Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

The News We Like Are Not the News We Visit: News Categories Popularity in Usage Data

Zied Ben Houidi,[1*] Giuseppe Scavo,[1,2*] Stefano Traverso,[3] Renata Teixeira,[2] Marco Mellia,[3] Soumen Ganguly[1*]
[1]Nokia Bell Labs*, {zied.benhouidi}@gmail.com
[2]Inria, {first name.last name}@inria.fr
[3]Politecnico di Torino, {first name.last name}@polito.it

Abstract

Most of our knowledge about online news consumption comes from survey-based news market reports, partial usage data from a single editor, or what people publicly share on social networks. This paper complements these sources by presenting the first holistic study of visits across online news outlets that a population uses to read news. We monitor the entire network traffic generated by Internet users in four locations in Italy. Together these users generated 80 million visits to 5.4 million news articles in about one year and a half. This unique view allows us to evaluate how usage data complements existing data sources. We find for instance that only 16% of news visits in our datasets came from online social networks. In addition, the popularity of news categories when considering all visits is quite different from the one when considering only news discovered on social media, or visits to a single major news outlet. Interestingly, a substantial mismatch emerges between self-reported news-category preferences (as measured by Reuters Institute in the same year and same country) and their actual popularity in terms of visits in our datasets. In particular, unlike self-reported preferences expressed by users in surveys that put "Politics", "Science" and "International" as the most appreciated categories, "Tragedies and Weird news"' and "Sport" are by far the most visited. We discuss two possible causes of this mismatch and conjecture that the most plausible reason is the disassociation that may occur between individuals' cognitive values and their cue-triggered attraction.

1 Introduction

The interest in the exchange and consumption of news is one of the oldest human habits (Stephens 2006). What has changed over time is the medium to disseminate news: from word-of-mouth, to written, to printed, to broadcasted and finally to online publishing and social media sharing. The shift from print to online offers new opportunities to study people's access to news, which human scientists, but also journalists and news organizations can leverage to understand readers' interests and better adapt their palette of content.

The literature is rife with data-driven studies of online news consumption (see Sec. 2). These studies characterize various aspects, for example the virality of news (Kourogi et al. 2015), news coverage in social media compared to traditional media (Olteanu et al. 2015), or user preferences compared to editors' suggestions (Boczkowski 2010; Abbar et al.). However, as with any empirical work, these studies are limited by what their input data allows to observe.

We identify four types of data sources used so far to characterize online news. (i) The first is published content. For example, projects like GDLET (Leetaru and Schrodt 2013) and EventRegistry (Leban et al. 2014) crawl online news articles worldwide and make them available to researchers. Although this approach faithfully captures the "supplier" part, such data does not convey what users really consume. (ii) The second source is what is publicly shared in popular social networks (Osborne and Dredze 2014). Although this approach nicely captures how users relay and react to news, it misses what happens outside the social network. (iii) The third source is collecting *usage data* statistics from an individual online newspaper (Boczkowski 2010; Abbar et al. ; Dezsö et al. 2006), a comprehensive view but limited to one news outlet. (iv) Finally, researchers and practitioners counter the partial visibility of these sources with more traditional sources like user surveys. Institutes like Pew Research and Reuters regularly issue survey-based reports (Mitchell and Page 2014; Newman, Levy, and Nielsen 2015) tracking various aspects of news consumption habits worldwide. Although they offer insights on larger aspects of human behavior and preferences, it is not clear how precise user explicit feedback is in capturing actual usage of online news.

In this work, we uncover a different source of data and contrast it, to the extent possible, to the previously described four existing data sources. In particular, we extract and analyze online *news visits* made by a population of users. We define a news visit as a click to a web page containing a single news article. In contrast to prior work focusing on a single news outlet (e.g., Dezsö et al. (2006)), we observe all news visits of individuals in a population of users across any online news website they visit when connected to their Internet provider network. More precisely, we extract online news visits by passively observing traffic traversing a network link, from which we extract the pages containing news that users visited. We perform our analysis on data collected for about one year and a half at vantage points we installed in Italy: three are located in the network of a large Internet Service Provider (ISP) and the forth is located in a large university campus. During this period we observe 80 million visits to 5.4 million

*Work performed when the authors were at Nokia Bell Labs.

distinct articles overall.

In this paper, we focus in particular on the popularity of news *categories*. We define a category as the general theme of a news article, as directly assigned by the publisher, as opposed to the specific story or event it covers. News categories correspond thus to the sections under which news editors often group their articles (e.g. Sport or International). We quantify the popularity of news categories in terms of number of visits and put it in perspective with their popularity considering various sources: (i) news discovered on Facebook, (ii) news visits to major news outlets, (3) published articles, and finally, (iii) user preferences expressed in surveys (Sec. 5). To estimate the view that social media and major news outlets have, we analyze the referral of the news visits, i.e., the previous page from which the visit came, to study how users discover news articles and quantify the importance of social networks and major publishers in the overall consumption volumes (Sec. 4). We regularly crawl all news outlets in Italy for a duration of three months to get a dataset of all published articles. Finally, to better assess the significance of the category popularity, we study their stability across locations (spatial) (Sec. 7) and across time of day and months (temporal) (Sec. 6).

We summarize our findings as follows.

• Usage data obtained from the network complements existing data sources well. At most 11% of news visits in our datasets come from social networks (Sec. 4). As such, studying only news shared on social media captures a small fraction of news visits—as already pointed by Flaxman, Goel, and Rao (2016). Popularity of categories as inferred from news discovered on Facebook does not match the popularity obtained from the analysis of all data. In particular, "Tragedies and Weird" news are over-represented on Facebook, while "Sport" is under-represented. Similarly, the popularity of categories we observe when considering all visits does not match that observed individually by a major national news outlet either.

• "Tragedies and Weird" and Sport are the most popular categories of news in our datasets (around 16% of visits each). Politics is surprisingly less popular with 3.5% of visits. Overall, despite notable differences, the popularity of many categories follows the supply, or what is published by editors. Surprisingly, Reuters' survey estimates do not match our data. For instance, respondents report International and Politics news as some of the most popular categories, far beyond Sport and Weird news, suggesting a dissociation between what users prefer and what they actually consume or are attracted by (Sec. 5).

• We discuss two possible causes to explain this dissociation. The first is social desirability bias whereby users would show a preference for categories that give a better image of themselves. The second, inspired by the incentive-sensitization theory of addiction (Berridge and Robinson 2016; 1998) in neuroscience, postulates that what we observe is a legitimate difference between what people "want" (the motivational part of a given reward) and what they "like". Our discussions with two experts (the professor of political communication and editor of the Reuters survey-based report, and the biopsychologist who formulated the incentive-sensitization theory of addiction) indicates that the second cause is more likely (Sec. 8).

• News-category popularity is overall stable across locations, despite demographic differences between the populations we study (Sec. 7). Albeit overall stable, the popularity of certain categories slightly varies depending on time of day or certain events of the year (e.g., vacations, beginning and end of the sports season and major events like the Paris terrorist attacks) (Sec. 6).

2 Related Work

Since the introduction of online newspapers, researchers in different communities have seen new opportunities to study news consumption. In this paper, we study online news with a focus on the difference between various data sources.

