Proceedings of the 2022 Truth and Trust Online Conference (TTO 2022)

October 12-14, Boston, MA, USA and Online

Production and Manufacturing by
Hacks Hackers
c/o Fletcher Heald and Hildreth PLC
1300 17th St. North, 11th floor
Arlington, VA 22209 United States

Order copies of this TTO proceedings from:

Hacks Hackers
c/o Fletcher Heald and Hildreth PLC
1300 17th St. North, 11th floor
Arlington, VA 22209 United States
`admin@truthandtrustonline.com`

Preface

After two years of being fully online, due to the COVID-19 pandemic, the fourth annual Conference for Truth and Trust Online (TTO), took place in a hybrid mode: in person, at the Boston University, in Boston (Massachusetts), and online, on October 12-14, 2022. This was the first time that TTO experimented with such a model, which presented a new set of challenges for organizers accustomed to either fully in person or fully online events.

TTO's mission is clear in its name: making progress toward an **online** world that is truthful and trustworthy. To realize this mission, TTO provides a unique forum for individuals from academia, industry, non-profit organizations, community organizers, and other stakeholders to come together to discuss the problems surrounding participation in online platforms with respect to truth and trust, together with socio-technical approaches for understanding and addressing them.

All these stakeholders bring different perspectives, which taken altogether, contribute to deepening our collective knowledge on truth and trust online. People in the frontlines, such as NGO staff, community organizers, and public servants share the problems they encounter while navigating the online ecosystem and discuss the impact of policies by governments and platforms, intended to increase online safety and health. Academics introduce new algorithms for automating fact-checking or new insights into how certain phenomena surrounding truth and trust occur in the wild. Industry representatives discuss the challenges of scaling automated solutions or thwarting adversarial attacks on the platforms. Listening to one another enhances how each group can approach their daily tasks.

To facilitate this multifaceted communication, the conference provides different kinds of sessions: new research published for the first time at TTO, invited panels with experts from different paths, demo sessions for new tools, and talks that summarize lessons learned. In putting together the program for the conference, we also strived to build in time for in-person interactions that will lead to new collaborations. One such session happened in the morning of October 13, in a workshop led by Dr. Jo Lukito (UT Austin), that brought together people interested in global mis/dis-information and approaches to address it in geographical areas that are underrepresented in our research community, also known as the Global South.

The Call for Contributions to the conference, as in the past three iterations, invited contributions of two kinds: (1) technical papers that have not been published elsewhere; and (2) talk proposals that either summarize already published research for the purposes of further dissemination or that introduce new ideas for the community to engage with.

We received 22 submissions for the technical track and 25 submissions for the talk proposals track. Each paper in the technical track received three (mutually anonymous) reviews from our program committee members and each talk proposal received two (and occasionally three) reviews. Five technical papers were accepted for inclusion in this proceeding. One accepted paper was withdrawn from the conference one month prior to its start. This left four papers for the proceeding. The final acceptance rate was then 4/21 = 19%. Meanwhile, 8 out of 25 talk proposals were accepted for full presentation at the conference, an acceptance rate of 32%. An additional group of 12 technical papers and talk proposals were invited for a poster session. Due to the challenges of the hybrid model, some of these posters were presented as pre-recorded videos with a live Q&A session on October 12, 2022, the virtual-only day of the conference. The rest of the posters were presented in-person as short lightning talks during a dedicated session on October 13, 2022 at Boston University. These proceedings include only the four accepted technical papers in their entirety, since they represent new research being published for the first time, as well as the title and authors of the accepted full and lightning talks.

We are very grateful to the TTO 2022 program committee members, composed of 40 individuals, for their dedication in selecting high-quality research for TTO, while providing constructive and encouraging feedback to all submissions. It is their generosity in volunteering their time that makes conferences such as TTO possible.

Eni Mustafaraj and Gianluca Stringhini,
TTO-2022 PC Co-Chairs

Organization

This conference is a truly unique collaboration between academia, industry, and practitioners.

General Chairs:
Jennifer Mathieu (Graphika)
Clara Tsao (TSPA)

Program Chairs:
Eni Mustafaraj (Wellesley College)
Gianluca Stringhini (Boston University)

Marketing and Publicity Chairs:
Justin Davis (Spectrum Labs)
Nevin Thompson (Credibility Coaltion)

Community Engagement Chair:
Ace Wang (Wellesley College)

Website Chair:
Keenan Jones (University of Kent)

Operations Manager:
Ahmed Medien (Hacks/Hackers)

Program Committee:

Ace Wang (Wellesley College)

Amy Zhang (University of Washington)

Andreas Vlachos (University of Cambridge)

Arkaitz Zubiaga (Queen Mary University of London)

Chen Ling (Boston University)

Christo Wilson (Northeastern University)

Clara Tsao (TSPA)

David Corney (Full Fact)

Diana Maynard (The University of Sheffield)

Dilrukshi Gamage (Tokyo Institute of Technology)

Elena Kochkina (Queen Mary University of London)

Emiliano De Cristofaro (University College London)

Emilio Ferrara (University of Southern California)

Fabiana Zollo (Ca’ Foscari University of Venice)

Firoj Alam (Qatar Computing Research Institute)

Francesca Spezzano (Boise State University)

Gerhard Weikum (Max Planck Institute for Informatics)

Gianluca Demartini (University of Queensland)

Giovanni Zagni (Pagella Politica/Facta)

Guy Aglionby (University of Cambridge)

Harith Alani (Knowledge Media Institute)

Jennifer Mathieu (Graphika)

Justin Davis (Spectrum Labs)

Kalina Bontcheva (The University of Sheffield)

Keenan Jones (University of Kent)

Kristen Johnson (Michigan State University)

Lee Mwiti (Africa Check)

Manoel Horta Ribeiro (Swiss Federal Institute of Technology Lausanne)

Naeemul Hassan (University of Maryland)

Neema Kotonya (Imperial College London)

Nguyen Vo (Worcester Polytechnic Institute)

Oana Cocarascu (King’s College London)

Piotr Przybyła (Institute of Computer Science, Polish Academy of Sciences)

Preslav Nakov (Hamad Bin Khalifa University)

Ricardo Campos (Instituto Politecnico de Tomar)

Sibel Adali (Rensselaer Polytechnic Institute)

Suzanna Ersoy (Government of Canada (UNSDG))

Tariq Alhindi (Columbia University)

Tristan Caulfield (University College London)

Zhijiang Guo (University of Cambridge)

Accepted Talks for Full Presentation

Is Personalised Content Moderation a Good Idea?
Bertie Vidgen and Paul Röttger

vera.ai - Verification assisted by Artificial Intelligence
Kalina Bontcheva, Symeon Papadopoulos, Nikos Sarris, Maria Bielikova, Miguel Colom, Fabio Giglietto, Denis Teyssou, Luisa Verdoliva, Jochen Spangenberg, Layla Gaye, Danae Tsabouraki, Patrick Aichroth, Zlatina Marinova, Alexandre Alaphilippe and Richard Rogers

The Data Access and Transparency (DATA) Index: Evaluating Transparency in Online Social Platforms
Shayne Longpre, Cameron Hickey, Manoel Ribeiro and Deb Roy

Identification and characterization of misinformation superspreaders on social media
Matthew DeVerna, Rachith Aiyappa, Diogo Pacheco, John Bryden and Filippo Menczer

The Audience Initiative: Assessing what drives trust and public value in online media
Maha Taki, Sonia Whitehead and Alasdair Stuart

Public reasoning towards truth and community trust: regulating online deliberative spaces for participatory AI and data governance
Willow Wong

Designing for Trust and Truth in Digital Intimacy
Vaughn Hamilton, Gabriel Kaptchuk, Allison McDonald and Elissa Redmiles

Crowdsourced Fact-Checking: How is BirdWatch Doing?
Mohammed Saeed, Gianluca Demartini and Paolo Papotti

Accepted Talks for Lightning Presentation

If Fact-Checking Were Easy, Would People Fact-Check?
Piers Howe, Bradley Walker, Andrew Perfors, Nicolas Fay and Yoshihisa Kashima

