What we learned about *The Gateway Pundit* from its own web traffic data

Zhouhan Chen[1*], Haohan Chen[2], Juliana Freire[1], Jonathan Nagler[1], Joshua A. Tucker[1],
[1]New York University
[2]The University of Hong Kong
[*]Corresponding author. E-mail: zhouhan.chen@nyu.edu

Abstract

To mitigate the spread of false news, researchers need to understand who visits low-quality news sites, what brings people to those sites, and what content they prefer to consume. Due to challenges in observing most direct website traffic, existing research primarily relies on alternative data sources, such as engagement signals from social media posts. However, such signals are at best only proxies for actual website visits. During an audit of far-right news websites, we discovered that *The Gateway Pundit* (*TGP*) has made its web traffic data publicly available, giving us a rare opportunity to understand what news pages people actually visit. We collected 68 million web traffic visits to the site over a one-month period and analyzed how people consume news via multiple features. Our referral analysis shows that search engines and social media platforms are the main drivers of traffic; our geolocation analysis reveals that *TGP* is more popular in counties where more people voted for Trump in 2020. In terms of content, topics related to 2020 US presidential election and 2021 US capital riot have the highest average number of visits. We also use these data to quantify to what degree social media engagement signals correlate with actual web visit counts. To do so, we collect Facebook and Twitter posts with URLs from *TGP* during the same time period. We show that all engagement signals positively correlate with web visit counts, but with varying correlation strengths. For example, total interaction on Facebook correlates better than Twitter retweet count. Our insights can also help researchers choose the right metrics when they measure the impact of news URLs on social media.

Introduction

Misinformation, or false news, is a major threat on today's Internet (Tucker and Persily 2020; Grinberg et al. 2019; Vosoughi, Roy, and Aral 2018). However, it remains challenging to measure the consumption of false news URLs. Since there is no single metric to quantify the spread of information, the choice of metrics can affect downstream analysis and alter final conclusions. There are two approaches to measuring false news consumption: indirect and direct.

For indirect measurement, a common method is to collect social media posts containing the URL of interest, calculate engagement signals, and use those metrics as a proxy for URL popularity (Center for an Informed Public et al. 2021; Guess, Nagler, and Tucker 2019; Guess et al. 2021). Indirect measurements reveal how people *share* news URLs, but not how people actually *visit* those URLs (Sacher and Yun 2017).

The alternative approach is to collect data directly related to URL visit statistics. Only a few studies were able to gather direct measurement data. For example, Chalkiadakis et al. (2021) collect visit data to fake news sites from third party services such as SimilarWeb and CheckPageRank to assess user engagement. In another study, Fourney et al. (2017) gather browsing data from Microsoft Internet Explorer and Edge to analyze visiting patterns to fake news sites before the 2016 US Election. To the best of our knowledge, no one has previously explored web traffic data collected on the server side in academic research related to misinformation. Web traffic data has rich features that alternative sources of data do not possess. The challenging part, though, is that even when websites record their traffic, few make the data publicly available.

During an audit of popular far-right news websites, we discovered that *TGP* makes its website traffic available to the general public. *TGP* is a right-wing news site in the United States. The website is crucial for the study of misinformation due to several reasons. First, *TGP* has an increasing user base. The website has the second largest percentage of traffic surge among right-wing news sites from December 2019 to December 2020 (Majid 2021). Second, *TGP* publishes a large amount of misinformation (Harling 2021; Faris et al. 2017). According to Center for an Informed Public et al. (2021), it is "one of the top-three most cited domains in tweets spreading false and misleading narratives about voter fraud [in the United States] in 2020." Third, *TGP* is highly influential. For example, *TGP* articles were cited by former US President Trump's lawyer and referenced in Trump's Impeachment Defense Memo.[1] All of these reasons make *TGP* an ideal case study to understand online fake news consumption behavior.

Given this opportunity, we collect the entire web traffic to *TGP* for one month from February 4, 2021 to March 3, 2021. We collect a total of 68 million website visits.

[1]https://int.nyt.com/data/documenttools/trump-defense-impeachment-trial/3a17fbb266bf3bf5/full.pdf

Our analysis is two-fold: we first explore available features within the web traffic data to understand how people consume news from *TGP* ; we then collect additional social media posts to quantify correlations between social media engagement signals and actual web visit counts. Our substantive findings include:

1. Search engines such as Google, Duckduckgo and Bing account for 88.5% of external referral traffic to *TGP* home page. However, social media platforms including Twitter, Facebook, Telegram and Gab account for more than 42% of external referral traffic to article pages at *TGP* .
2. Geo-location modeling shows that *TGP* is more popular in counties that voted for Trump in 2020 US election. Topic modeling reveals that articles that mention "2020 US election fraud" are visited by 29% more users compared with articles from other topics.
3. Social media engagement signals positively correlate with actual website visit counts. Not all metrics are the same: Facebook metrics achieve a stronger correlation than Twitter metrics.