Comparison of various news sources. Some of the related work compared various sources of news in terms of coverage. Zhao et al. (2011) use topic models to compare the topics in news shared in Twitter with those of the New York Times for a period of three months. The authors found that Twitter covers more personal life and pop culture and that users tweet less about world events, but retweet a lot, which causes the news to spread. Olteanu et al. (2015) compare the coverage of news about climate change in traditional news media and in Twitter. They show that the scope of the traditional news media they consider and Twitter are different. Chakraborty et al. (2016) also compare the topical coverage of the New York times, Twitter and Facebook, and highlight other differences. Kwak et al. (2018) compare the gap between "media attention"– measured as popular topics at a news aggregator – and "public attention" – measured with Google Trends – on international news coverage across hundreds of countries. Finally, Boczkowski (2010) relied on usage data from a specific news outlet to measure the gap between what publishers think is interesting and what users really read. Similarly, Abbar et al. () measure the gap between the geographical interests of users and the geographical coverage of Al-Jazeera.

In this work, we not only compare news categories discovered on social media and those published by news outlets, but we additionally compare these categories to (i) categories considering all news visits of a population, (ii) those considering only visits to major news outlets and, finally, (iii) those preferred by users in surveys. We similarly find differences between published articles and those visited from Facebook. In addition to prior work, we find that focusing on a single news outlet, although very popular, does not capture the overall behavior of users when including all outlets. Hence, what is learned from a large single outlet cannot be representative of all traditional news media. Unlike prior work, ours spans visits to all news outlets, for different populations of users and for a long period of time. This allows us to observe how habits change over time and regions.

Usage of surveys. Surveys remain a valuable tool to understand user behavior online (Mitchell and Page 2014; Newman, Levy, and Nielsen 2015; Prior 2009; Lee and Chyi 2014). Research institutes like Reuters or Pew Research regularly issue survey-based reports tracking various aspects of

user behavior. Here, we contrast popularity inferred from usage data with self-reported preferences in surveys (Newman, Levy, and Nielsen 2015).

Similarly to our work, Lee and Chyi (2014) question the perceived noteworthiness of consumed news. The authors rely on a survey to show that only about one third of the content produced by the mainstream news media is perceived as noteworthy. Our work shows that what users value might not necessarily match what they will actually view.

Surveys are also known to suffer various biases. For what concerns news, in a prominent work, Prior (2009) has confirmed anecdotal evidence suggesting the inability of surveys to accurately measure news media exposure. In particular, by directly comparing Nielson's audience measurement estimates to survey-based estimates, Prior showed that self-reports tend to overly inflate media exposure, by a factor of three on average and up to eight for certain demographics. Our work complements his by comparing survey estimates and usage data based on the finer granularity of what users value (news-category preferences) and not only exposure. We reveal indeed a major difference between self-reported preferences and actual views.

News on social media. Less directly related to our work, in recent years, social media has risen as a means to share news. This has led to the explosion of social media data-driven studies of news. We report few examples. Kwak et al. (2010) show that 85% of the topics in Twitter are related to news headlines. Osborne and Dredze (2014) compare the performance of Twitter, Facebook and Google Plus in terms of news coverage and latency. They find that Twitter is faster while the other two offer more diversity. Saez-Trumper, Castillo, and Lalmas (2013) identify different kinds of biases in news shared on Twitter. Wang and Mark (2013) study the news consumption from social media in China. They characterize different types of audience for different types of news. Kourogi et al. (2015) extract features from the headlines and text of tweeted news and use an SVM ranker to infer features that can predict the virality of a news article. Morgan, Lampe, and Shafiq (2013) show that users sharing news on Twitter present no bias based on their perceived ideology of the news outlets, i.e., they see no bias due to selective exposure in their datasets. Flaxman, Goel, and Rao (2016) study whether social media helps increase exposure to diverse perspectives or lead to ideological segregation by creating filter bubbles. In their analysis, they find that most news consumption comes from people directly visiting mainstream newspapers web sites, with only a small fraction of news visits coming from Facebook. In this paper, we confirm that only a small fraction of news consumption in our datasets comes from social media and highlight a difference in the popularity of certain news categories when considering overall visits and those coming from Facebook.

3 Usage data: a new observation point

In this paper, we rely on datasets that gives us a privileged vantage point to study news categories' consumption habits. In this section, we present our datasets, the ethical issues around their use and how they complement existing sources.

Trace	Network	City	Households	Surfers	Distinct news
ISP-City1	ISP	City1	12,193	285,038	1M
ISP-City2-a	ISP	City2	1,760	116,730	0.8M
ISP-City2-b	ISP	City2	11,520	493,184	1.8M
Campus-City2	Campus	City2	6,635	1,102,348	1.8M
Published-articles	-	-	-	-	80k

Table 1: Description of datasets (collection from January 2015 to May 2016).

3.1 Datasets and methodology

Our analysis relies on a record of news visits we extract from passive Internet traffic observation on four different networks with tens of thousands of users who access the Internet from their PCs, smartphones and tablets. We first describe our methodology to extract news visits from the raw traffic.

Data collection. We collected data from four networks: at one university campus in Italy (*Campus-City2*) and in a large residential ISP in two different Italian cities (*ISP-City1*, *ISP-City2-a*, and *ISP-City2-b*). In the first scenario, we monitor the link that connects the campus LAN and WiFi networks to the Internet. In the residential scenarios, we monitor traffic from one district in one city (*ISP-City1*), and two different districts in another city (*ISP-City2-a* and *ISP-City2-b*). Here, users connect to the Internet thanks to ADSL or FTTH access, which is shared through WiFi inside their houses.

The traces were obtained by running Tstat (Finamore et al. 2011). Tstat processes network packets and extracts raw HTTP requests. More precisely, for each HTTP request, it extracts the requested URL, the referral (which captures the URL of the previous page visited by the user, if any), an anonymized IP address, the timestamp of the visit, and the user agent (which describes the browser or the app originating the request). All in all, we analyze tens of billions of HTTP connections spanning the period from January 1st 2015 to May 15th 2016.

We rely on a fifth dataset, *Published-articles* obtained by crawling the 667 major news outlets in Italy for a period of 3 months. This allows us to contrast what is published with what is consumed. All in all, we collect 80k articles.

From HTTP traces to visits. HTTP traces contain a large number of requests to images, scripts, or ads that compose webpages, but that do not correspond to the page users actually visited. Hence, we first need to extract visits to webpages. For this, we use a Hadoop cluster and adapt algorithms from our earlier work (Ben Houidi et al. 2014) to extract user visits all HTTP requests.

Identifying news articles. From the set of visited URLs, we keep only those URLs that correspond to well-known online newspaper sites. Defining which webpage corresponds to a news article is subjective and tricky. In particular, lists of news outlets used in prior studies (Bakshy, Messing, and Adamic 2015; Ribeiro et al. 2018) focus mainly on English-speaking news outlets and hence are unfit to identify news articles in our Italian-based dataset. To overcome this issue, we rely on the Google News authority and consider all publishers that it indexes as potential news outlets. To build our list, we crawl Google News Italy once every 20 minutes during one month and build a list of unique outlets. The resulting

list has 667 online outlets that we make available online (onl online April 2019). Google News does a thorough job in indexing news. Our manual inspection of outlets in the list shows that it covers all the popular online outlets in Italy including some influential blogs. Our analysis, however, may miss some foreign or unpopular niche newspapers. Table. 1 provides for each of the four locations the number of distinct news articles that we observe.