Developing disinformation detection models in low-resource contexts: Czech news article dataset for source-level credibility
Matyáš Boháček

2020 U.S. Internet Election Memes: Political Propaganda With More Than a Creative Information Disorder Twist
Stephanie McVicker and Ann Jabro

Developing Dialogue Systems to Combat Hate Speech: A Baseline Framework for Effective Counterspeech
Jonathan Wang and Marcus Tomalin

Trust-by-Proxy: Building Institutional Confidence Through the Use of Blockchain and Social Media Influencers
Tomás Guarna and Eric Gordon

Automated Fact-Checking: Challenges and Opportunities
Maximilian Reimer and Vinay Setty

Applying 'inoculation theory' to mass media to build resilience to mis- and disinformation at scale
Yvonne MacPherson and Alasdair Stuart

The Fake News Arms Race: How AI Can Create – and Detect – Fakes
Or Levi

Increasing the credibility of vaccine-related information on Wikipedia through collaboration
Netha Hussain, Alexandra Bornhoft and Ariel Cetrone

Labelling a Catastrophe: Digital Databases, Evidence, and Accountability in Syria
Amre Metwally

Real-Time Detection of Ephemeral Bot Attacks: The Case of Fake Trend Bots on Twitter Turkey
Tugrulcan Elmas

Technical Papers

YouTube COVID-19 Vaccine Misinformation on Twitter: Platform Interactions and Moderation Blind Spots

David S. Axelrod
Indiana University
Bloomington
Luddy SICE
daaxelro@iu.edu

Brian P. Harper
Indiana University
Bloomington
Luddy SICE
bpharper@iu.edu

John C. Paolillo
Indiana University
Bloomington
Luddy SICE
paolillo@indiana.edu

Abstract

While most social media companies have attempted to address the challenge of COVID-19 misinformation, the success of those policies is difficult to assess, especially when focusing on individual platforms. This study explores the relationship between Twitter and YouTube in spreading COVID-19 vaccine-related misinformation through a mixed-methods approach to analyzing a collection of tweets in 2021 sharing YouTube videos where those Twitter accounts had also linked to deleted YouTube videos. Principal components, cluster and network analyses are used to group the videos and tweets into interpretable groups by shared tweet dates, terms and sharing patterns; content analysis is employed to assess the orientation of tweets and videos to COVID-19 messages. From this we observe that a preponderance of anti-vaccine messaging remains among users who previously shared suspect information, in which a dissident political framing dominates, and which suggests moderation policy inefficacy where the platforms interact.

1 Introduction

During the COVID-19 pandemic, Social media platforms acted rapidly to staunch misinformation on their platforms. This has led to policies addressing misinformation that are inconsistent internally and between each other(Krishnan et al., 2021). Twitter updated its COVID-19 misinformation policy at the beginning of 2021 and throughout the year to address vaccine misinformation (Twitter, 2021). YouTube's COVID-19 policies were similar, but also added a separate vaccine misinformation policy in late September(YouTube, 2021a,b). Though different in format and exceptions, both policies obligate the respective platforms to moderate vaccine misinformation.

While misinformation is regularly deleted on both platforms, new misinforming content appears in a back-and-forth process from which external researchers may only catch glimpses. Assessing the quality and efficacy of anti-misinformation policy is therefore quite difficult, not least because of interactions between different social media platforms and how they may amplify each others' misinformation. While platform policies and their implementation are centered around the holdings of a specific corporation, users may use multiple platforms irrespective or ownership, and so may readily exploit inter-platform differences.

What then is the shared role of Twitter and YouTube in the spread of vaccine misinformation? How do the interfaces between the two platforms and their policy differences contribute to it? And what responsibility do users bear in the process of circulating such misinformation? This paper addresses these questions through a mixed-methods analysis of a set of tweets sharing video links to likely vaccine misinformation.

2 Background

Previous studies have suggested a link between social media misinformation and vaccine hesitancy behaviors (Loomba et al., 2021; Pierri et al., 2022), and while there may be reason to be cautious about proposing causal links (Valensise et al., 2021), it is clear that an anti-vaccine echo chamber resides among social media. COVID-19 research on YouTube has often attempted to assess whether or not its content is dominated by misinformation, with results generally indicating that news coverage predominated over misinformation on YouTube in 2020 (Knuutila et al., 2020; Paolillo et al., 2022; Marchal and Au, 2020; Andika et al., 2021; Li et al., 2020), though misinformation was always present. Other studies assess the quality of medical communication on the YouTube platform, finding mixtures of good and bad information quality (Szmuda et al., 2020; Basch et al., 2020). More recently, research has shifted from addressing the pandemic itself to assessing the state of vaccine messaging and its role in combating misinformation (Jennings et al., 2021; Laforet et al., 2022). The situation is broadly similar for research on Twitter, with early attempts to identify misinformation about the virus and the pandemic (Gallotti et al., 2020; Mourad et al., 2020) shifting toward vaccine misinformation (Yousefinaghani et al., 2021), often assessing the role of bots and the presence of active anti-vaccine campaigns (Shi et al., 2020; Sharma et al., 2022).

Rather than treat different platforms as separate entities, it is better to conceptualize them as parts of a platform ecosystem (Van Dijck et al., 2018), and cross-platform work is important for understanding both how platform systems interact with each other and how users experience individual platforms. Unfortunately, while good cross-platform work has been done (Cinelli et al., 2020), platform dyads contain individual patterns of interaction that are not necessarily visible on a grand scale. With respect to YouTube and Twitter, links to YouTube videos within COVID-19 tweets have been previously observed (Yang et al., 2021), and Ginossar et al. (2022) found evidence to suggest that Twitter links to YouTube videos were effective means of spreading misinformation during the first half of 2020, often using prior conspiratorial or anti-vaccine content on YouTube to do so. The present paper continues the work of observing this cross-platform interaction, albeit with different methods of data collection and analysis.

3 Methods

To address our questions, we conducted an exploratory analysis around a dataset likely to contain vaccine misinformation collected at the interface of Twitter and YouTube. This interface is hosted almost entirely on Twitter: while active YouTubers often have Twitter accounts and link to them, Twitter's feature of shortening and including links allows YouTube videos to be shared in discourses they would not otherwise reach. Twitter's public data API allows one to search tweets for a given time period, possibly filtering them for sites they link to, and YouTube's API allows one to ascertain if a video link still exists on the platform. Through these means, it is possible to construct a view of the YouTube-Twitter platform interface that reflects interaction around COVID-19 vaccine misinformation.

Following acquisition of a relevant dataset, two general research methods are available: quantitative exploration (e.g., via cluster analysis and network analysis), and qualitative analysis. The approach followed here is to employ different stages of quantitative and qualitative analysis to support each other. This is necessary as the data collected are of a substantial scale that is difficult to approach, and only systematic examination of the videos and tweet texts by researchers can tell us what they actually mean for COVID-19 vaccine discourse.

Since our dataset covers a broad time period (Jan 4 to Dec 31, 2021), the state of information around COVID-19 vaccines and treatment changes greatly during the sample period, meaning it is likely that patterns of video production on YouTube and/or sharing on Twitter also change. Hence, our first step is to organize the time period into approachable groupings of dates based on the video tweeting patterns over time. This information was then presented in the form of a web interface providing links to the relevant videos alongside the tweets from the database; after viewing selected videos, we could then make qualitative judgments in the interface regarding their content. These judgments were then analyzed for consistency and employed in subsequent network analyses. An additional set of analyses were conducted on the term distribution in the tweets; these did not reveal vaccine-related topics as had been hoped, but rather formulaic patterns used for marketing and spam, which were prevalent in our dataset.