We hope those findings can inform researchers to select social media metrics judiciously even if they do not have access to server-side traffic. We also hope to motivate researchers to look for data sets beyond social media platforms. Lastly, based on our analysis, we want to work with the industry partners to design robust intervention strategies to mitigate the spread of fake news URLs. For example, social media companies can add friction when they redirect traffic to far-right sites; search engines can promote articles that have been fact-checked, and to down-rank articles that come from historically low-quality sources such as *TGP*.

Method

In this section, we first explain how we collect web traffic data from *TGP* for one month. We then give an overview of the collected data, and address issues related to data integrity, missing data, and data privacy.

Data collection

During an audit of *TGP*, we discover that the website openly collects and publishes its visitor traffic via StatCounter, a web traffic collection service. There is a button on the home page of *TGP* that leads every visitor to a dashboard with detailed traffic data. Additionally, any user can download real-time visitor traffic by sending an HTTP GET request to a URL endpoint, which we refer to as the *download URL*. This URL includes two important parameters, which we call *StartTime* and *EndTime*.[2] Table 1 shows features associated with each visit.

[2]The stat is available at https://statcounter.com/p9449268/summary/?guest=1. The *download URL* follows the following pattern: https://statcounter.com/p9449268/csv/download_log_file?range=StartTime--EndTime. *StartTime* and *EndTime* must be in ISO format, such as *2021-04-12T02:18:41*. For a formal definition of ISO format, refer to https://www.w3.org/TR/NOTE-datetime.

feature	description	example
datetime	time of the visit	2021-02-02 18:23:31
ip	IPv4 adddress	10.11.123.12
os	operating system	IOS
url	url visited	thegatewaypundit.com
isp	ISP	Verizon
country	country of IP	USA
city	city of IP	Houston
region	state of IP	Texas
referrer	previous url	google.com
page title	title of the article	Expert claims...
browser	browser name	Safari
resolution	device resolution	375×667

Table 1: Features in *TGP* web traffic data set.

During our testing phase, we find that no matter what *StartTime* and *EndTime* we set, the downloaded CSV file always contains traffic captured during the most recent 20 minutes. To collect website traffic continuously, we automate a Chrome Browser to visit the *download URL* every 15 minutes, from February 3, 2021 to March 3, 2021. We choose a 15 minute interval because it is below the 20-minute interval with a safe margin. One side effect is that our data has duplicates. To remove duplicates, we identify that each website visit is uniquely defined by the combination of five features: *datetime, url, ip, os*, and *browser*. Therefore, we only keep the first record if multiple records have the same five-feature combination.

Data integrity

To validate that our collection method captures the entire traffic, we compare the daily number of visits reported by StatCounter against the number calculated from our collection after de-duplication. Our data set has a completeness ratio of more than 99.8% on a daily basis. We define the completeness ratio as our number of visits divided by StatCounter's number of visits. The lost entries are possibly caused by parsing errors or corrupted network packages. We believe that this small number of missing entries (less than 0.2%) will not affect trends we observe.

Missing data and bot traffic

Even though we capture the entire web traffic, our data source (StatCounter) has several inherent limitations. One potential issue is under-counting. For example, anyone who blocks HTTP and HTTPS requests to StatCounter will not have their visits logged by the server. This can happen if people install certain anti-tracking plug-ins. Unfortunately, it is impossible to know exactly how many users install anti-tracking tools, as those tools are designed to hide web visit history.

Another problem is the presence of bot traffic. Bots are programs that automatically visit web pages. According to its documentation, StatCounter does not record most bots or crawlers, because clients have to actually load javascript for their hit to be logged in the system. Sophisticated bots that

emulate human behavior can still bypass the detection. Even though the amount of missing data and advanced bot traffic is undetectable, we believe that those irregularities will not affect the overall trends we report from our analyses.

Data privacy

To address concerns regarding data privacy, we first note that the *TGP* web traffic data is publicly available, and does not contain personally identifiable information such as name, phone number, cookie, session ID, device ID, or email address. For geo-location, we are only provided with the city that the incoming IP address belongs to. Additionally, we aggregate our results and do not report any individual traffic.

Insights from 68 million web visits

In this section, we take a multi-pronged approach to analyze our data from one-month of web visits. We first give an overview of the data set. We then analyze referrer links to understand which sites bring users to *TGP*. Next we leverage geo-location information to validate if people who visit *TGP* come from areas that voted more favorably for Donald Trump in the 2020 US Presidential Election. Finally, we cluster all articles into groups of topics to quantify what types of stories are more likely to go viral.

