, we have rescaled the entropy values to range between 0 (for the lowest entropy in the sample) and 1 (for the highest entropy in the sample). Following previous findings, more intermediated news sessions are associated with greater diversity of outlets (green columns). But we see the opposite finding when we focus on the diversity of topics (red columns).

Figure 4: Topic and outlet diversity across access modes



6.6 The higher the user's share of news from social media or search, the less diverse the set of topics they are exposed to relative to the reference category of direct access. For news aggregators we do not find a significant negative association between news aggregator sessions and topic diversity.

6.7 These results are consistent with concerns about how different modes of news access curate and present news. A person who goes on a news outlet's homepage visits only one outlet but will see a variety of headlines on a wider range of topics, much as they would looking at the front page of a print newspaper.

6.8 A social media platform on the other hand might identify the interests of the user and try to find articles to satisfy those interests, and in doing so cover a wider range of news outlets, but ultimately a narrower range of topics. This finding is consistent with growing concerns that social media drives echo chambers and is consistent with the high level of ideological segregation in news browsing that other research has observed on Facebook. We discuss these issues in more detail in our main report[49] and in our 2022 Discussion Document.[50]

6.9 In relation to the result for search engines, there is a separate factor at play, since these are driven by user inputs – ie a user indicates in the search term what topics they are interested in. This could explain why we observe the lowest diversity scores for search-based news sessions.

6.10 The effect of news aggregators on topic diversity is negative, but not significant. The lack of significance combines a smaller point estimate (-0.10) compared to the estimates for social media and search engines (-0.20 and -0.47 respectively) and a wider confidence interval which reflects the small share of news aggregators' news sessions in the overall sample – our baseline algorithm only identified 0.6% of news sessions as being from aggregators. News aggregators tend to use a combination of editorially-driven curation of news content and recommender systems on their services and in this respect they may be more similar to a direct news source than social media and search engines.

Table 4: Regression tables for equation *(1)*

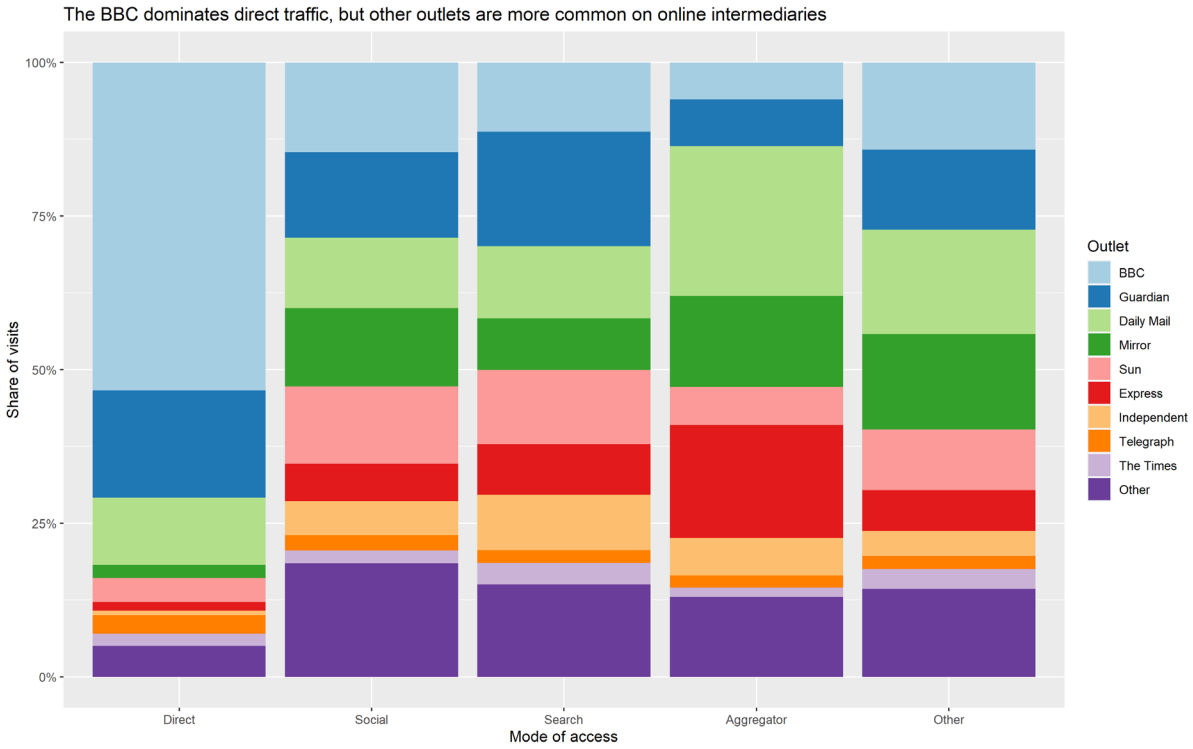
	Topic diversity		Outlet diversity	
Access modes	Estimates	Confidence Interval (95%)	Estimates	Confidence Interval (95%)
(Intercept)	0.66	[0.65 ; 0.68]	0.11	[0.10 ; 0.13]
Social	-0.20	[-0.25 ; -0.15]	0.50	[0.45 ; 0.55]
Aggregator	-0.10	[-0.24 ; 0.03]	0.68	[0.53 ; 0.83]
Search	-0.47	[-0.50 ; -0.43]	0.69	[0.65 ; 0.73]
Other	-0.21	[-0.23 ; -0.19]	0.28	[0.26 ; 0.31]
Observations	3,755		3,755	