Our tweets come from the CoVaxxy project, an effort that collected IDs for vaccine related tweets using almost 80 keywords (DeVerna et al., 2021) through Twitter's `statuses/filter` v1.1 API endpoint.[1] In order to examine how YouTube videos are shared in Twitter content that deals with vaccines we first retrieved the full Twitter data and metadata for each tweet, and then identified the URLs embedded in them, extracted unique video identifiers, and queried the YouTube API for the video status using the `videos:list` endpoint. Since previous research showed that inaccessible

[1] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview

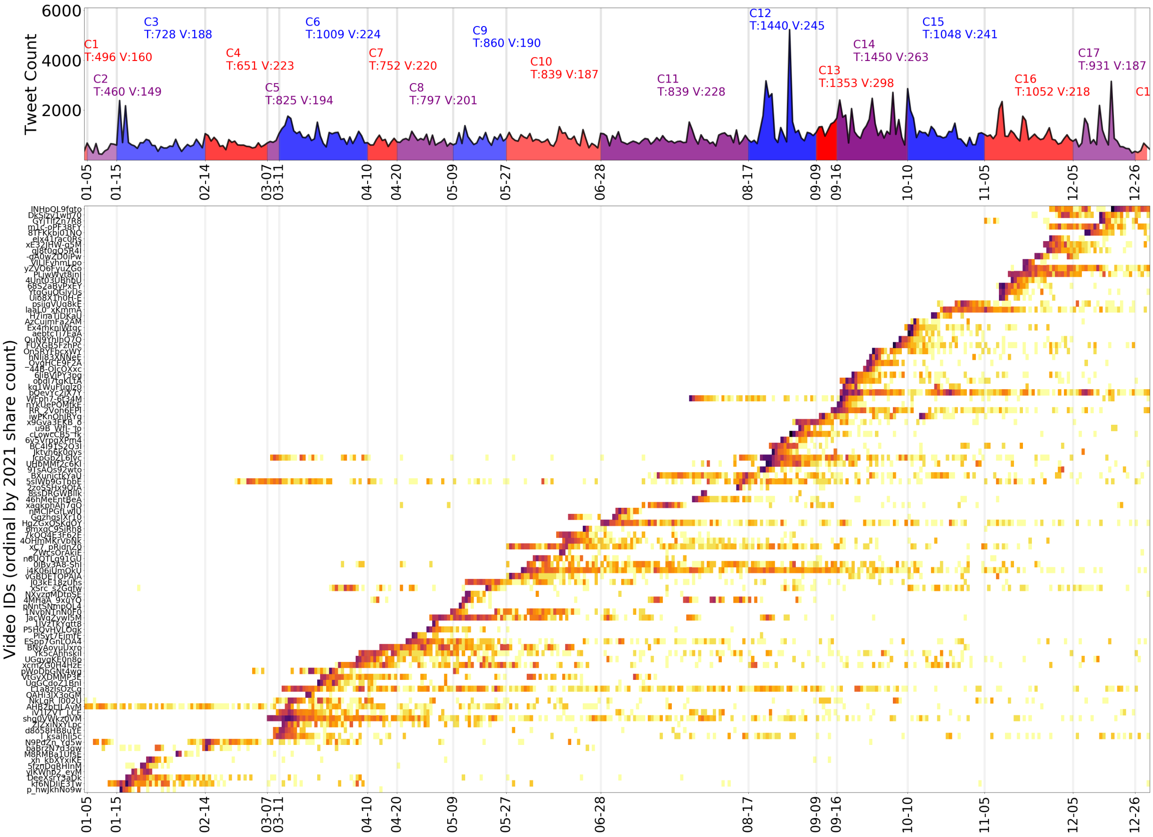


Figure 1: Top: Time series for tweets linking to YouTube videos in our dataset. The time series is segmented into date groupings (see section 4), labeled with cluster identities and counts of the tweets and videos in each. Dates shown are the last day in each date cluster. Vertical lines extend down into the heat map to guide the eye. Bottom: Heat map of 100 most shared videos (2021). Heat map color encodes daily frequency for each video. Left-hand y-axis lists video IDs ordered by the date of their maximum daily share count.

videos contain a high proportion of anti-vaccine content such as the “Plandemic” conspiracy documentary (Yang et al., 2021), we treat any videos removed by Twitter and publicly unavailable as suspicious. Overall, the fraction of vaccine-related tweets linking to YouTube videos in the CoVaxxy dataset was relatively small, with a daily median of 0.52%. However, among these links, a daily median of 10.95% were to inaccessible videos. Over the year, there was a decreasing trend in inaccessible videos with a peak of 45% in July. Because it is estimated that it takes an average of 41 days for YouTube to remove videos violating their terms (Knuutila et al., 2020), we checked the status of videos at least 2 months after the video was last posted on Twitter. Since unavailable content is not available for full investigation, we focus on related but available content by selecting tweets with available videos from users who also shared one of the unavailable videos in our dataset.

4 Grouping dates in the sample

The sampling procedure described above resulted in a set of 339,763 tweets (126,244 original and 213,519 retweets) containing 34,819 distinct YouTube video links; this is the cross-platform dataset we seek to explore. Our first question concerns whether there are discernible video tweeting patterns over time. We approached this by examining the timestamps of tweets sharing a common video; this is visualized as a heat map of the 100 most viewed videos in Figure 1, in which the diagonal pattern clearly indicates that each video is shared on or close to a specific date (generally close to its publication date, though there are exceptions), while in some cases a horizontal dashed line of points indicates that a video might be shared over some longer set of dates, not always close together in time.

To extract this pattern, a PCA was conducted on the video-date incidence matrix, which yielded three potentially usable dimensions accounting for

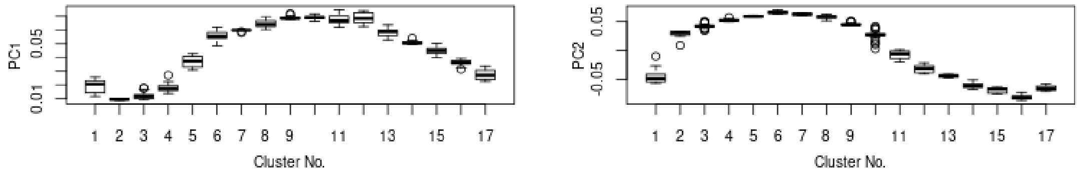


Figure 2: Boxplots of the date groupings on PC1 and PC2, showing each resulting date range to be distinct in terms of video sharing from the others in the sample.

20.0% of the shared variance. This was followed by cluster analysis of the dates on the PCs using the R package mclust (Scrucca et al., 2016); this approach permits us to evaluate a broad range of cluster solutions within a principled, model-based framework. A 17-cluster variable volume, shape and orientation (VVV) solution was selected as having optimal BIC; the clusters break the sequential dates into 17 contiguous groups, with the exception of cluster 1, which groups together the earliest and latest dates on account of having fewer videos shared with other dates (Figure 2). Hence, date clusters other than 1 suggest epochs within the Twitter sharing of vaccine-related content from YouTube either as periods of common activity or activity lapses between topics with concentrated attention.

5 Video Coding

To probe the nature of the date groupings, a content analysis interface was constructed in which a cluster and a date within the cluster could be chosen to present a list of videos to examine, as well as to provide access to the tweets sharing the video. To limit the coding work to manageable levels while focusing on important videos, we committed to coding only those videos tweeted or retweeted at least 5 times on any given day in 2021. Three coders participated in coding the sample, each coding alternating clusters throughout the year. In total, 5,201 videos out of the total 34,819 videos were coded. Observed thematic continuity and overlap across date clusters ensured that coders had a chance to view and discuss the full range of content in our dataset.

Five binary coding check boxes were displayed on each video, to indicate whether it had vaccine-related content **v**, whether it expressed positive and/or negative valence messages toward vaccination **p** and **n**, whether it had unusual characteristics potentially requiring discussion among the coders **q**, and whether the video possessed a non-publisher context panel **cx**. For coding **v**, we accepted ancillary topics like discussions of vaccine mandates and vaccine procedures whether or not claims about the efficacy of vaccination were directly made. The valance of videos **p** and **n** was understood to mean what impression of vaccination the video presents. This could be simultaneously positive and negative, as was common in the context of videos about vaccine debates. The fourth category **q** is intended for potential future work with this dataset and so may be ignored for the present. The fifth category **cx** assessed the presence of context panels — in the YouTube player interface, context bars are provided for user information such as regularly misinformed topics, and we wished to assess how consistently YouTube flagged misinforming videos. YouTube also uses context panels to inform users about state-owned media, but we ignored these cases, as they do not relate to potentially misinforming content so much as the quality of sources. In contrast to the other codes, little interpretation is required for coding **cx**.