Overview of our data set

Our data collection contains 68,268,818 unique visits, from February 3, 2021 to March 3, 2021. Figure 1 plots the number of visits per hour. Since more than 95% of the visits come from the United States, we see a regular and circadian pattern where the traffic increases during the day, and decreases during the night. The daily peak hourly visit is around 200,000. The only exception is one hour in February 13, 2021, with a recorded visit of nearly 300,000. February 13, 2021 is the day Donald Trump was acquitted on impeachment charges, a highly publicized political event in the United States. The two most visited articles published that day are both about impeachment acquittal. Finally, we find that more than 80% of visits come from mobile devices such as IPhone and Android. We do not find differences in the distribution of pages viewed by mobile versus desktop devices.

Finding 1: Search engines and social media sites are the main drivers of traffic to *TGP*

Knowing what websites bring people to *TGP* helps us to identify the source of traffic and to design intervention strategies to slow down the spread of low quality or false news. To reconstruct traffic flows, we use the *referrer* column in our web traffic data. When a browser navigates to URL B from URL A, it usually includes a string called referrer in the HTTP request (in our example A is the referrer of B). Among 68,268,818 visits, 35,296,042 (52%) have referrers. For visits that do not have referrers, either users visit a URL directly, or the browser strips the referrer, which can happen when certain privacy-enhancing features are turned on. To aggregate referrers that belong to the same site, we normalize each referrer URL to its domain name,

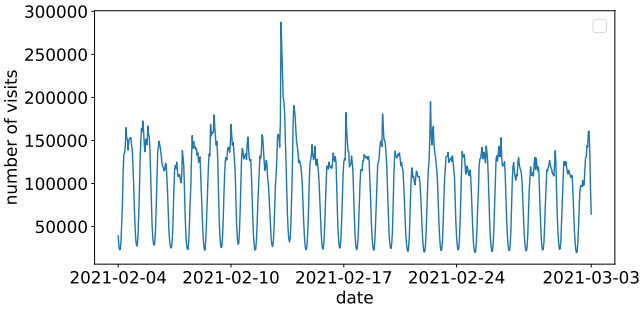

Figure 1: Number of visits per hour, from February 4, 2021 to March 3, 2021. There is a peak on February 13, 2021, the day Donald Trump was acquitted on impeachment charges. Our data shows that the two most visited articles during that day both covered this event.

removing hostname, path, and other query parameters. We consider two referral behaviors based on the destination URL: sites that bring users to the home page, and sites that bring users to an article page. A **home page** URL points to domain *thegatewaypundit.com*, while an **article page** URL has the form *thegatewaypundit.com/ARTICLE-NAME*. Each type of referral behavior has its own characteristics, which we analyze separately.

Websites that bring users to the home page. Figure 2 shows the top 20 domains that bring visitors to *TGP* home page. The most significant traffic driver is search engine sites, which account for 88.5% of external referral traffic. Among them, *Google.com* is the top driver of home page traffic (66%). The anonymous search engine *duckduckgo.com* is the fourth (13%), and the Microsoft-developed *bing.com* ranked the fifth (9%).

The second most popular referrer to the home page is other *TGP* article page. This shows when people browse articles on *TGP*, they usually navigate back to the home page from different article pages. The third referrer is the *TGP* home page. This is likely caused by people clicking links to the home page when they are already at the home page. Further down the list are far-right and conservative news sites such as *drudgereport.com*, *63red.com* and *protrumpnews.com*.

We also identify referrers from suspected phishing domains. One such domain is netlix.com, ranked number ten. The domain name was previously at the center of a lawsuit. According to a legal complaint filed by Netflix in 2009, the video streaming company claimed that the domain name "netlix" looked too similar to "netflix," and requested netlix.com to be transferred to Netflix.[3] The court rejected the order, and netlix.com still belongs to its original owner. The website originally redirected users to *TGP*. As of February 15, 2022, the website redirects users to mydailychoice.com.

Websites that bring users to an article page. Figure 3 shows the top 20 domains that bring visitors to an article page. The top two referrers – the *TGP* home page and *TGP*

[3]https://www.adrforum.com/domaindecisions/1287043.htm

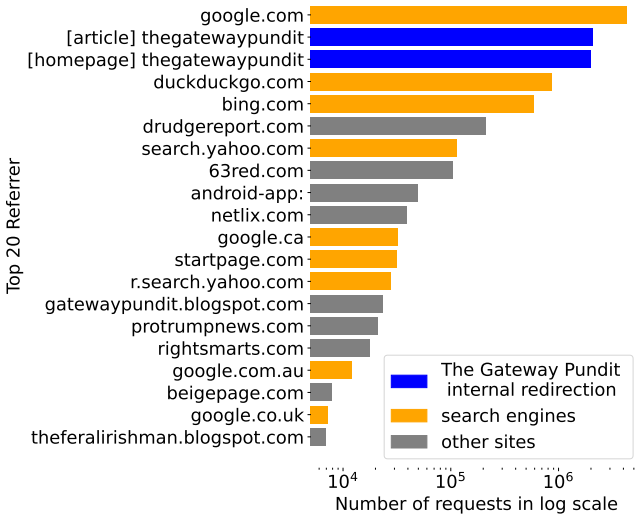

Figure 2: Top 20 domains that bring users to *TGP* home page. 88.5% of external referral traffic comes from search engines such as Google, Duckduckgo, Bing and Yahoo.