Assigning categories to articles. Finally, we label news articles with categories. Assigning a category to a news article is also difficult and subjective. To counter this, we rely on the sections under which editors publish articles. So, similarly to prior work (dos Rieis et al. 2015), whenever available, we extract the category of news assigned by the news editor. Using this method, we identify first 167 unique section names in the four datasets. Since different editors may give different names to the same categories (e.g., use of plural and singular), we manually merge similar categories. Similarly, we group sub-categories into larger ones by hand. We hence obtain 40 distinct categories. This allows us to confidently label more than 70% of news articles. The remaining 30% are hard to automatically label. By manually inspecting them, we observed that most of them come from minor unpopular websites that do not adopt proper labeling of sections.

To shed light on the composition of some ambiguous categories, we manually verify three categories that sometimes overlap with others:

• *International:* Many articles are published under the International section, but it is difficult to further pinpoint the category of the news article beyond the fact that it happened abroad. We manually label 200 randomly chosen articles from this category and find that 60% is about "Tragedies/Weird" news and 16.5% of International politics. An additional 11% are a mixture between politics and tragedies (wars, fraud, terrorism attacks, etc.).

• *Ed. Columns:* Many articles are grouped by editors under various names like "columns", "opinions" or (special) "editions". For the ease of the presentation, we group all these articles under the label "Ed. Columns". Similarly, labeling a 200-article sample, we find that such columns cover a large number of topics, with "Tragedies/Weird" having 25% of articles and Politics 17%.

• *Tragedies/Weird:* One popular section referred in Italian as "cronaca" is notorious for reporting mainly Crime stories. We manually label 200 randomly chosen articles from this category and find that it contains 42% of Crime-related articles, 31% are about natural disasters and accidents, and around 10% are weird news. For the ease of the presentation, we refer to this category as "Tragedies/Weird", but it contains a majority of crime news.

We manually review samples of the rest of the categories as well to confirm that articles are correctly labeled. We also publish the scripts we use to group editor sections into categories as well as the 80K articles in *Published-articles* dataset and their associated categories (onl online April 2019) to allow further verification of our methods.

Identifying users: Households and surfers. One challenge for our analysis is that the traces have no per-user identification. A single user may appear multiple times in the traces, because she may connect from multiple devices at home. Conversely, multiple users may share the home gateway's IP address, e.g., members of a household connected to the Internet through the same ADSL/FTTH gateway. To overcome this, we study "users" at two granularities: surfers and households. We identify a *household* by the anonymized IP address. Because of the static assignment by the ISP of IP addresses to home gateways, this corresponds to a household in ISP traces.[1] We define a *surfer* as the concatenation of the anonymized IP address and the user agent. A surfer captures a particular user surfing the web from a particular browser or mobile app. As documented in our prior work (Vassio et al. 2018), we observe that smartphones and tablets account for more than 50% of devices in the datasets and generate around 20% of page visits. Note that with this definition, the same person might still appear as multiple surfers. This is particularly likely if we take into account the long duration of our traces and the fact that user agents change with software upgrades. Neither surfers nor households defines precisely a user, we will use them both with caution to approximate per-user behavior.

Table. 1 shows the numbers of observed households and surfers in the traces.[2] Finally, unless otherwise stated, we consider the number of visits to a particular article as the number of distinct surfers that clicked on the article. Note that counting the popularity of articles considering the number of distinct households instead of the number of distinct surfers does not considerably change the results: only very few articles are read by more than one surfer in a household.

3.2 Ethical Issues

Privacy protection mechanisms implemented in Tstat have been devised in close collaboration with ISPs' privacy officers and legals (Trammell et al. 2014). Tstat processes packets in real time and obfuscates any Personally Identifiable Information (PII). IP addresses are anonymized using consistent and irreversible hashing functions – so that it is impossible even for the ISP network administrators to link traffic to a customer identity. In this work, we instrumented Tstat to collect the minimum information required for our analysis. It only logs URLs and referrals and user agent strings. Moreover, we instrument Tstat to sanitize URL to avoid exposing users' identifiers such as email addresses, by, e.g., removing query parameters after the "?" character. Web cookies and locations are not logged. Note that Tstat has no visibility on encrypted traffic (HTTPS), where the sensitive information concentrates.[3] Our analysis ignores the traffic

[1] In the case of the campus traces, it often corresponds to a dynamically allocated IP address.

[2] The number of observed surfers is large due to the presence of multiple devices behind the same IP, but mostly because of software upgrades. Over a week, we observe on average only three user agents per IP.

[3] Although more and more web sites use HTTPS, all the news outlets relied on HTTP at the time of our measurements.

generated by users who activate the Do-Not-Track flag in their clients (observably less than 1%).

Our traffic monitoring activity and data collection obtained the approval of the Security and Privacy Offices of the campus and the ISP in which we deploy our traffic monitoring probes. We discussed the opportunity to make customers aware of the data collection. Given the anonymity and the technical precautions we adopted to preserve users' privacy, this was considered to be not required. With the 2018 European General Data Protection Regulation (GDPR), however, this type of data collection does require explicit user consent. Thus, future data collection efforts of this type will become more challenging in Europe.

Finally, the data is only available to the researchers for the sake of pursuing their research objectives, which involves only the study and publication of aggregated results. Hence, none of the analysis or the results we present in the paper can be linked back to a single user. The data analysis (which was conducted post-GDPR) was subject to a privacy impact assessment, conducted with the data protection officer of Inria. Given that the datasets were already available, that they do not contain any PII, and that the analysis builds on aggregated trends, the conclusion was that no additional formal ethical review was required for the analysis.

3.3 Limits

Our methodology and datasets have some limits that we summarize hereafter.

• First, our perspective is spatially constrained: we can monitor news browsing behavior (on both mobile and laptop) only when users are connected to the networks we monitor, which correspond to either their home or work. As such, we miss online news consumption when users connect to the Internet using cellular networks, e.g., when they read the news from their mobile while travelling or commuting.

• Second, our dataset focuses only on actual clicks to news articles and can not capture when users get exposed to online news without visiting the news article, for instance by seeing the preview as shared on a social network.

• Third, we lack a precise way to identify a user. To this end, we mainly focus our analysis on properties exhibited by aggregated populations of users. Whenever needed, we carefully rely on the notions of surfer and household to approximate users. This limitation also prevents us by design from obtaining demographic information about users to be used in the analysis.

• A fourth limitation is that this dataset comes from a single country. Therefore, our findings might not generalize to other regions and other cultures.

• Finally, especially when later compared to surveys, our dataset lacks of provable guarantees of representativeness of users in the studied country. In general, whenever applicable, we tested the statistical significance of our results. Given the large scale of our dataset, when comparing empirical CDFs for instance, even barely visible differences between CDFs were significant up to 3% of significance level. So while we can assess how our results and data (which is considerably large both in time and space) are significant and representative for the populations we study, we have no guarantees that such populations are representative of the country. For instance, although the populations are diversified from various cities and various environments (e.g., campus vs. work vs. residential), our dataset does not include populations from rural scenarios. We also have, for obvious privacy reasons, no access to information about the demographics of the populations we study. To alleviate this issue, we study the stability of our results across various locations/sample populations and time spans. We also compare various statistics obtained from our usage data to those available for the studied country. For instance, the penetration rate of both Facebook and Twitter is similar to the stats available country-wide (as shown in Sec. 4). Also, not shown for brevity, available stats on market shares of various web browsers and devices are comparable to what we observe in our data.