The content coding was analyzed using a combination of PCA and linear modeling. This permitted us to identify patterns of correlation among the codes while simultaneously verifying consistency across coders. We conducted a centered, scaled PCA on the five video codes for the 5201 videos. Two PCs with variance exceeding 1 were retained: PC1 accounts for 51.5% of the total variation, whereas PC2 accounts for 32.3% (total 83.9%, residual 16.1%). As can be seen in Figure 3, on PC1, in order of decreasing strength, **v** (vaccine content), **cx** (context panel), **p** (positive valence), and **n** (negative valence) are shifted left (negative), whereas **q** (flag for further discussion) is on the extreme right (positive) end. This suggests that coding q is negatively correlated with vaccine-related content: as coders found more content there relevant to potential future discussions. PC2 separates the **n-p** dimension, though **cx** is loaded reasonably close to **p**. The lower proportion of shared variation suggests that **n** and **p** were less strongly inversely related than **q** and **v**. The loading of **cx** suggests that vaccine-related content was likely to bear a YouTube context panel, but most especially when a positive attitude is expressed than negative; both are otherwise less strongly correlated with **v** than **cx**. Hence, the coded videos are characterized by two dimensions: vaccine versus non-vaccine content, and positive or negative valence. Valence is more strongly associated with vaccine-related content, with positive and negative tending to exclude each other and context panels tending to appear on positive valence videos.

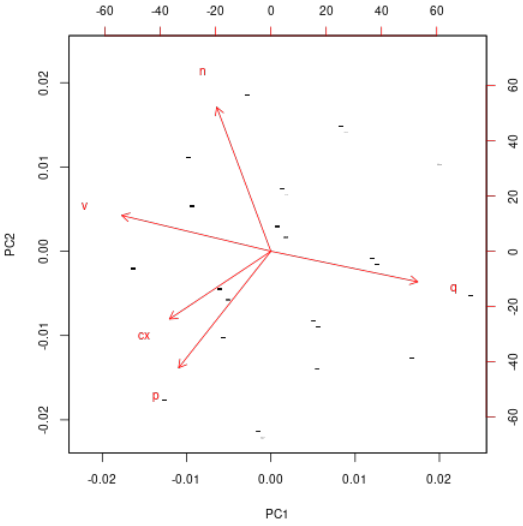


Figure 3: Biplot of the PCA for our coding results. Lines indicate association from videos with our codes.

5.1 Inter-rater reliability

Linear model (least squares) regressions of PC1 and PC2 with respect to coder were conducted to test for inter-coder reliability. So long as the contribution of coder to the explained variance of the regression is small, we can be assured that coders are consistent. As the PCs are orthogonal, they are tested individually; while coder is significant in both, the N is large and the proportion of variance accounted for by coder is very small (the adjusted R-squared is 0.9% for PC1, for PC2 it is 1.96%). So that we may readily inspect the inter-coder differences, we use sum contrasts (Table 1) and report the conditional means for each coder, centered on zero. For PC1, only the positive parameter is significant, meaning there is a significant difference between coders on either side of zero, but not between the two on the negative side. From this, it appears that coder 2 may have greater use of **q** or less use of **v** than the other coders. For PC2 both coder parameters are significant; while coder 3 is close to coder 2 in value, they appear to differ in use of the valence codes. Coder 1 appears to have significantly more **p/cx** than the other two. This could be a difference in either the coding of valence or in the way that YouTube has handled context panels across the different dates of our sample.

PC	**Coder1**	**Coder2**	**Coder3**
1	-0.0446	0.2406	-0.1960
2	0.2371	-0.1484	-0.0887

Table 1: Conditional means for coder models of PCs

As there are significant differences in the coding among the different coders, we would like to understand the consequence of this. Since coder is the sole categorical variable in these models, it shifts the intercepts for the videos coded on PC 1 and PC 2, and the residuals provide proxy scores for the videos independent of coder. These can also be transformed back to the original coding scale for each variable by rotating them with the PCA variable loadings, adjusting for the mean and standard deviation of the original variables, and setting a threshold of 0.5 to clip the values to 0 and 1. This gives a predicted consensus coding for each video across coders which can be checked against the original codes for consistency. When this is done, we end up with the differences for each code in Table 2. The column labels **f→f**, **f→t**, etc. indicate cases where the original coding of f corresponds to the predicted consensus coding of f or t, etc.

var	**f→f**	**f→t**	**t→f**	**t→t**
v	1255	21	33	3892
p	3353	42	61	1745
n	2937	30	19	2215
q	3744	46	139	1242
cx	2637	458	736	1378

Table 2: Pairwise comparisons of corrected and original codes.

The corrected **v**, **p**, and **n** are very close to their original values, **q** and **cx** less so. In Table 3, we summarize the corrections in terms of the individual coders, collapsing the values for **f→f** and **t→t** into a single row for each coder labeled **=**. As before, we find small numbers of corrections (predicted inter-coder differences) across variables **v**, **n** and **p**, and larger differences for **cx** and **q**. We conclude from this that we can safely use the original coding of **v**, **n** and **p** without alteration; in the subsequent diffusion analysis we do this by using PC2 without correction. Context panels (**cx**) require greater care in interpretation; coders coded different groups of dates, which could have different rates of deployment of the relevant context panels on YouTube. However, considering the greater interpretation required for **v**, **n** and **p** compared to the simple identification of a box existing in **cx**, this suggests that it is plausible that the differences in coding **cx** stem from the dataset rather than coding errors.

	Δ	**v**	**p**	**n**	**q**	**cx**
1	f→ t	14	36	27	15	231
	t→ f	3	37	6	47	404
	=	2716	2660	2700	2671	2098
2	f→ t	0	0	0	23	96
	t→ f	4	6	2	18	139
	=	1311	1309	1313	1274	1080
3	f→ t	7	6	3	8	131
	t→ f	26	18	11	74	193
	=	1120	1129	1139	1071	829

Table 3: Per-coder change comparisons (Δ) between original and corrected variables.

6 Linking Clusters

Finally, we constructed a cluster analysis according to content (video) diffusion paths. To start we construct a user-link bipartite network where links are represented by the original shortened URLs. Shortened URLs can be generated by the system many times per video and so index a specific sharing path for a video. We apply the Louvain community detection algorithm (Aynaud, 2020) to the link-link projection of 88,958 nodes and 9,043,668 edges, arriving at 2,305 community solution. These communities are then treated as meta-nodes and used these nodes to construct a community-video bipartite network, on which we again apply the Louvain algorithm, arriving at 10 clusters of diffusion communities that share videos in common between them. Figure 4 shows these 10 clusters as meta-nodes with edges between cluster and self loops weighted according to the number of links between and within a cluster, respectively. The average loading of the cluster's constituent nodes onto PC2 from the content analysis is mapped as the gradient color, with darker shades being more toward the loading of **n** and lighter being more toward the loading of **p**. Table 4 lists the top three most-shared videos among each of these clusters.

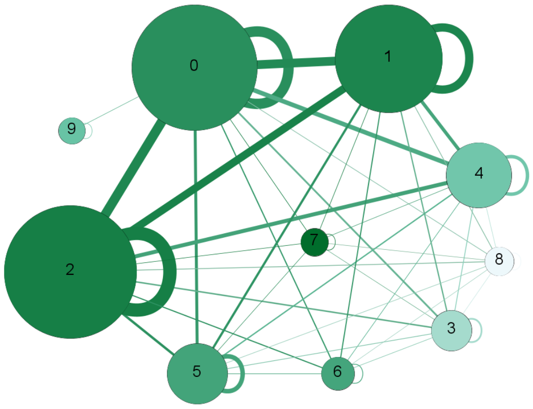


Figure 4: 10 clusters represented by meta-nodes with edges and self-loops representing links between clusters and internal linking, respectively. Clusters were determined by applying the Louvain community detection algorithm on the links-links projection of the links-video network. Node color brightness encodes average loading of constituent nodes onto PC2.