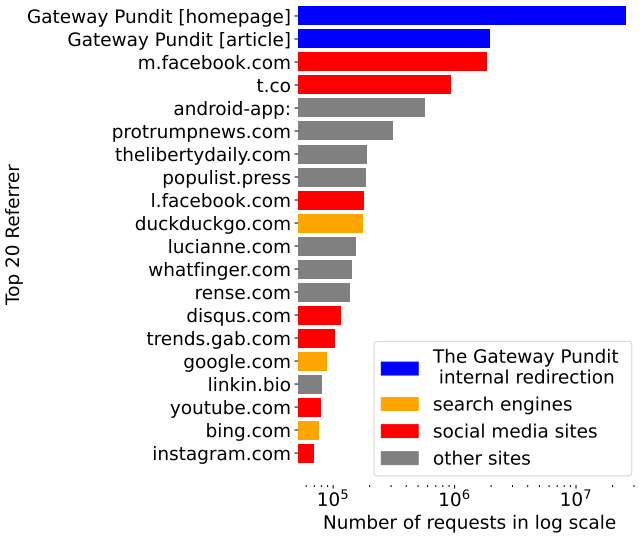

Figure 3: Top 20 domains that bring users to an article page. When we exclude internal referrers, social media platforms such as Twitter, Facebook, Telegram, Gab bring 42% of external referral traffic.

article pages – are both internal traffic. This indicates that (a) most users first land on the home page before clicking on an individual article, and (b) some users click a new article page while browsing an existing article page, since different articles are link to each other. After we exclude internal traffic, we identify three groups of websites:

1. **Social media platforms** including Twitter, Facebook, and emerging platforms such as Telegram and Gab. Together they account for 42% of external referral traffic.

2. **Conservative news sites** such as *protrumpnews.com*, *thelibertydaily.com*, *populist.press* and *whatfinger.com*. Those sites repost articles from *TGP* on a regular basis.

3. **Search engines** such as Google and Duckduckgo.

To further understand how much role each social media platform plays in driving the traffic, we plot the daily number of visits with referrers from four different social media platforms, shown in Figure 4. The overall trend shows that Twitter and Facebook drive more traffic than Telegram and Gab, which is probably not all that surprising given how many more users the first two platforms have in the United States compared to the latter two. Nevertheless, this finding highlights the fact that not all of the traffic to *TGP* is coming from the right-wing online ecosystem. Daily traffic volume fluctuates and can be affected by external events. For example, Jim Hoft, founder of *TGP*, was suspended by Twitter on February 6, 2021.[4]. The suspension might be related to the decline of traffic from Twitter on that day

[4]https://www.forbes.com/sites/ajdellinger/2021/02/06/twitter-suspends-gateway-pundit-jim-hoft/?sh=761c0cff3653

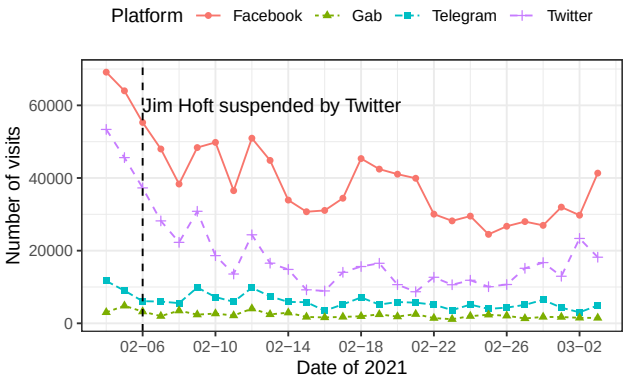

Figure 4: Daily referrers by platform. Facebook and Twitter are the top two traffic drivers. On February 6, 2021, Jim Hoft, founder of *TGP*, was suspended by Twitter. This event might be related to the decline of Twitter traffic in early February.

Finding 2: Visitors to the site are more likely to be from areas that voted for Donald Trump during the 2020 US Presidential Election

Our web traffic data records IP and city-level geo-location labels for every request. To better understand the audiences' political preferences, we leverage the geo-location information to answer the following question: given the fact that *TGP* is a right-leaning publication, is it more popular in counties where more people voted for Trump?

To answer this question, we fit a regression model to estimate how changes in percentage of Trump voters per county (x) affect changes in percentage of *TGP* visitors per county

(y). We assume that each unique IP address is one unique visitor. In reality, our assumption might not always be true. For example, multiple people in a household can share the same IP, or one person can visit the site from multiple IP addresses. Even though those limitations exist, IP address is the most accurate proxy to real human traffic in our data set. IP is also commonly used in security research to generate threat intelligence from traffic logs (Fourney et al. 2017).