[49] Ofcom, 2024, Online news: research update.
[50] Ofcom, 2022, Discussion document: Media plurality and online news, ('Discussion Document, 2022').

	Topic diversity	Outlet diversity
R2 / Adjusted R2	0.181 / 0.180	0.301 / 0.301

6.11 Diving deeper into the results, the distribution of outlets by access mode (Figure 5) indicates that lower outlet diversity in direct access is driven, in part, by the fact that the BBC takes up a very large share of articles accessed directly. Our finding here is in line with previous research (Fletcher, Kalogeropoulos, & Nielsen, 2023).

Figure 5: Shares of outlets across access modes



Source: Ofcom analysis of Ipsos Iris panel-only data, 15 September – 15 October 2021.

6.12 Overall, online intermediaries appear to be sending users to a wider range of news outlets but in doing so are not increasing the diversity of the topics that they access.

Robustness of the results

6.13 To verify the robustness of our results, we considered several modifications to our baseline methodology. In earlier testing, we also considered using alternative sentence embedding models and different specifications for identifying the access mode of user sessions. We decided to use sentence transformers because of their state-of-the-art performance in the academic literature and because they did not require us to specify additional parameters to generate the embeddings. The different specifications for identifying access modes made no substantive difference to our regression results, which is consistent to what we found in our 2022 Discussion Document,[51] so we decided to proceed only with the base scenario.

[51] Discussion Document, 2022. See annex 5.

6.14 We also tested whether our results still hold if we used an alternative method to quantify content diversity by taking the mean of the pairwise distances between the sentence embeddings of all the headlines that each user accessed as in Möller et al. (2020). We chose this modification because in early testing we found that our results were most sensitive to the hyperparameters of UMAP (dimensionality reduction) and HDBSCAN (clustering) in our topic modelling. Since the distance between sentence embeddings does not involve dimensionality reduction or clustering, we can avoid this source of instability altogether.

6.15 The results of this alternative method are in Table 5 below. The results echo the main finding that diversity is lower if news articles are accessed through social media and search engines.