Cluster 0, 1, 2, 4, and 5 dominate interaction within this model. Cluster 0 is characterized by mixed anti-vaccine content from multiple sources that otherwise overlaps with that of other clusters. This cluster represents the sharing behavior of twitter users who are not especially tied to particular domains within vaccine discourse beyond being broadly anti-vaccine. Cluster 1 contains YouTube videos from content creators who try to present themselves as not anti-vaccine despite producing content that is regularly consumed by anti-vax audiences. The Jimmy Dore Show and Dr. John Campbell are representative of this cluster. Cluster 2 predominantly contains content related to Project Veritas, a channel premised on exposing corporate or left-wing "conspiracies." This content is very clearly anti-vaccine, but from conspiratorial perspectives.

Cluster 3 contains largely the channel Workout

Solutions Health Fitness, which produces generally benign content on YouTube but employs a novel Twitter strategy of frequent self-promotion with high anti-vaccine messages. Cluster 4 contains more politically-minded anti-vaccine figures, like Russel Brand or Paul Joseph Watson. Cluster 5 includes doctors or other experts arguing against some part of pandemic public health guidelines. These doctors, such as Dr. Geert Vanden Bossche or Dr. Peter McCullough, are interviewed on many different YouTube channels. Cluster 6 contains mostly news reports that are often pro-vaccination or neutral on their own, but the tweets linking to them are at times arguing against these videos. Cluster 7 contains videos discussing vaccine or virus-related topics in technical detail that are not consistently pro or anti vaccine. Cluster 8 contains videos from 7NewsAustralia, an Australian news broadcaster who consistently promotes vaccination. Cluster 9 consists of Christian fundamentalist antivax narratives, largely from the channel A Voice in the Desert.

Cluster 1, 2, 4, and 5 can be seen as different archetypes of popular anti-vaccine discourse in our dataset, with 1 representing a more mainstream hesitancy, 2 representing conspiratorial discourse, 4 representing political discourse intersecting with anti-vaccine messages, and 5 representing the fusion of apparent expertise with anti-vaccine messages. These are held together by Cluster 0, representing the Twitter sharing of videos by people who sharing broadly anti-vaccination messages.

7 Discussion and Conclusion

We began our investigation by sampling tweets at the Twitter-YouTube interface among users affected by vaccine-related moderation, among whom there is a propensity of anti-vaccine and anti-authority messages. Some use of pro-vaccine videos is made, but typically in refutation; other pro-vaccine videos are pushed to the edges of the discourse. Hence, we observe that the discursive impetus which led to moderating anti-vaccine information in the first place continues in spite of those efforts. Sharing specific anti-vaccination information may become harder under moderation, but the same users will often still share closely-related content.

Anti-vaccine discourse appears to persist through a predominantly political framing. Well-known figures taking anti-authoritarian stances on public health measures provide a framework into which dissident researchers, doctors, nurses, etc. can slot topically-relevant messages. These messages sow further doubt, and thereby these actors potentially rise to a new kind of prominence. John Campbell, Peter McCullough and Geert Vanden Bosche are three examples of such professionals whose public communication profiles have been raised as vaccination skeptics. Anti-vaccine content further differentiates along a political-medical axis among the users sharing it.

At times, the connection of video content to vaccination is more tangential, e.g., in the case of Workout Solutions (cluster 3), whose Twitter account contributed an astounding 30,417 tweets to our dataset, only 6,920 of which were retweets; the next nearest user had only 2,916 tweets. These tweets were highly formulaic, consisting of recombined phrases along with numerous hashtags and mentions for Canadian political figures (the account owner and related YouTube channel are in Ontario). Here, it appears that someone attempted to draw attention to his YouTube channel, possibly unsuccessfully, by linking to ongoing anti-lockdown/anti-vaccination political discourses using some kind of third-party Twitter app.

It also appears that moderation at the interface between YouTube and Twitter is not entirely successful. In part, this can be attributed to failures to fully implement the moderation stances adopted by the individual platforms. One such failure is the tendency for pro-vaccine messages to be flagged by context bars on YouTube while misinforming content is left un-flagged. It is possible that the means by which the context bars are applied is too simplistic, e.g. keyword searches in titles, descriptions and transcripts that are readily evaded by mutating the keywords. Alternative means, e.g. tracking known misinforming personalities, would potentially be much more effective, and more closely resembles how editorial discretion is exercised in print and broadcast media, but may aggravate such figures' dissident political stances.

Regardless, the policy differential between Twitter and YouTube *vis a vis* COVID-19 vaccination causes leakage. Other strategies at the disposal of a platform such as YouTube, such as search down-ranking of videos, are entirely side-stepped by cross-platform linking, where the original platform no longer has control over the spread

Cl#	Links	Video Title (truncated)
0	580	Mass Vaccination in a Pandemic - Benefits versus Risks: Interview with Geert...
	252	Vaccines and Related Biological Products Advisory Committee – 9/17/2021
	239	Spike protein inside nucleus enhancing DNA damage? - COVID-19 mRNA...
1	78	EXPLOSIVE Truth About Vaccines & COVID w/Inventor Of mRNA Vaccine...
	50	Spike protein inside nucleus enhancing DNA damage? - COVID-19 mRNA vaccines...
	50	Kyle's vaccine complication
2	157	Carnicom Institute Disclosure Project - Overview with Clifford Carnicom
	104	Pfizer Scientists: 'Your [COVID] Antibodies Are Better Than The [Pfizer]...
	100	Johnson & Johnson: 'Kids Shouldn't Get A F*cking [COVID] Vaccine;'...
3	67	Vaccines and Related Biological Products Advisory Committee – 9/17/2021
	55	Back to School Morning Outdoor Weighted Cardio Rogue Fitness Fat Boy Sled
	53	Bring Back DDT & Dr Conover's Antibiotics For Wolbachia Co-Infections
4	52	Vaccine Passports: THIS Is Where It Leads
	51	Million March for Freedom Rally - London
	42	Krystal Ball: Bill Gates Is LYING TO YOU On Vaccine Patent Protection
5	122	Mass Vaccination in a Pandemic - Benefits versus Risks: Interview with Geert...
	29	#ScreenB4Vaccine: An Interview between Hooman Noorchashm MD, PhD and...
	25	Peter McCullough, MD testifies to Texas Senate HHS Committee
6	34	Vaccines and Related Biological Products Advisory Committee - 10/22/2020
	22	Eric Clapton: CANCELLED for exposing COVID-19 Vaccine
	10	'Natural Immunity' Lawsuit Over COVID-19 Vaccine Mandate Ends in Surprising
7	18	The Inventor of mRNA Vaccine Technology: Dr Robert Malone
	5	Elon Musk on mRNA "You could turn someone into a freaking butterfly with the...
	2	Italy Lawmaker Cunial Demands Arrest of "Vaccine Criminal" Bill Gates
8	20	Rare inside tour of German lab creating mass Pfizer COVID-19 vaccines from...
	16	Prime Minister Scott Morrison among the first Australians to receive COVID-19...
	14	ATAGI recommends Pfizer vaccine be offered to Australian children as young as 12...
9	7	Jesus and the Mark of the Beast
	4	The Scientific Method And Jesus
	4	The Truth About Christianity That Nobody Tells You

Table 4: Cluster number (Cl#) with the top 3 linked videos (listed by their titles).

of the content. Consequently, there appears to be a need for closer inter-platform cooperation in establishing and implementing moderation policies, although this too is likely to encounter political resistance from people who believe that either platform might unfairly moderate their messages.

Our cross-platform approach has a number of limitations. Identifying topics through term clusters highlighted spam strategies used on Twitter. With respect to YouTube, video deletion for reasons other than terms of service violations raised further issues. For example, pro-vaccine videos from Indian state media entered into our sample due to the channel's decision to delete certain livestreams, potentially raising the number of pro-vaccine videos in our dataset. Similarly, videos hosted on YouTube but marked as "unlisted" may be found through Twitter; the reasons that content creators delist their videos are not necessarily consistent or clear. More generally, Twitter and YouTube users can always use the platform in unusual ways, and more careful culling from broad queries like that of CoVaxxy might be needed. Future work could address some of these data collection concerns, especially if differentiating the reason for the deletion of the original YouTube videos could be ascertained. This is not yet provided for by the YouTube API. Alternatively, future work could investigate this and similar datasets through more detailed content analysis approaches.