We then collect county-level 2020 US Presidential Election results, including the total number of voters and number of voters who voted for Trump. Finally we aggregate city-level count into county-level count. We fit the following linear regression model:

$$y = ax + b, \text{where for each county } i\text{:}$$

$$y_i = \frac{\text{unique number of visitors from county } i}{\text{total number of voters from county } i}$$

$$x_i = \frac{\text{\# voters who voted for Trump from county } i}{\text{total number of voters from county } i}$$

Figure 5 shows the scatter plot of x and y, the fitted regression line and confidence intervals. The coefficient a is 0.037 and the intercept b is 0.058. The r-squared value is 0.17. The coefficient indicates a positive correlation between x and y, supporting our initial hypothesis that *TGP* is more popular in counties where more people voted for Trump. Unfortunately, we do not know the casual relationship from our observational data. Do people read news from *TGP* because they support Trump, or do they support Trump because they read news from *TGP*? Future research can help investigate those questions.

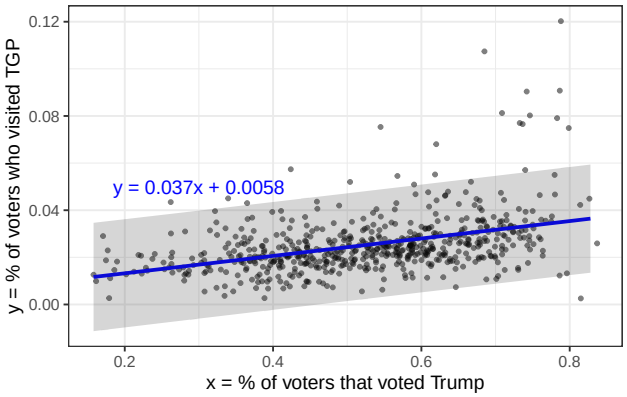

Figure 5: Scatter plot of county-level % voters who voted for Trump (x-axis) in the 2020 US Election versus % voters who visited *TGP* (y-axis). The red line is expected value of y given x. The yellow and green lines are confidence intervals. The regression model has a coefficient of 0.037 and an intercept of 0.058, suggesting a positive correlation between x and y.

Finding 3: Topics related to "election fraud" and "capital riot" receive more visits than others

During the one-month period of our study, *TGP* published 1070 articles. Some stories go viral, others do not. What topics are discussed? What makes one topic goes viral? To better understand those connections, we use topic clustering technique to group *TGP* articles into distinct topics. We then analyze topic distribution to identify viral topics.

How do we extract topics? Each article published on *TGP* comes with a one-sentence title with references to key names and events. For example, one article published on February 18, 2021 is titled "*Maricopa County Audits Are Proving to Be a Waste of Time and Money, They Were Never Created to Identify the Suspected Election Fraud in the County*." Given the rich information from the title, we use non-negative matrix factorization (NMF) to cluster 1070 titles into different topics. Previous studies have used NMF to discover meaningful topics from short-length corpus such as tweets (Tanash et al. 2015). In our case, the input to NMF is a title-word matrix, where each entry is the tf-idf weight of a word in a title. NMF factorizes this matrix into a word-topic matrix, and a topic-title matrix.[5] The number of topic is a user-defined parameter. After experimenting with different values, we set the parameter at 10, as the resulting topics are coherent and distinct from each other. Figure 6 shows keywords associated with each topic. Conspiratorial topics such as voting fraud are very popular (Topic 3,8,10). Other topics are impeachment (Topic 1, 7), capital riot (Topic 5), COVID-19 vaccine (Topic 9) and US politics (Topic 2, 4, 6).

What topics receive more visits? For each article, we count the number of unique visits based on number of unique IP addresses. We then group article-level counts into topic-level. Figure 6 shows for each topic, the distribution of article-level number of visits. If we sort topics by their median number of visits, three out of the top four most visited topics relate to the "2020 US election fraud," an unverified claim pushed by far-right news media. The second most visited topic mentions "capital riot," "police fbi" and "antifa" – all related to the United States Capitol attack on January 6, 2021. The popularity of those topics indicate that readers of *TGP* had a huge appetite for articles about politics and election.

Comparing web traffic data with social media engagement signals

As noted previously, existing research on news consumption mostly focuses on how news URLs are shared on social media platforms, especially Twitter and Facebook. While social media signals can tell us how people *share* news, they do not answer how many people actually *visit* each URL. Fortunately, we can leverage our data to assess the relationship between sharing and viewing of social media, and therefore shed light on whether the former is actually a useful proxy for the latter. More specifically, we can examine whether

[5]We use the Python sklearn package to implement topic classification. Reference: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