	ISP-City2-b		*ISP-City1*		*ISP-City2-a*	
	Apr'15	Apr'16	Apr'15	Apr'16	Apr'15	Apr'16
Self Referral	58%	57%	60%	57%	69%	64%
Facebook	11%	11%	9%	9%	4%	5%
Google Search	10%	10%	9%	9%	9%	9%
Direct Browsing	12%	13%	14%	15%	10%	14.5%
Google News	1.25%	1%	0.8%	0.8%	0.6%	0.9%
Twitter	0.27%	0.27%	0.16%	0.18%	0.12%	0.13%

Table 2: News referral share in residential areas (based on April 2015 and April 2016)

4 Weight of social media and online publishers

Both social media and online publishers represent standard sources to study news consumption. To understand how usage data complements these sources and infer their importance, we study the news' referrals, i.e., the pathways from where the visit to the news article came. Using this approach, we can measure the share of visits in our datasets that comes from online social networks and online publishers.

Tab. 2 presents the results for the three residential areas for two separate months, April 2015 and April 2016. The table shows a ranking of the most important sources and their respective shares.

First, note that "Self referral" means that the visit comes from a webpage within the same website. The "Direct Browsing" class corresponds to visits with no referral. This may be due to direct visits (e.g., bookmarks) or often to links obtained via alternate channels like email or messaging apps that do not pass the referer, something sometimes called the "Dark social Web".[4] Note also that we verified that Google, Facebook and Twitter always passed the referral at the time of the data collection. Hence, we fortunately do not miss their referral traffic and can precisely estimate their share. Finally, it is worth noting that the impact of HTTPS on our referral analysis is limited. In the worst case, the referral share of other possible unknown HTTPS websites is upper bounded by 10-15%, which is the share of "Direct Browsing" (no referral).

[4]https://www.theatlantic.com/technology/archive/2012/10/dark-social-we-have-the-whole-history-of-the-web-wrong/263523/

Overall, with the exception of the few variations that we discuss below, the results are stable across locations and especially across months. Not shown on the table, we verify that the results change only slightly between months.

Regardless of the location, most of the visits come from the homepage of the online newspaper itself. Indeed, 57% to 69% of all news visits are self referrals. This result implies that users mostly rely on visiting the website of their favorite online newspaper to discover news.

Far below, the next most frequent referrals are "Google search"[5] (around 10%), "Direct Browsing" (12-15%) and "Facebook" at around 10%, with the exception of *ISP-City2-a*(around 5%), which we explain later. Twitter's share is much smaller (0.1-0.27%). Google News accounts for around 1% despite 22% of surveyed people in Italy claim to use it to read news (Newman, Levy, and Nielsen 2015).

Two sources of data show visible variations across locations and across months. The first is the "Direct Browsing" which steadily increased in one year. We believe that this is due to the increase of HTTPS adoption, which implies that less referers are passed; more visits appear to us as "No referer". The second is Facebook, which has around 10% in *ISP-City2-b*and *ISP-City1* but only 4 to 5% in *ISP-City2-a*. We conjecture that this is due to the fact that *ISP-City2-a* aggregates traffic for an area which has a lot of office buildings. In a working environment, people may be less likely to use Facebook to discover news. This dataset, as opposed to the others, has indeed higher activity during the day and week days and lower activity at nights and weekends.

Finally, the relatively small share of social networks (11% at most) is somewhat surprising. To verify that our datasets do not introduce a bias concerning social networks, we study the percentage of active Facebook and Twitter users.[6] We find that 65% of surfers in our datasets visited Facebook in 17 months, and only 7% visited Twitter. These percentages are similar to the available statistics of Facebook and Twitter usage in Italy (Manson online May 2016; Statista online May 2016). This result also confirms the study of Flaxman, Goel, and Rao (2016), which found a small fraction of news visits coming from Facebook.

Notice that the immediate referral metric might underestimate the weight of social media. For instance, users coming from Facebook and landing on a news article may end up visiting a number of other articles on the site. We analyze the whole chain of referrals starting from Facebook. We find that the vast majority of visits leaving Facebook stops only one page away. This behavior is documented in more details in our prior work (Vassio et al. 2018).

Since online newspapers are the most influential news pathways with 55% of referral traffic, we study also the percentage of visits that each of them drives individually. We find a large disparity between news outlets. While 80% of online newspapers are referral for less than 0.1% of all news visits, 31% of visits comes from the top 5 newspapers. To better understand this, we show in Fig. 1 the top 20 newspapers in terms of referral share in *ISP-City1* (red bars), together with the referral shares of the same newspapers in *ISP-City2-b* (blue bars). The top newspapers are similar, but present differences in their ranking due to regional preferences. For instance, *outlet-city1-1* and *outlet-city2* are very popular local online newspapers based in City1 and City2, respectively. We explore such a spatial effect in more details in Sec. 7.

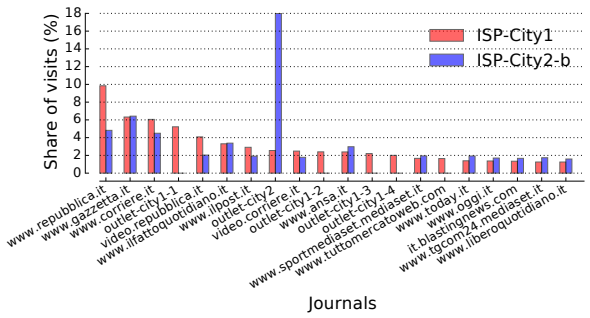


Figure 1: News referral share of the top 20 newspapers in *ISP-City1*, in perspective to the same newspapers in *ISP-City2-b*.

Takeaway. *Although many studies of news consumption are based on news shared in social networks, direct access to online newspapers remains the main source for most users in our datasets. Only 16% of clicks to news articles in our dataset come from social networks. Understanding how and what news are shared is clearly important per se, but these news might only represent a small fraction of what is consumed. Depending on the geographical locations, certain news outlets account alone for up to 18% of all news referral traffic.*

We will further assess in Sec. 5 if data obtained from a single major news outlet or Facebook can be representative of news-category usage behavior at large.

5 Popular News Categories

We now measure the popularity of news categories in our datasets and put it in perspective with the popularity that can be inferred from various other sources.

The first row of Table 3 presents the top 11 most *popular* categories by aggregating data from all our four locations. We define the popularity here by the percentage of visits that each category gets. The top 11 categories capture together around 85% of all visits. We see that the most visited categories are Tragedies/Weird, Sport, and Ed. Columns. Politics is surprisingly unpopular with less than 3.5% of the overall visits. We will show later that, despite notable variations, these results represent a behavior that is stable across time (Sec. 6) and space (Sec. 7).

[5]News visits from Google Search suggest that users either were intentionally looking for news about a particular event, or used the search engine to find the homepage of a newspaper.

[6]Traffic towards these services is encrypted using HTTPS (thanks to the refer field, we can only see visits that *come from* these websites because they always pass the referrer when transitioning from HTTPS to HTTP). Hence, for this task, we use the Tstat TCP and DNS logs available with our datasets.