Acknowledgments

This work is supported in part by the Knight Foundation and the National Science Foundation (grant ACI-1548562). We also want to thank John Bryden and the rest of the CoVaxxy team for facilitating access to CoVaxxy data and Kai-Cheng Yang for helpful feedback.

References

Reynold Andika, Chien T Kao, Christopher Williams, Young J Lee, Hassan Al-Battah, and Richard Alweis. 2021. Youtube as a source of information on the covid-19 pandemic. *Journal of Community Hospital Internal Medicine Perspectives*, 11(1):39–41.

Thomas Aynaud. 2020. python-louvain x.y: Louvain algorithm for community detection. https://github.com/taynaud/python-louvain.

Corey H Basch, Grace C Hillyer, Zoe C Meleo-Erwin, Christie Jaime, Jan Mohlman, and Charles E Basch. 2020. Preventive behaviors conveyed on youtube to mitigate transmission of covid-19: cross-sectional study. *JMIR public health and surveillance*, 6(2):e18807.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10.

Matthew DeVerna, Francesco Pierri, Bao Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Cristopher Torres-Lugo, Kai-Cheng Yang, Fil Menczer, and John Bryden. 2021. Covaxxy: A global collection of english twitter posts about covid-19 vaccines. *Proceedings of the International AAAI Conference on Web and Social Media*.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, 4:1285–1293.

Tamar Ginossar, Iain J Cruickshank, Elena Zheleva, Jason Sulskis, and Tanya Berger-Wolf. 2022. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in youtube videos shared in early twitter covid-19 conversations. *Human Vaccines & Immunotherapeutics*, pages 1–13.

Will Jennings, Gerry Stoker, Hannah Bunting, Viktor Orri Valgarðsson, Jennifer Gaskell, Daniel Devine, Lawrence McKay, and Melinda C Mills. 2021. Lack of trust, conspiracy beliefs, and social media use predict covid-19 vaccine hesitancy. *Vaccines*, 9(6):593.

Aleksi Knuutila, Aliaksandr Herasimenka, Hubert Au, Jonathan Bright, Rasmus Nielsen, and Philip N Howard. 2020. COVID-related misinformation on YouTube: The spread of misinformation videos on social media and the effectiveness of platform policies. Oxford, UK: Project on Computational Propaganda. https://comprop.oii.ox.ac.uk/research/posts/youtube-platform-policies/.

Nandita Krishnan, Jiayan Gu, Rebekah Tromble, and Lorien C Abroms. 2021. Research note: Examining how various social media platforms have responded to covid-19 misinformation. *Harvard Kennedy School Misinformation Review*, 2(6):1–25.

Priscila E Laforet, Corey H Basch, and Hao Tang. 2022. Understanding the content of covid-19 vaccination and pregnancy videos on youtube: An analysis of videos published at the start of the vaccine rollout. *Human Vaccines & Immunotherapeutics*, pages 1–8.

Heidi Oi-Yee Li, Adrian Bailey, David Huynh, and James Chan. 2020. Youtube as a source of information on covid-19: a pandemic of misinformation? *BMJ global health*, 5(5):e002604.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348.

Nahema Marchal and Hubert Au. 2020. "coronavirus explained": Youtube, covid-19, and the sociotechnical mediation of expertise. *Social Media+ Society*, 6(3):2056305120948158.

Azzam Mourad, Ali Srour, Haidar Harmanani, Cathia Jenainati, and Mohamad Arafeh. 2020. Critical impact of social networks infodemic on defeating coronavirus covid-19 pandemic: Twitter-based study and research directions. *IEEE Transactions on Network and Service Management*, 17(4):2145–2155.

John Paolillo, Brian Harper, and David Axelrod. 2022. Covid-19 information on youtube in the early pandemic. In *Proceedings of the 55th Hawaii International Conference on Systems Science*.

Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. *Scientific Reports*, 12(1):1–7.

Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. 2016. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289.

Karishma Sharma, Yizhou Zhang, and Yan Liu. 2022. Covid-19 vaccine misinformation campaigns and social media narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 920–931.

Wen Shi, Diyi Liu, Jing Yang, Jing Zhang, Sanmei Wen, and Jing Su. 2020. Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter. *International Journal of Environmental Research and Public Health*, 17(22):8701.

Tomasz Szmuda, Mohammad Talha Syed, Akshita Singh, Shan Ali, Cathrine Özdemir, and Paweł Słoniewski. 2020. Youtube as a source of patient information for coronavirus disease (covid-19): a content-quality and audience engagement analysis. *Reviews in Medical Virology*, 30(5):e2132.

Twitter. 2021. Covid-19 misleading information policy.

Carlo M Valensise, Matteo Cinelli, Matthieu Nadini, Alessandro Galeazzi, Antonio Peruzzi, Gabriele Etta, Fabiana Zollo, Andrea Baronchelli, and Walter Quattrociocchi. 2021. Lack of evidence for correlation between covid-19 infodemic and vaccine acceptance. *arXiv preprint arXiv:2107.07946*.

José Van Dijck, Thomas Poell, and Martijn De Waal. 2018. *The platform society: Public values in a connective world*. Oxford University Press.

Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. 2021. The covid-19 infodemic: Twitter versus facebook. *Big Data & Society*, 8(1):20539517211013861.

Samira Yousefinaghani, Rozita Dara, Samira Mubareka, Andrew Papadopoulos, and Shayan Sharif. 2021. An analysis of covid-19 vaccine sentiments and opinions on twitter. *International Journal of Infectious Diseases*, 108:256–262.

YouTube. 2021a. Managing harmful vaccine content on youtube.

YouTube. 2021b. Vaccine misinformation policy - youtube help.

The role of online attention in the supply of disinformation in Wikipedia

Anis Elebiary[†] and **Giovanni Luca Ciampaglia**[† ‡]
[†] University of South Florida, Tampa, FL
[‡] University of Maryland, College Park, MD
anis@usf.edu, gciampag@umd.edu

Abstract

Wikipedia and many User-Generated Content (UGC) communities are known for producing reliable, quality content, but also for being vulnerable to false or misleading information. Previous work has shown that many hoaxes on Wikipedia go undetected for extended periods of time. But little is known about the creation of intentionally false or misleading information online. Does collective attention toward a topic increase the likelihood it will spawn disinformation? Here, we measure the relationship between allocation of attention and the production of hoax articles on the English Wikipedia. Analysis of traffic logs reveals that, compared to legitimate articles created on the same day, hoaxes tend to be more associated with traffic spikes preceding their creation. This is consistent with the idea that the supply of false or misleading information on a topic is driven by the attention it receives. These findings improve our comprehension of the determinants of disinformation in UGC communities and could help promote the integrity of knowledge on Wikipedia.

1 Introduction

In recent years several Internet websites have become the hubs for communities where users can produce, consume, and disseminate content without central oversight. Examples of these *user-generated content* (UGC) websites include major social media platforms, like Facebook or Twitter, or global online knowledge production communities like Wikipedia, which is known as a model for the production of vast reliable, high-quality knowledge (Yasseri and Menczer, 2021).

However, a negative consequence of the popularity of UGC websites is that their low barriers to access, combined with the lack of supervision from experts or other gatekeepers, results in the proliferation of false or misleading information on the Web as a whole (Wardle and Derakhshan, 2017; Lazer et al., 2018).

False or misleading content often spreads on social networking platforms (Amoruso et al., 2020; Castillo et al., 2011; Zareie and Sakellariou, 2021; Grinberg et al., 2019; Guess et al., 2019, 2020; Allcott and Gentzkow, 2017), but there are growing concerns that other UGC communities like Wikipedia may be vulnerable to these threats too (Sáez-Trumper, 2019). This is especially worrisome since Wikipedia is one of top most visited internet websites (Similarweb LTD, 2022) and a popular source of knowledge (Okoli et al., 2014). Wikipedia contains over 50 million articles in more than 300 languages; in February 2022, the English language edition of Wikipedia alone received 781M visits (from unique devices) and was edited over 5M times (Wikipedia contributors, 2022c; Wikimedia Foundation, Inc., 2022d). Hence, preserving the integrity of Wikipedia is of paramount importance for the Web as a whole (Sáez-Trumper, 2019).