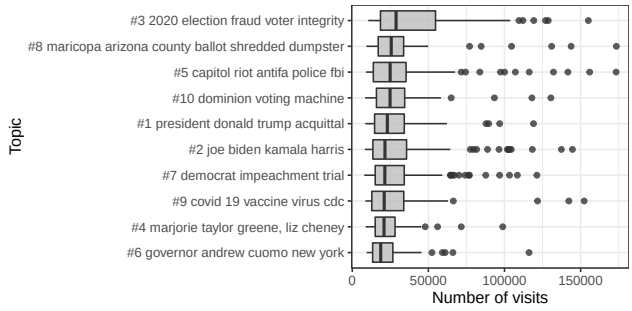

Figure 6: Distribution (box plot) of article-level number of visits, separated by topic. In terms of the median value, the most visited topics are related to the 2020 US election fraud (#3, #8, #10) and the 2021 US Capital riot (#5).

there is any correlation between social media sharing behavior and actual news consumption behavior for *TGP* articles, and, if so, assess how strong the correlation is. To answer those questions, we first collect Facebook and Twitter metrics to measure popularity of *TGP* links shared on each platform. We then test the correlational strength of different metrics against website visit count to identify good traffic estimators.

Collecting posts from Facebook and Twitter

Among 1070 online articles published by *TGP* during our one-month data collection, 1020 received more than 10,000 unique web visits. To ensure the stability of our experiment, we focus on those 1020 URLs and discard URLs with lower web visit counts. We then use the Crowdtangle API to collect Facebook posts that contain any one of the 1020 URLs published by *TGP*. Crowdtangle is a data intelligence service that tracks aggregated engagements and interactions of posts from Facebook pages and groups (both public and private). We use Twitter Academic API to collect all original and public tweets that contain any one of the 1020 URLs.[6] For each URL, we calculate the seven metrics shown in Table 2. Our comparison is not exhaustive – there are many metrics associated with a social media post. The metrics we choose in this section are the most often used by researchers.

metric	source
# unique visits (all referrer)	web traffic dataset
# unique visits (from facebook.com)	web traffic dataset
# unique visits (from twitter.com)	web traffic dataset
total number of FB reactions	Crowdtangle API
total number of FB interactions	Crowdtangle API
total number of likes	Twitter API
total number of retweets	Twitter API

Table 2: We calculate seven metrics to quantify the popularity of an article URL.

[6]CrowdTangle API: https://help.crowdtangle.com/en/articles/1189612-crowdtangle-api; Twitter Academic Research API: https://developer.twitter.com/en/products/twitter-api/academic-research

Measuring correlations

We calculate Pearson correlations between the *log* of each social media metric and (a) the *log* number of visits from any referrer and (b) the *log* number of visits from platform-specific referrer. Pearson correlation is used to summarize the strength of the linear relationship between two variables (Freedman, Pisani, and Purves 2007). Because both social media metrics and website visit counts follow a logarithmic distribution, where only a small number of URLs receive very large amount of traffic, we take log of all variables. We show all pairwise correlations in Table 3. More detailed scatter plots are available in the Appendix.

We first observe that Facebook metrics correlate better with traffic that only originated from Facebook than traffic that originated from all sites. The same is true for Twitter metrics. For example, Table 3 shows that the Pearson correlation between total Facebook interaction and the number of visits from facebook.com is 0.939, while the correlation is only 0.595 for the number of visits from all referrers. This decrease is due to the fact that each social media platform can only capture URL sharing activities within its platform.

We also observe that Facebook metrics correlate better with web visit counts than Twitter metrics. For example, when we focus on the first column of Table 3, we see that Facebook interactions have a higher correlation with web visit counts than Twitter retweet counts. The former metric has a Pearson correlation of 0.595 while the latter has a correlation of 0.465. Why is there a discrepancy? One reason may simply be that significantly more Americans use Facebook than Twitter, and thus the former provides a better snapshot of the overall online population than the latter. Understanding what other factors affect the correlation is a subject for future research.

	# unique visits from all referrers	# unique visits from only FB or Twitter
FB reaction	0.511	0.857
FB interaction	0.595	0.939
Twitter retweet	0.465	0.574
Twitter like	0.435	0.542

Table 3: Pearson correlations between social media metrics and website visit counts.

Discussion and limitations

How do our findings about *TGP* inform future research that won't have access to such detailed traffic data? First, we show that there are caveats when using social media metrics as a proxy for URL popularity. Different platforms and metrics yield different correlation strengths. To fully understand the scope and intent of a news URL, researchers should collect data from multiple platforms and report multiple metrics if possible.

The second insight is that social media companies should be more transparent about outgoing traffic from the platform. Retweet count or interaction count is not the same as actual web page visits. Our analysis shows that social network platforms and search engines play an important role

in bringing users to *TGP*. Unfortunately, neither the Twitter Academic API nor the Crowdtangle API provides URL click statistics. As a result, researchers without access to server logs usually have no idea how often social media platforms actually redirect traffic to an external website. While it is true that some false information is consumed within the platform by simply reading a headline and blurb, a much fuller picture of the role played by social media platforms in exposing users to false information would be greatly enhanced by information about whether people left the platform for the full article.