Usage data (all visits)	**Sport** 16.6%	**Tragedies/Weird** 15.8%	**Ed. Column** 12%	**Entertainment** 7.8%	**International** 7%	**Economy** 6.7%	**People** 4.9%	**Photo Gallery** 4%	**Science** 3.9%	**Politics** 3.5%	**Region** 2.6%
Usage data (visits from Facebook)	**Tragedies/Weird** 23.7%	**Ed. Column** 10.2%	**Entertainment** 8%	**People** 7.8%	**Region** 5.3%	**Sport** 4.3%	**Photo Gallery** 4%	**International** 3.9%	**Health** 3.3%	**Politics** 3.1%	**Science** 2.2%
Usage data (visits Repubblica.it)	**Ed. Column** 20.3%	**Tragedies/Weird** 18.5%	**Sport** 12.6%	**International** 10.1%	**Entertainment** 9.9%	**Travel** 7%	**Economy** 5.1%	**Politics** 3.6%	**Technology** 3.5%	**Cars and motos** 2.9%	**Science** 2.4%
Survey data (users)	**International** 49%	**Politics** 46%	**Local** 44%	**Science** 44%	**Region** 35%	**Health** 35%	**Sports** 33%	**Economy** 30%	**Culture** 26%	**Weird** 19%	**People** 16%
Published data (articles)	**Tragedies/Weird** 15.2%	**Ed. Column** 10.6%	**People** 9.1%	**Sport** 6.5%	**Videos** 5.8%	**Entertainment** 5.2%	**Economy** 5.1%	**Politics** 4.4%	**International** 3.5%	**Region** 3.2%	**Cars and motos** 2.3%

Table 3: Popularity of news categories according to various sources (rounded to the nearest tenth). Note that the last two sources are not directly comparable with the first three due to difference of units: visits (first three) vs. articles vs. users.

Surfers (k=2)	**Surfers (k=5)**	**Households (k=5)**	**Households (k=20)**
256,108 surfers	99,470 surfers	28,799 households	15,504 households
Tragedies/Weird 36.9%	Sport 37.7%	**Tragedies/Weird 67.9%**	**Tragedies/Weird 55.6%**
Sport 32.2%	**Tragedies/Weird 36.1%**	Ed. Columns 62.2%	Sport 50.1%
Ed. Columns 30.6%	Ed. Columns 28.7%	Sport 55%	Ed. Columns 45.4%
Entertainment 23.7%	Entertainment 20%	Entertainment 54.6%	Entertainment 36.3%
Economy 18.7%	Economy 16.8%	People 47.4%	Economy 26.6%
People 17.7%	International 15.9%	Economy 42%	People 24.8%
International 15.9%	People 12.4%	International 35.2%	International 24.5%
Science 11.1%	Photo Gallery 10.4%	Photo Gallery 31%	Photo Gallery 18%
Photo Gallery 10.2%	**Politics 8.5%**	Science 29%	**Politics 13.7%**
Politics 9.9%	Science 8.3%	Region 24.7%	Science 13.4%
Region 9%	Region 6.7%	Technology 24.1%	Region 11.8%

Table 4: Simulation of Top 11 most popular categories (percentage of interested surfers and households).

5.1 Popularity according to various sources

We now put in perspective the popularity we observe in usage data with the one that could be observed by other sources. For social media, we focus on Facebook and extract the subset of visits to news articles with Facebook as referral. We measure the popularity of various categories and report the results in the second row of Table 3. For usage data obtained from major news outlets, we similarly extract the subset of visits to such outlets. We present the results for Repubblica.it in the third row of the table. In the fourth row, we report the popularity of news categories as expressed by users in the Reuters survey in Italy (Newman, Levy, and Nielsen 2015). Note that with the exception of "Local" (somewhat included in Region in our case) and " Weird" (somewhat included in "Tragedies/Weird") news for which we do not have a perfect direct match in our editor-based categories, remaining categories have a suitable match in our dataset. Finally, we apply the same principle on the *Published-articles* dataset to measure category popularity among published articles in the last row.

First, looking at news discovered on Facebook yields interesting differences with the overall usage data. For example, Sport news seem under-represented on Facebook with only 4% of visits originating from Facebook, whereas it represents 16% of visits in usage data. Tragedies/Weird on the other hand is over represented with 23% of popularity in Facebook against 16%. Other categories like Health (less than 1% when considering all visits) and People seem over represented on visits originated from Facebook. One possible explanation for these differences is that what people think valuable to share with others does not necessarily match what they themselves view. This explanation seems inline with a recent study in psychology that analyzed the neural correlates of message propagation (Falk et al. 2013). The study has found that people especially mentalized about what will be appealing to others when propagating a message. However, another explanation could be that the population that uses Facebook to discover news is different compared to the rest of the population. Further investigation is needed to understand these differences.

Second, comparing with the popularity as seen by major news outlets, we also find noticeable differences. In the reported case of Repubblica.it, few categories like Ed. Column, International, Travel, Technology, Cars and Motos are particularly more popular compared to the entire data. This result suggests, at least, that relying on usage data from a single news editor is not representative of user consumption habits at large. Other outlets exhibit other differences.

Our next comparison point is with surveys. At first sight, our usage data seems far from Reuters' (Newman, Levy, and Nielsen 2015) survey-based preference estimates for the same year in Italy.[7] Although 46% of users report being interested in Politics and 44% in Science, each of these categories accounts for less than 4% of the visits. The survey estimates are however expressed in terms of percentages of users. We perform various simulations to investigate in more details the differences between reported self-preferences and actual usage data in Sec. 5.2.

Finally, to put these numbers further in perspective, we study the popularity of news categories in terms of number of *published* articles (and not visits). We remind that a direct comparison is thus not possible due to the difference of units. Interestingly, overall, the percentages of published articles seem to perfectly match the percentages of visited articles for many categories. There are, however, as in the previous cases, noticeable exceptions. Especially, Sport and to a lesser extent International and Science have higher popularity in terms of number of visits compared to published articles. Sport has indeed 16% of visited articles despite only 6.5% of published articles. International has also 7% of visits despite only accumulating 3.5% of published articles. Not shown in the table, only 1.9% of published articles are about Science but almost 4% of the views relate to Science. At the opposite side, the People category has a higher number of published articles (9%) than the actual fraction of visits it attracts (4.85%).

Takeaway. *First, usage data shows that Tragedies/Weird and Sport are by far the most accessed news categories. Despite being liked in surveys by almost one out of two users, Politics and Science attract less than 4% of visits. Second, none of the existing comparable sources of data can capture precisely the news category popularity at large when considering all visits from all users. Each of these sources yields a different popularity distribution that reflects the peculiarity of the data source.*

[7] Note that the survey data has a different unit (users) compared to our usage data so a direct comparison is not possible at this stage.

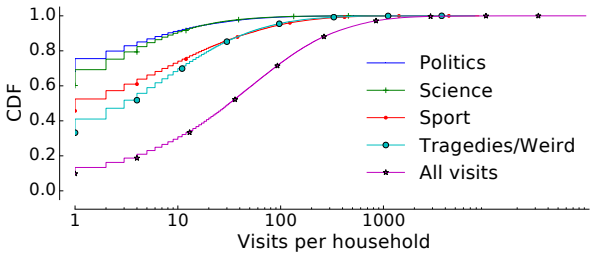


Figure 2: Distribution of number of visits per household for different categories. Results obtained by aggregating data collected in 2015 from *ISP-City1*, *ISP-City2-a* and *ISP-City2-b*.