There are many potential threats to the integrity of knowledge in Wikipedia (Sáez-Trumper, 2019). One common threat comes from vandalism, which is “a deliberate attempt to compromise the integrity of the encyclopedia, often through the insertion of obscenities, insults, nonsense or crude humour, or by page blanking” (Wikipedia contributors, 2021).

Vandalism, however, is not the only threat to the integrity of Wikipedia’s content. Whereas vandalism focuses on defacing existing entries, there exists evidence showing that Wikipedia is also targeted by *hoaxes*, whose aim is to create whole new entries about fake, fictitious topics. An example of a famous Wikipedia hoax is the entry *Jar’Edo Wens*, a fake Australian Aboriginal deity, which went undetected for almost 10 years before being debunked and deleted (Dewey, 2015). But hoaxes remain a threat to Wikipedia's content integrity to this day. Recently, one of the largest such incidents the platform has ever seen has been discovered on the Chinese Wikipedia: a user named *Zhemao* wrote 206 fake entries, starting from 2019 until 2022, about Russia's history in the Middle Ages (Moon, 2022).

Hoaxes are thus not to be confused with vandalism; although vandalism is a much bigger threat in scope and size compared to hoax articles, hoaxes constitute a more subtle threat, which has received less attention compared to vandalism.

A crucial question that remains unresolved is what drives the creation of hoaxes on Wikipedia. Because their original authors are aware that these articles are false, hoax articles are different from mere *misinformation*, but should rather be considered instances of *disinformation* (Wardle and Derakhshan, 2017; Lazer et al., 2018). As such, understanding the factors that determine the supply of hoaxes on Wikipedia could shed light on disinformation in general, including broader threats to the integrity of the Web, like state-sponsored propaganda (King et al., 2017; Zannettou et al., 2019; Golovchenko et al., 2020) and conspiracy theories (Starbird, 2017).

To bridge this gap, in this paper, we study the role of *online attention*, in the form of individual page views, in the supply of disinformation in Wikipedia. The idea of an economy of attention was first introduced by Simon (1971), who observed that human attention is a limited resource that needs to be allocated (Goldhaber, 1997). Here, to quantify the flow of collective attention to individual topics of knowledge, we take advantage of the unique Wikipedia traffic dataset and API. Specifically, in this work we seek to answer the following questions:

Q1. Does online attention toward a topic increase the likelihood of disinformation being created about it?

Q2. Operationally, is there a relationship between traffic to Wikipedia and the production of hoax articles?

To answer these questions, we collected a list of known hoax articles (Wikipedia contributors, 2022a) along with their creation timestamps and content. To control for potential confounding factors in the distribution of traffic to Wikipedia over time, for each hoax, we considered a cohort consisting of all the legitimate (i.e. non-hoax) Wikipedia articles that were created on the same day as the hoax. Similar to Kumar et al. (2016), we find that hoaxes differ from legitimate articles in key appearance features, but do not strongly differ in the number of hyperlinks they contain. Next, for each article (either hoax or non-hoax), we parsed its content and extracted all the outlinks, i.e. its neighbors in the Wikipedia hyperlink network. The presence of a link between two Wikipedia entries is an indication that they are semantically related. Therefore, traffic to these neighbors gives us a rough measure of the level of online attention to a topic *before* a new piece of information (in this case an entry in the encyclopedia) is created.

Finally, we measure the relative change of traffic in the 7-day period before and after the creation of a hoax and compare this change to that of the articles in its cohort. To preview our results, we find that, on average, online attention tends to precede the creation of hoaxes more than it does for legitimate articles. This observation is consistent with the idea that the supply of false and misleading information on a topic is driven by the attention it receives.

In the rest of the paper we discuss related work (Section 2), and then describe our methodology (Section 3): the details of the data collection process, the comparison between features of hoaxes and legitimate articles, and the pre-processing of the Wikipedia traffic data. Section 4 discusses the techniques used to quantify online attention and its relationship to the hoax creation, and the statistical procedures performed to asses the results. Finally, section 5 summarizes our findings and future directions.

All code and data needed to replicate the findings of this study are available on Github at github.com/CSDL-USF/wikihoaxes.

2 Related Work

Over the years Wikipedia has developed an impressive array of socio-technical solutions to ensure the quality of its content. Early work on Wikipedia has shown that most acts of vandalism are repaired manually by the crowd of contributors within a matter of minutes (Viégas et al., 2004). In addition to human interventions, automated tools like ClueBot NG play a crucial role in keeping the encyclopedic entries clear from damage (Geiger and Halfaker, 2013; Halfaker and Riedl, 2012). On top of these methods, there exist other preventive measures such as patrolling recent changes, creating watchlists, blocking frequent vandalism creators, and using editorial filters. Finally, multiple research attempts have been conducted to aid in both the manual and the automatic detection of vandalism (Potthast et al., 2008; Adler et al., 2010; Smets et al., 2008; Harpalani et al., 2011).

Despite this wealth of work, little is known about Wikipedia hoaxes. Kumar et al. (2016) collected a sample of known hoaxes from resources compiled by the Wikipedia community, and studied their longevity, along with other characteristics. They found that one in a hundred hoaxes remain undetected for more than a year, with 92% of the cases detected within the first day. They also observed that, although only 1% of all hoaxes remain undetected for more than a year, those that stay undetected have a higher chance over time of remaining so. Finally, they showed that, on average, hoaxes have a lower density of internal links and receive less traffic than legitimate (i.e., non-hoax) articles (Kumar et al., 2016).

Traffic to Wikipedia has been used before to study collective attention. García-Gavilanes et al. (2017) studied the patterns of attention to Wikipedia in the wake of airplane crashes. They found that the traffic to entries about *previous* airplane crashes was greater than that of the current crash, i.e. the one that triggered the attention surge (García-Gavilanes et al., 2017). Ciampaglia et al. (2015) studied traffic patterns during the creation of new Wikipedia entries (i.e., not just hoaxes) and observed that the creation of new information about a topic is preceded by spikes of attention toward it, as measured by traffic to neighboring entries (Ciampaglia et al., 2015). This is consistent with a model in which the demand for information on a topic drives the supply of novel information about it. Consequently, measuring traffic to Wikipedia entries can help us get a step closer to understanding why and when hoaxes are more likely to be produced.

3 Data and Methods

We first describe how the dataset of hoaxes was collected and the process of building the cohort of each hoax.

3.1 Data Collection

Prior work has relied on a broad definition of 'hoaxes' that leverages the 'New Page Patrol' (or NPP) process (Kumar et al., 2016). Unfortunately, access to these data was not public due to the nature of the NPP process. Therefore, in the present work we relied on a smaller, public list documenting known hoaxes discovered by Wikipedia editors outside of the NPP process (Wikipedia contributors, 2022a). To be included in this list, a discovered hoax must meet either of the following two characteristics: (*i*) they have gone undetected for more than a month after patrolling (Kumar et al., 2016), or (*ii*) they were discussed by reliable media sources.

To collect this list, we queried the Wikipedia API using the 'prefix search' endpoint (MediaWiki contributors, 2022a) to collect the titles of the hoaxes residing in the administrative list maintained by Wikimedia under the prefix 'List of Hoaxes on Wikipedia'. The total number of titles retrieved was $N_h = 190$. We then used the Toolforge (Wikitech contributors, 2021) to query the database replica of the English Wikipedia for the creation date of each hoax article, defined as the timestamp of the first revision recorded in the database. Figure 1 (left) shows a summary of the number of hoaxes created over time, with the majority of hoaxes appearing in the period 2005–2007, and a decline starting in 2008. This observed behavior can be in part explained by the fact that the Wikipedia community started patrolling new pages in November of 2007 (Kumar et al., 2016; Wikipedia contributors, 2022b) and is also consistent with the well-known peak of activity of the English Wikipedia community (Halfaker et al., 2013).