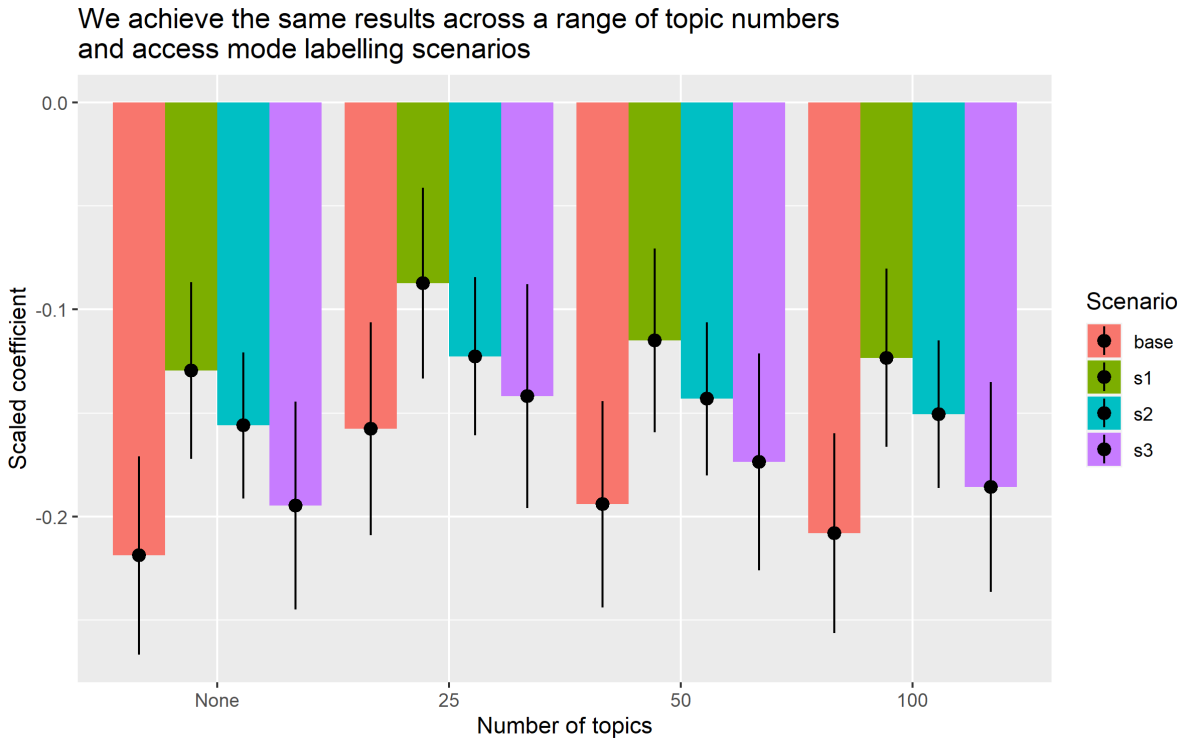
Table 5: Robustness of results

	Pairwise distances	
Access modes	Estimates	Confidence Interval (95%)
(Intercept)	1.29	[1.28 ; 1.30]
Social	-0.15	[-0.19 ; -0.11]
Aggregator	-0.18	[-0.28 ; -0.08]
Search	-0.27	[-0.29 ; -0.24]
Other	-0.16	[-0.18 ; -0.14]
Observations	5,186	
R2 / Adjusted R2	0.090 / 0.089	

6.16 We also re-estimated our baseline regression model using a) a variety of different algorithms for identifying the access mode[52] and b) a range of different constraints for the number of topics. For the sake of simplicity, we will only present the estimated coefficients for the social media share of news sessions, but we also found statistically significant results for search as in the baseline model. Since the number of topics systematically impacts the mean entropy, we also scaled the entropy values to range between 0 and 1. The results are presented in Figure 6. The estimated coefficient on social media is significant and negative in all cases.

[52] These alternative access mode classifications differ from the benchmark classification by varying the maximum time that we permit a news session to last (one hour in the benchmark classification) or the maximum number of steps which we allow a news article to be away from a homepage visit (five in the benchmark classification). See also Data section for a description of the benchmark classification.

Figure 6: Scaled coefficients on social media for robustness checks



6.17 Overall, these checks confirm that our results are robust to a variety of alternative specifications for topic modelling and alternative approaches to measuring the diversity of news consumption.

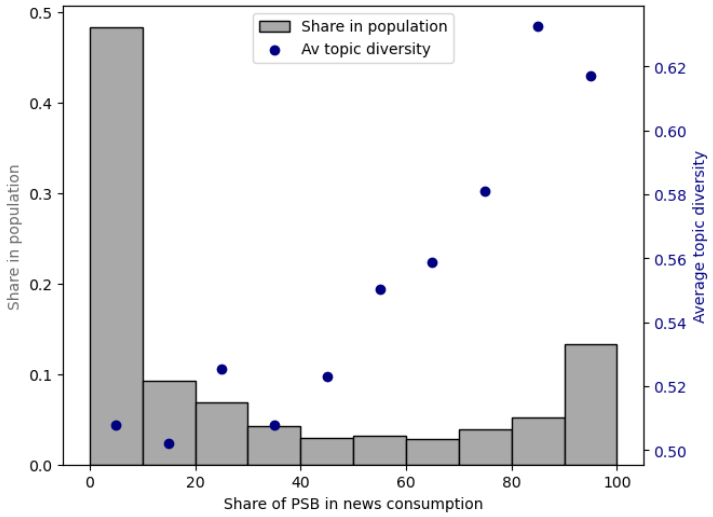
PSBs and diversity

6.18 Given the importance of the BBC in directly accessed news (see Figure 5), we also looked at how topic diversity relates to the share of online content from the BBC and other PSBs (ITV and Channel 4) in a person's news diet.[53] PSBs have a statutory requirement to include news programming of high quality and covering national and international matters and might therefore expose their audiences to a wide range of news topics. Unlike the BBC, the commercial PSBs do not have statutory requirements relating to the provision of news online. However, the BBC, ITV, STV, S4C and Channel 4 all provide written news articles online, alongside video content.

6.19 We group our sample into ten segments of equal width depending on the share of their news which come from PSBs – irrespective of their access mode. The lowest PSB segment covers shares between 0% to 10%, and the highest-PSB segment covers shares between 90% and 100%. Figure 7 plots for each segment the share of people in this segment (the grey bars, scale on left y-axis) and the average topic diversity of people in the bin (the blue dots, scale on right y-axis).