We next further investigate the difference between usage data and self-reported preferences in surveys.

5.2 Usage Data versus Self Reports

Although the Reuters survey was not conducted on the same exact set of users in our datasets, it was done on a small, yet representative sample of the population in Italy in the same year of our collection. To put both results in perspective and better contextualize our findings, we perform a more thorough comparison in this section.

The Reuters survey counts the popularity in terms of number of users, we count it in terms of total visits. To make a fairer comparison, we run different experiments with the aim of estimating the percentages of users that are interested in each category based on our datasets. We consider surfers and households and assume that a surfer (household) "is interested" in a given category if it has viewed at least k times a news article from that category. We vary the threshold from 2 to 20 visits (in the entire 17 months) and compute each time the fraction of interested surfers (households) in each category, separately for each location and aggregated across locations.[8] For a fair comparison with the survey and to consider a worst case estimation, we obtain the fraction by dividing this number, not by the total number of surfers (households), but by the volume of surfers (households) which have shown interest in at least one category.

Intuitively, when the threshold is low, category popularity is inflated, which gives us an upper bound on the number of surfers and households interested in each category. We report the results for two thresholds, for the entire same year as the survey, in Table 4. With $k = 2$ for surfers and $k = 5$ for households, we find that at maximum 10% of surfers and 24.08% of households are "interested" in Politics. These drop to 8.55% and 13.7%, respectively, when considering higher thresholds.

Overall, unlike the survey report, Tragedies/Weird and Sport are still by far the most "interesting" categories for both surfers and households. Despite our lack of precise user identification, results in Table 4 show a strong mismatch between what users report to prefer and what they actually consume. We further discuss possible reasons in Sec. 8.

[8]We exclude the campus trace because of the difficulty to define surfers and households due to dynamic allocation of IP addresses.

Finally, to complement the above simulation, we study the distribution of visits per household. Fig. 2 shows the empirical distributions of visits to each category across households. The figure shows that there is a large variation amongst households for all categories. Yet, Politics and Science span much less users compared to Sport and especially Tragedies/Weird news.

Takeaway. *Unlike their preferences in surveys, users seem to consume news articles from more "catchy" categories such as Tragedies/Weird, Sport and Entertainment, while Politics, Science and Economy are far less popular.*

This result could reveal a social desirability bias that survey institutes should better account for. Other explanations are also plausible: (i) users actually prefer certain categories but cannot resist the "urge" of clicking on other appealing categories, (ii) users consume what the supply provides. The latter is, however, less plausible. First, there exists categories for which the demand does not match the offer (e.g., People has more articles than visits). Second, editors try to publish what they estimate attractive for users. We will further discuss these possibilities in Sec.8.

The percentages of category popularity in our data are representative of the populations we study. Sub-sampling both surfers and households from this population does not change the popularity of categories. However, one question for us is how stable are these results in both space and time. We will explore this question in the next sections.

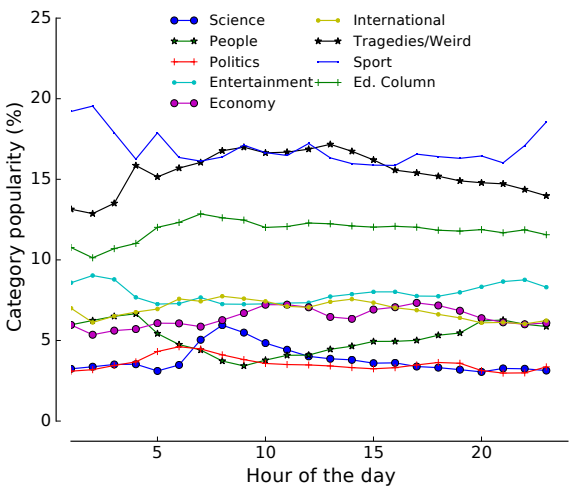
6 Temporal stability

We now study how stable is the popularity of categories over different hours of the day and across months.

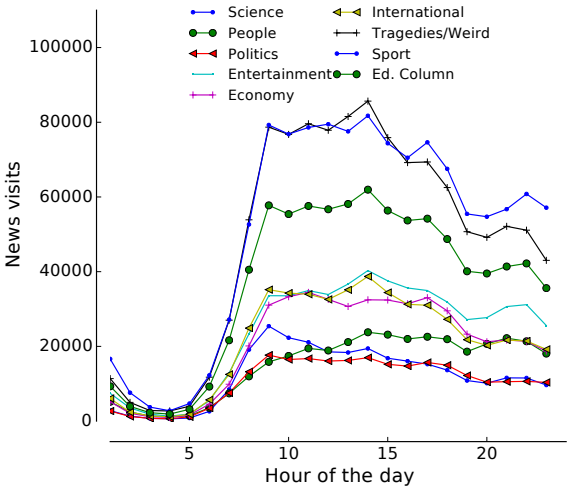
6.1 Over time of day

Fig. 3(a) shows the percentages of visits per news category at various hours of the day. Fig. 3(b) complements it by showing the absolute numbers of visits for each category. Both figures are based on aggregated data, across the 17 months and various locations. We notice that, for the periods with the highest activity, i.e., from 9:00am till 11:00pm, the percentages of each category seem to be overall stable. There are however few curious variations. "Tragedies/Weird" attracts more visits than Sport in both percentages and total number of visits around midday, before leaving the first place to Sport during the afternoon. This trend continues till late at night where Sport reaches its highest share compared to "Tragedies/Weird" (e.g. 19% vs 13% at midnight). This could be due to the fact that sport events usually happen at night. Another noticeable variation is Science which reaches a peak around 6% between 8:00am and 9:00am then lays around 3% the rest of the day. Interestingly, the last two trends are observed in each location separately, including in residential areas, and resist to sampling (across households).

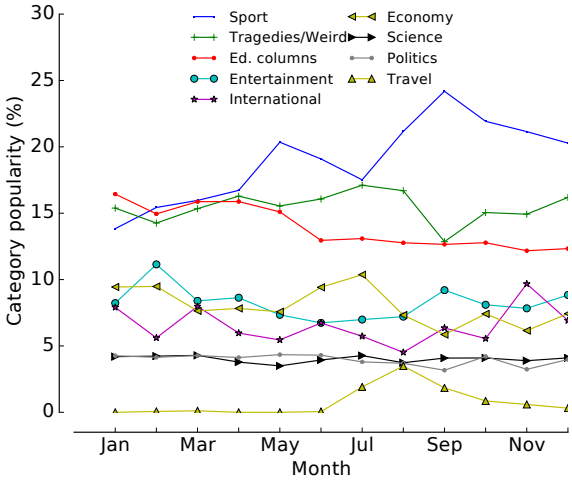
Finally, notice the low number of users between 1:00am and 8:00am. Statistics during this period need thus to be interpreted with caution. Nonetheless, various locations curiously exhibited similar trends during this time: a much higher proportion of Sport compared to Tragedies/Weird and a rise of Entertainment.