In terms of future directions, we plan to collaborate with industry partners that have direct access to web traffic data, such as web tracking companies or hosted service providers. One challenge is to ensure data privacy. Researchers can explore techniques such as differential privacy or federated machine learning to overcome the data sharing difficulty.

One major limitation of our research is our inability to analyze other comparable web traffic data sets. We reached out to several media organizations to request similar data, but in no case were we successful in receiving access to similar data. As a result, our work here focuses solely on *TGP*. Accordingly, our conclusions should not be overgeneralized and might not apply to other news outlets. Moreover, to date we have been unable to benchmark our findings here against a more mainstream news outlet such as *The New York Times* or *The Wall Street Journal*.

Another limitation is that our paper does not look into the time dynamics within the period – most of the analysis is done by aggregating values in the entire one-month period. Future research would benefit from analyzing how visits to low quality news source change over time. Do they, for example, increase during election campaigns? Do they go down during periods of international crises like the Russian invasion of Ukraine? While one way to tap into time trends is to slice our current data set into smaller pieces, another would be to keep collecting new data sets from *TGP*. With more data sets that span multiple periods, we will be able to observe trends that are dependent on time.

Related Work

Measuring how people consume misleading or fake news is an important but challenging research area. Previous work mostly studies the spread of fake news on social media platforms (Center for an Informed Public et al. 2021). For example, Vosoughi, Roy, and Aral (2018) collect tweets containing links to fake news sites, and concludes that fake news spread faster and further than traditional news. In another study using Twitter data, Grinberg et al. (2019) claim that "fake news accounted for nearly 6% of all news consumption, but it was heavily concentrated on a small percentage of users." Similarly, Guess, Nagler, and Tucker (2019), Guess et al. (2021) collect Facebook posts to understand news consumption behavior, and (Shao et al. 2018) deployed a system called Hoaxy to analyze the diffusion of articles from low-credibility sources on Twitter.

While social media engagement signals can tell us how people share news on different platforms, they do not necessarily translate into web traffic to the news site (Sacher and Yun 2017). One way to bridge this gap is to directly gather data from volunteers via browsing extensions. For example, Ognyanova et al. (2020) ask participants to install a browser extension to measure their exposure to fake news. However, this approach is usually expensive and the sample size is small.

To understand population-level news consumption behavior, there is an urgent need to collect "unique datasets with increased validity (Pasquetto et al. 2021)." Web traffic data is a direct measurement of news consumption. In one study, Chalkiadakis et al. (2021) assesses user engagement by collecting traffic data from tracking services such as SimilarWeb and CheckPageRank. Fourney et al. (2017) gather browsing data from Microsoft Internet Explorer and Edge, and analyze visitor patterns to a list of fake news domains before the 2016 US Election.

Different from all previous approaches, we focus on collecting the entire web traffic to a single but important news site (*TGP*). Our data set enables us to validate and extend previous traffic-based analysis. As far as we know, we are the first to test correlations between social media engagement signals and web traffic counts, by combining Twitter and Facebook posts with web traffic data.

References

Center for an Informed Public; Digital Forensic Research Lab; Graphika; and Stanford Internet Observatory. 2021. *The Long Fuse: Misinformation and the 2020 Election*. URL https://purl.stanford.edu/tr171zs0069.

Chalkiadakis, M.; Kornilakis, A.; Papadopoulos, P.; Markatos, E. P.; and Kourtellis, N. 2021. The Rise and Fall of Fake News sites: A Traffic Analysis. *CoRR* abs/2103.09258. URL https://arxiv.org/abs/2103.09258.

Faris, R.; Roberts, H.; Etling, B.; Bourassa, N.; Zuckerman, E.; and Benkler, Y. 2017. Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. *Scientific Reports* URL https://ssrn.com/abstract=3019414.

Fourney, A.; Racz, M. Z.; Ranade, G.; Mobius, M.; and Horvitz, E. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, 2071–2074. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3132847.3133147. URL https://doi.org/10.1145/3132847.3133147.

Freedman, D.; Pisani, R.; and Purves, R. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* .

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363(6425): 374–378. doi:10.1126/science.aau2706. URL https://science.sciencemag.org/content/363/6425/374.

Guess, A.; Aslett, K.; Tucker, J.; Bonneau, R.; and Nagler, J. 2021. Cracking Open the News Feed: Exploring What U.S.

Facebook Users See and Share with Large-Scale Platform Data. *Journal of Quantitative Description: Digital Media* 1. doi:10.51685/jqd.2021.006. URL https://journalqd.org/article/view/2586.

Guess, A.; Nagler, J.; and Tucker, J. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5(1). doi:10.1126/sciadv.aau4586. URL https://advances.sciencemag.org/content/5/1/eaau4586.