[53] In terms of news consumption the BBC is by far the largest among the PSBs.

Figure 7: Topic diversity and share of PSBs in news consumption



6.20 A large share of the sample does not obtain (or obtains little) news from PSBs, and the diversity of their news diet is relatively low. We also observe a bi-modal distribution of the sample: having a balance of PSB and non-PSB news is relatively rare, and most people either get little to no news *or* most or all of their news from PSBs. The figure strongly suggests that a high share of PSBs correlates with high topic diversity. We corroborate this finding by running regression model (2). The result is presented in table 6. The predicted topic diversity for someone who does not read online news from any PSB is 0.50, whereas the predicted diversity for someone who entirely relies on PSBs for online news is 0.61 (0.50 + 0.11).

Table 6: Regression tables for equation (2)

	Topic diversity	
	Estimates	Confidence Interval (95%)
(Intercept)	0.50	[0.49 ; 0.51]
PSB share	0.11	[0.09 ; 0.12]
Observations	3,134	
R2 / Adjusted R2	0.045 / 0.045	

6.21 These findings are robust to the chosen number of topics: The topic model algorithm freely estimates the number of distinct topics.[54] We thus also ran the algorithm restricting it to identify 25, 50, or 100 topics exactly which all produced very similar results. We also checked the sensitivity to our results to including Sky within a group alongside the BBC, Channel 4, and ITV. This did not make an important difference to our results.

[54] The algorithm identifies 120 topics.

7. Discussion and conclusion

7.1 Using a sample of the UK population for whom we had data on their app usage and browsing histories for one month in 2021, we have analysed how the diversity of their news diets relate to the way they discover news online. In particular, we have looked at diversity in terms of outlet concentration, and in terms of topic concentration. We distinguished mainly between news discovery through OIs on the one hand and through the news outlets' homepages and apps on the other.

7.2 Our research confirms the previous finding in the literature that news discovery through OIs is associated with higher outlet diversity. However, we also find that news discovery through OIs (in particular, social media and search engines) is associated with lower topic diversity. This latter finding points to a need to refine our understanding of news diversity and re-evaluate the view that OIs are neutral or beneficial in adding to the diversity of news viewed by news consumers.

7.3 We also find that people that get a larger proportion of their online news from a PSB have a higher diversity of topics in their news diet and that people that make little or no use of PSBs online have a lower diversity of news topics.

7.4 We acknowledge some limitations of our research. Our sample is not a random cross-section of the population, and we have analysed a snapshot of news consumption in a context which is dynamic: news consumption habits and how OIs interact with them is and has been in flux. As such, our study might not generalise to the population of the UK, or over time. Further, while web-tracking data as employed in this study has opened up new research opportunities, it remains imperfect in capturing the entirety of a person's news consumption. In particular, we do not observe the extent to which people's offline consumption substitutes or complements their online consumption.

7.5 Perhaps most importantly, the research does not prove there is a causal relationship between the use of OIs and PSBs and news topic diversity. Our findings are certainly compatible with a causal interpretation, which could have important policy implications. [55] However, they can also be interpreted in different ways. For example, people with relatively narrow news interests might not choose to visit news outlets' homepages, or perhaps people combine different discovery methods for different purposes.

[55] See for example, Mattis et al, 2022, *Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design*. New Media & Society. Helberger, 2015, *Merely facilitating or actively stimulating diverse media choices? Public service media at the crossroad*. International Journal of Communication.

A1. Responding to this Economic Discussion Paper

How to respond

A1.1 If you would like to respond to the analysis in this Economic Discussion paper, or on the use of these analytical tools in general, you can reply using any of these options.

A1.2 You can respond by email to edp.responses@ofcom.org.uk. If your response is a large file, or has supporting charts, tables or other data, please email it to edp.responses@ofcom.org.uk, as an attachment in Microsoft Word format, together with the cover sheet.

A1.3 Responses may alternatively be posted to the address below, marked with the title of the EDP:

Economics and Analytics Group

Ofcom

Riverside House

2A Southwark Bridge Road

London SE1 9HA

A1.4 We welcome responses in formats other than print, for example an audio recording or a British Sign Language video. To respond in BSL:

- send us a recording of you signing your response. This should be no longer than 5 minutes. Suitable file formats are DVDs, wmv or QuickTime files; or
- upload a video of you signing your response directly to YouTube (or another hosting site) and send us the link.

A1.5 We do not need a paper copy of your response as well as an electronic version. We will acknowledge receipt of a response submitted to us by email.