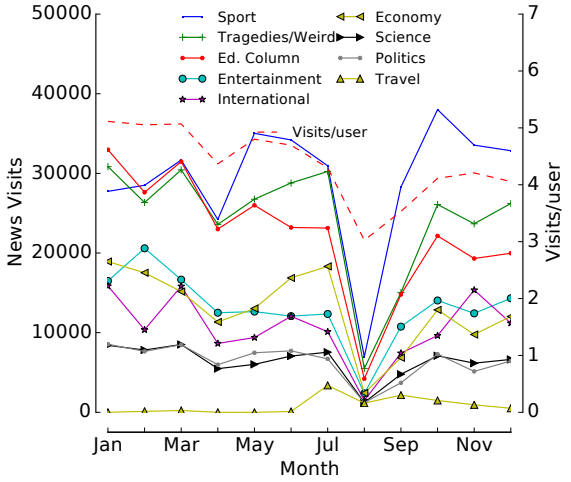
(a) Time of day (Percentages)



(b) Time of day (total views)



(c) Over months (Percentages)



(d) Over months (Total views)

Figure 3: Category popularity evolution over time.

6.2 Over months

Fig. 3(c) shows the percentage of visits to each news category at each month of 2015 in one of the locations. Fig. 3(d) complements it by showing the absolute numbers of visits, as well as the mean number of visits per active surfer during the month, a metric that approximates the activity of users during the month. The major decrease in August is due to holidays and people moving out of the cities.

Analyzing the figures, we find that, although being overall stable, the popularity of a couple of categories varies depending on cyclic or special events of the year. For instance, Sport, which has the most noticeable variation, peaks around May and September, which correspond to the end and the beginning of the sports season. Travel and, not shown, Photo Gallery exhibit a peak around the summer vacation. The latter, by manual inspection, is boosted by photos of celebrities during their summer vacation. Entertainment peaks in February, boosted by a famous country-wide festival. Finally, International shows a peak around November 2015, the period of the terrorist attacks in Paris. During this period, the number of visits per surfer slightly increased but the absolute number of visits for all the remaining categories has decreased, suggesting that the Paris attacks event impacted the other categories. Finally, performing the same analysis on a weekly basis yield no unexpected noticeable behavior. Most of the time, the rankings and the overall popularity are comparable to the yearly data. Interestingly, a category like Politics was remarkably stable. During our 17 months of analysis, however, we were not aware of any major or local political event. To further explore this, we split the category popularity on a daily basis looking for peaks in the popularity of politics. We found 3 days in which Politics peaked up to 7% of popularity in all locations (and up to 10%, one of the days, in the campus trace). When further investigating, we

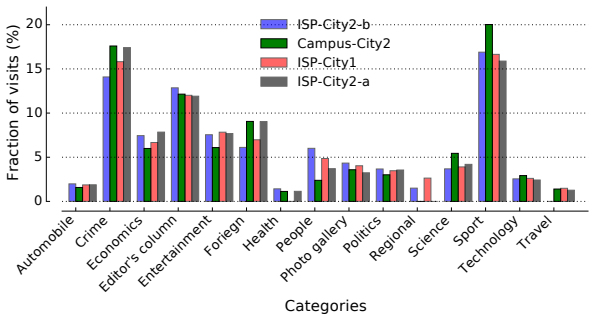


Figure 4: Most popular categories per location (% of visits).

found that these days preceded an Italian referendum in 2016, to which almost 16 million voters nation-wide participated. However, these small daily peaks did not have a remarkable influence on the monthly popularity of Politics.

Takeaway. *Although the exact popularity percentages of certain news categories slightly vary according to the time of day and regular or exceptional events of the year, they are overall stable in terms of ranking.*

7 Spatial stability

We now study the stability of category popularity across our locations and hence also across different demographics since the campus network contains mainly a young population of students.

Fig. 4 presents the top-10 categories for each location in terms of percentage of visits they attract. We see that news categories exhibit only slight differences across locations with few exceptions. *ISP-City2-a* and *Campus-City2* both have a slightly higher percentage of Tragedies/Weird. However, these two locations have, in contrast to the two others, a majority of active users during day time, a period of the day where Tragedies/Weird is often slightly more popular regardless of the location as we saw previously. *Campus-City2* users read more Sport and less People compared to the rest. Notice that in *Campus-City2*'s university the number of male students is much greater than the number of female ones because the university hosts mainly engineering courses. This fact can explain why Sport is more popular in *Campus-City2* than in other datasets. Finally, *ISP-City1* has a higher share of Regional news.

Takeaway. *Despite differences in geographical locations, category popularities seem to be rather stable across our four vantage points.*

8 Informal discussion

The comparison of users' self-reported preferences and actual usage data presented in Sec. 5 shows, for comparable categories, that there is a mismatch between preferred categories and popular ones. This difference is visible for both percentages of visits as well as the penetration in terms of users. Furthermore, the category popularity that we observe is not circumstantial. Indeed, with the exception of few events that slightly change the popularity of some categories, the popularity remains stable over time. Additionally, the popularity of categories is also rather stable across the various locations we study.

This important observation leads us to wonder why preferences expressed by users differ from usage data, i.e., users' actual behavior. Unfortunately, our dataset alone can not help answering this question. Hence, we build on prior work to informally discuss two possible causes, which we run by two experts. This section reports these informal discussions. Further research is required to evaluate which of these possible causes explains these differences.

The first possible cause is the so called *social desirability bias*, i.e., the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. Indeed, people might be "ashamed" of saying that they like "weird" news or news about violence or might have a better image of themselves if they show they are interested in more intellectual categories like Science and Politics.

The second possible cause is that this dissociation is a "legitimate" difference between what people *want*, and hence do (e.g., click on link), and what they actually *prefer* (e.g., an expression of interest in a given topic). This possible explanation is inspired by the incentive-sensitization theory of addiction (Berridge and Robinson 2016; 1998) in neuroscience. Individuals are driven by various rewards (e.g., food, sex, information or surprise). According to this theory, a reward has two components: *Like*, the pleasure one obtains from it, and *want*, i.e., how much one desires it, or how much effort one is ready to spend for it. Individuals may develop addiction to a given reward: The addictive behavior or substance hijacks the *want* system. The *want*, the motivational part of the reward becomes thus disconnected and higher than the *like*, the pleasure that the individual gets from the reward.[9] Assuming that users' expressed preference coincides with their perceived "pleasure", i.e., their *like*, then we conjecture that the dissociation we observe might be telling of the addictive nature of certain news categories. In other terms, it could be that people do not like (anymore), for instance sensational news, but ended up being "forced" to click on them in a cue-triggered way, like in addiction. However, this latter explanation could be too simplistic because what is subjectively perceived as preference is probably complex and cannot be reduced to the pleasure that individuals get from a "simple" reward like drugs. Other aspects like self-knowledge or self-image might enter also into play to guide people defining what is valuable. For instance, the need to maximize feelings of self-esteem can lead to the wish of reading noble or serious categories like Science and Politics.