Harling, A.-S. 2021. The Gateway Pundit NewsGuard Nutrition Label. https://www.newsguardtech.com/wp-content/uploads/2020/02/The-Gateway-Pundit-NewsGuard-Nutrition-Label.pdf. [Online; accessed 12-June-2021].

Majid, A. 2021. Top 50 largest news websites in the world: Surge in traffic to Epoch Times and other right-wing sites. https://www.pressgazette.co.uk/top-50-largest-news-websites-in-the-world-right-wing-outlets-see-biggest-growth/. [Online; accessed 14-June-2021].

Ognyanova, K.; Lazer, D.; Robertson, R. E.; and Wilson, C. 2020. *Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power*. URL https://doi.org/10.37016/mr-2020-024.

Pasquetto, I. V.; Swire-Thompson, B.; Amazeen, M. A.; et al. 2021. *Tackling misinformation: What researchers could do with social media data*. URL https://doi.org/10.37016/mr-2020-49.

Sacher, S. B.; and Yun, J. M. 2017. Fake News is Not an Antitrust Problem. *CPI Antitrust Chronicle* URL https://ssrn.com/abstract=3090649.

Shao, C.; Hui, P.-M.; Wang, L.; Jiang, X.; Flammini, A.; Menczer, F.; and Ciampaglia, G. L. 2018. Anatomy of an online misinformation network. *PLOS ONE* 13(4): 1–23. doi:10.1371/journal.pone.0196087. URL https://doi.org/10.1371/journal.pone.0196087.

Tanash, R. S.; Chen, Z.; Thakur, T.; Wallach, D. S.; and Subramanian, D. 2015. Known Unknowns: An Analysis of Twitter Censorship in Turkey. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, WPES '15, 11–20. New York, NY, USA: Association for Computing Machinery. doi:10.1145/2808138.2808147.

Tucker, J. A.; and Persily, N. 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*. SSRC Anxieties of Democracy. Cambridge University Press. doi:10.1017/9781108890960.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380): 1146–1151. doi:10.1126/science.aap9559. URL https://science.sciencemag.org/content/359/6380/1146.

Appendix

Correlations between social media metrics and website visit counts

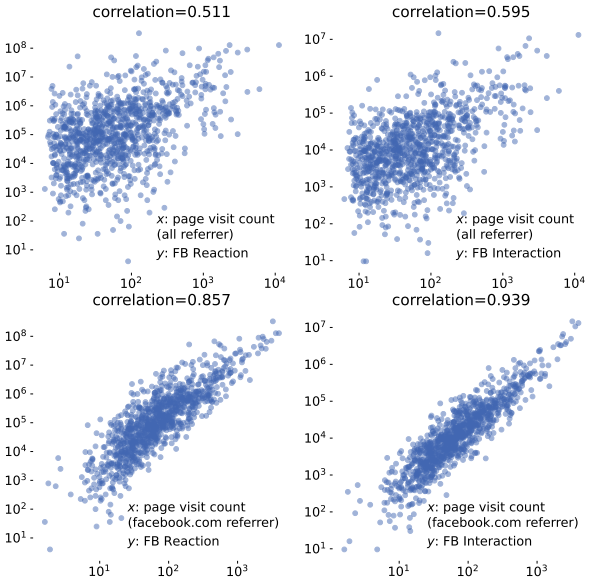

Figure 7: Four scatter plots visualizing correlations between Facebook engagement metrics (left column: reaction, right column: interaction) and page visit counts (top row: all referrer, bottom row: only *facebook.com*). In each scatter plot, a dot represents one *TGP* URL. Both axes are in log scale.

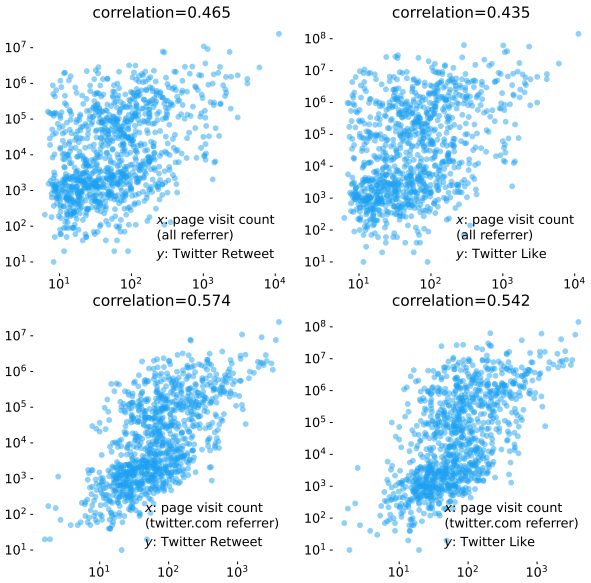

Figure 8: Four scatter plots visualizing correlations between Twitter engagement metrics (left column: retweet count, right column: like count) and page visit counts (top row: all referrer, bottom row: only *twitter.com*). In each scatter plot, a dot represents one *TGP* URL. Both axes are in log scale.