To shed more light on this, we contacted Prof. Rasmus Kleis Nielsen, who is one of the lead authors of the Reuters Institute report (Newman, Levy, and Nielsen 2015). We asked for his feedback about our findings and whether he thinks that the mismatch we observe could be due to a social desirability

[9]The groundbreaking work of Berridge and Robinson helped scientists understand the difference between "like" and "want". In particular, their work redefined the role of mesolimbic dopamine, previously thought to be a pleasure neurotransmitter, and ultimately understood as a motivation or incentive salience mediator.

bias. He stated that he is not surprised by the difference between surveys and usage data results. Interestingly, he provided an interpretation of our finding that goes in the direction of the second possible cause, i.e., what people do must not necessarily follow what they prefer. Concerning the response bias, he replied that surveys are designed in such a way to minimize any kind of biases, so he would not support the first hypothesis. Indeed, prior to fielding surveys, a large effort is spent by Reuters and its partners all over the world to test for various issues, including social desirability bias. He also pointed us to work by Prior (2009), which demonstrates that self-reports immensely over-estimate news exposure.

We additionally contacted Prof. Kent C. Berridge, one of the two biopsychologists/neuroscientists who formulated the incentive-sensitization theory of addiction. We asked for his feedback about the second possible cause. Prof. Berridge said he thinks our second conjecture is correct: the dissociation between individuals' cognitive values, on one hand, and cue-triggered attraction, on the other, is probably what causes the mismatch we observe. As such, for him, the expressed news preferences reflect people's cognitive judgments about value (one of which is probably influenced by self-image) whereas their actions reflect their motivational-driven choices. Finally, he further added that this is "probably a little different" from what happens in the mesolimbic dopamine system in the brain that causes a dissociation between *want* and *like* for the same thing, as it happens when engaging with addictive behavior.

Finally, it is worth noting, that regardless of the reasons, the popularity of "Tragedies/Weird" news in our datasets seems in line with the history of news consumption where sensational news have been the most popular categories since ancient times (Stephens 2006).

9 Conclusion

In this paper, we analyzed 80 million news visits to 5.4 million news articles, extracted from a dataset of 17 month long anonymized HTTP traces. We focused on the popularity of news categories in this dataset and put it in perspective with the category popularity in four other sources: what is discovered on Facebook, what is observed by major news outlets, what is published, and finally user preferences expressed in surveys. The results of our analysis show that news consumed in social networks represent a small fraction of the overall news consumption and that none of the existing sources can faithfully capture news-category consumption at large. Interestingly, our results further demonstrate that the analysis of usage data complements survey-based data with new insights. Indeed, by putting survey-based results in perspective with usage data, we observed a clear mismatch between news categories which users claim to prefer and those they actually visit. We conjectured that this can be explained by a dissociation between individuals' cognitive values and their cue-triggered attraction.

References

Abbar, S.; An, J.; Kwak, H.; Messaoui, Y.; and Borge-Holthoefer, J. Consumers and suppliers: Attention asymmetries. a case study of aljazeera's news coverage and comments.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Ben Houidi, Z.; Scavo, G.; Ghamri-Doudane, S.; Finamore, A.; Traverso, S.; and Mellia, M. 2014. Gold mining in a river of internet content traffic. In *TMA*.

Berridge, K. C., and Robinson, T. E. 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain research reviews* 28(3):309–369.

Berridge, K. C., and Robinson, T. E. 2016. Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist* 71(8):670.

Boczkowski, P. J. 2010. The divergent online news preferences of journalists and readers. *Communications of the ACM* 53(11):24–25.

Chakraborty, A.; Ghosh, S.; Ganguly, N.; and Gummadi, K. P. 2016. Dissemination biases of social media channels: On the topical coverage of socially shared news. In *Tenth International AAAI Conference on Web and Social Media*.

Dezsö, Z.; Almaas, E.; Lukács, A.; Rácz, B.; Szakadát, I.; and Barabási, A.-L. 2006. Dynamics of information access on the web. *Physical Review E* 73(6):066132.

dos Rieis, J. C. S.; de Souza, F. B.; de Melo, P. O. S. V.; Prates, R. O.; Kwak, H.; and An, J. 2015. Breaking the news: First impressions matter on online news. In *Ninth International AAAI Conference on Web and Social Media*.

Falk, E. B.; Morelli, S. A.; Welborn, B. L.; Dambacher, K.; and Lieberman, M. D. 2013. Creating buzz: the neural correlates of effective message propagation. *Psychological Science* 24(7):1234–1242.

Finamore, A.; Mellia, M.; Meo, M.; Munafò, M. M.; and Rossi, D. 2011. Experiences of Internet traffic monitoring with Tstat. *IEEE Network*.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80(S1):298–320.

Kourogi, S.; Fujishiro, H.; Kimura, A.; and Nishikawa, H. 2015. Identifying attractive news headlines for social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1859–1862. ACM.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600. ACM.

Kwak, H.; An, J.; Salminen, J.; Jung, S.-G.; and Jansen, B. J. 2018. What we read, what we search: Media attention and public attention among 193 countries. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 893–902. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Leban, G.; Fortuna, B.; Brank, J.; and Grobelnik, M. 2014. Event registry: learning about world events from news. In

Proceedings of the 23rd International Conference on World Wide Web, 107–110. ACM.

Lee, A. M., and Chyi, H. I. 2014. When newsworthy is not noteworthy: Examining the value of news from the audience's perspective. *Journalism Studies* 15(6):807–820.

Leetaru, K., and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2. Citeseer.

Manson, M. online May 2016. Number of facebook users in italy. http://goo.gl/14gFl1.

Mitchell, A., and Page, D. 2014. State of the news media 2014. *Pew Research Journalism Project*.

Morgan, J. S.; Lampe, C.; and Shafiq, M. Z. 2013. Is news sharing on twitter ideologically biased? In *Proceedings of the 2013 conference on Computer supported cooperative work*, 887–896. ACM.

Newman, N.; Levy, D. A.; and Nielsen, R. K. 2015. Reuters institute digital news report 2015. *Available at SSRN 2619576*.

Olteanu, A.; Castillo, C.; Diakopoulos, N.; and Aberer, K. 2015. Comparing events coverage in online news and social media: The case of climate change. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, number EPFL-CONF-211214.

online April 2019. Published news urls and categorization scripts. https://smartdata.polito.it/published-news-urls-datasets-and-categorization-scripts/ .

Osborne, M., and Dredze, M. 2014. Facebook, twitter and google plus for breaking news: Is there a winner? In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.

Prior, M. 2009. The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opinion Quarterly* 73(1):130–143.

Ribeiro, F. N.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. P. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth International AAAI Conference on Web and Social Media*.

Saez-Trumper, D.; Castillo, C.; and Lalmas, M. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 1679–1684. ACM.

Statista. online May 2016. Number of twitter users in italy. http://goo.gl/liXla2.

Stephens, M. 2006. *A History of News*. Bloomington, IN, USA: Oxfort University Press.

Trammell, B.; Casas, P.; Rossi, D.; Bar, A.; Houidi, Z.; Leontiadis, I.; Szemethy, T.; and Mellia, M. 2014. mplane: an intelligent measurement plane for the internet. *IEEE Communications Magazine* 52(5):148–156.

Vassio, L.; Drago, I.; Mellia, M.; Houidi, Z. B.; and Lamali, M. L. 2018. You, the web, and your device: Longitudinal characterization of browsing habits. *ACM Trans. Web* 12(4):24:1–24:30.

Wang, Y., and Mark, G. 2013. Trust in online news: comparing social media and official media use by chinese citizens. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 599–610. ACM.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*. Springer. 338–349.