Finally, to build the cohort of each hoax, we queried the Wikipedia database replica for all legitimate articles created on the same day. Since Wikipedia entries are often accessible through different titles, in collecting the cohort, we resolved all redirects created the same day as the hoax. Treating these redirects as separate entries would inflate the cohort size and could skew traffic statistics used later for estimating the level of online attention. Figure 1 (right) shows the effect that redirects have on the size of each cohort. In some cases, failing to account for redirects can increase the size of cohorts to up to $16,000$ articles.

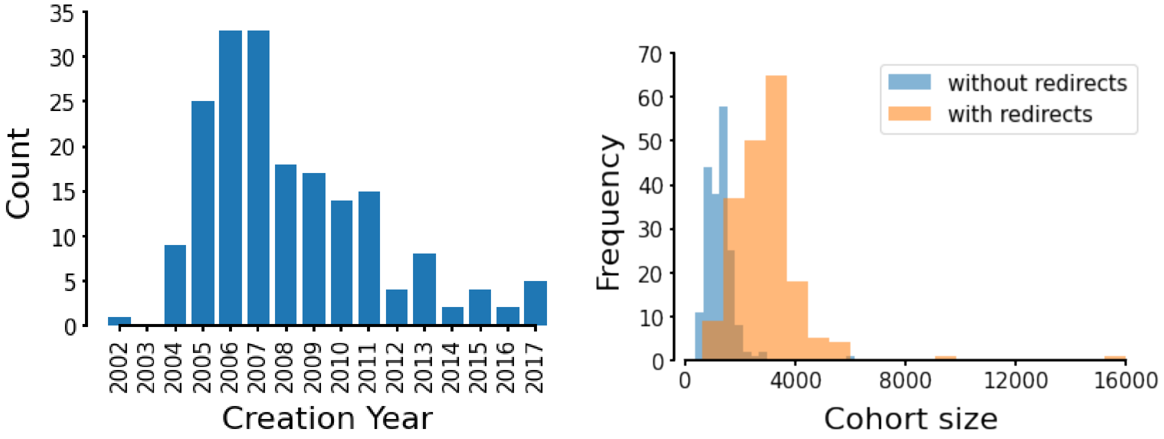


Figure 1: Left: Hoaxes detected in the English Wikipedia. Right: Cohort size distribution for hoaxes in our dataset before (solid blue) and after (solid orange) resolving redirects.

3.2 Appearance Characteristics Analysis

To understand the differences between each hoax and its cohort members, we analyzed their appearance features, inspired by the work of Kumar et al. (2016) who, in addition to appearance features, studied network, support, and editor features for both hoax and legitimate articles (Kumar et al., 2016). We considered the following four features: (*i*) the *plain text length* is the number of words in an article after removing all the wiki markup; (*ii*) the *ratio of plain to markup text* is the number of words obtained after removing all markup in the body of the article, divided by the number of words before removal; (*iii*) the *density of wiki-links* is the number of wiki-links per 100 words, counted before markup removal; and, finally, (*iv*) the *density of external links* is defined similarly as the density of wiki-links, but for links to external Web pages.

To be able to calculate these features for each hoax and its cohort, we consulted the API to extract their plain text using the *TextExtracts* extension instead (MediaWiki contributors, 2021). For the wiki markup we used the revisions API (MediaWiki contributors, 2022b). A regular expression was used to count the number of words in plain and markup text. Finally, to find the wiki and external links within each article we used *wikitextparser* (5j9, 2022).

Aside from the plain text to markup ratio, the chosen appearance features have very skewed distributions. To illustrate this point, fig. 2 shows the distribution of each score for five manually sampled cohorts in our data. For the plain text length, fig. 2 shows that the median is between 100 and $1,000$ words, yet there exist articles that reach and even exceed $10,000$ words. The same case persists in the wiki-link density — the median is under 10 links per 100 words, however some articles have up to 40 links, and similar for the other two features.

Thus, after collecting all the four features, we computed the modified z-score z' to compare different hoaxes together:

$$z' = \frac{x - \tilde{x}}{\mathrm{MAD}} \qquad (1)$$

Where x is a feature measured on a hoax, $\tilde{x}$ the median value of the feature on the non-hoaxes, and MAD the median absolute deviation of x with respect to $\tilde{x}$. We chose to use z' instead of the regular z-score since it is more resilient to outliers (Iglewicz and Hoaglin, 1993).

3.3 Analyzing Wikipedia Traffic Data

To analyze the traffic that the articles in our dataset receive, we used a dataset on traffic compiled by the Wikimedia foundation (Wikimedia Foundation, Inc., 2022b). The Wikimedia Foundation has published two main traffic datasets: the earlier *pagecounts-raw* (Dec. 2007–Aug. 2015), and the more recent *pageviews* (started Jul. 2015). Since most of the hoaxes in our dataset were created in the period between 2005 and 2011, we have decided to use the older pagecounts-raw data. This dataset contains the count of non-unique HTTP requests made for each article in an hourly time frame, collected by the proxy server of Wikipedia (Ciampaglia et al., 2015), along with request title and additional metadata. We pre-

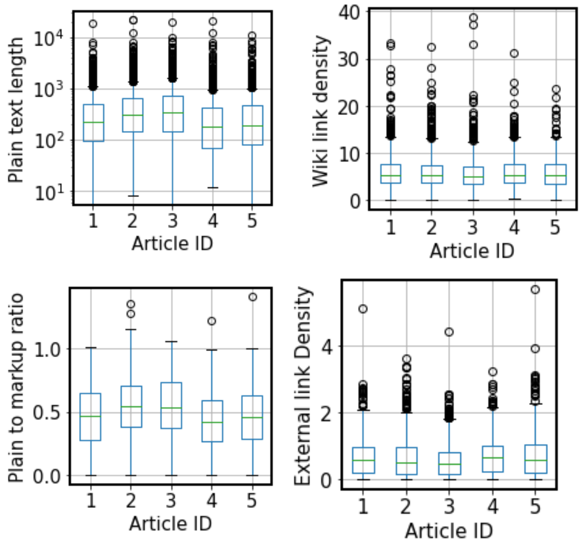


Figure 2: Distribution of appearance features for five manually sampled cohorts in our data.

processed pagecounts-raw to resolve redirects, filter unwanted entries, and clean illegal titles.

Pre-processing the data was performed over the following three steps. First, the raw data was filtered. The filtration process selected only entries related to the English Wikipedia project while removing all pages from namespaces other than the 'main' MediaWiki namespace. Second, the filtered data was cleaned from illegal titles. Illegal titles were discarded by removing characters which are not allowed in Wikipedia page titles (Wikimedia Foundation, Inc., 2022a, cf. 'Page Restrictions'). The hashtag sign '#' is considered illegal only if it is the first character in a title; otherwise it indicates a subsection within a page. Hence, a title including '#' is discarded only in the former case. In addition to removing illegal characters, we decoded common URL-encoded characters (e.g. '%20') and replaced any space with an underscore character. Third, to resolve redirects, the Toolforge was consulted to extract all the redirects within the main namespace of the English Wikipedia. The result was a cleaned and filtered hourly dataset of the view count for pages within the main namespace of the English Wikipedia.

4 Results

4.1 Appearance Features

We start by analyzing the appearance features of hoaxes relative to the non-hoaxes in their cohort. Figure 3a shows that most hoaxes have either similar or slightly smaller plain text length compared to that of their cohorts. We also observe the presence of several outliers, indicating that a subset of hoaxes in our sample tends to have unusually higher word counts. This is consistent with the results of Kumar et al. (2016), who observed that 'successful' hoaxes (i.e., that have gone undetected for at least 1 month) have a median plain text length of 134 words — almost twice as large as that of legitimate articles. However, the analysis of Kumar et al. (2016) differs from ours in multiple ways. First, as already mentioned, they used a different, larger set of hoaxes collected as part of Wikipedia's regular NPP process. Second, they used a matching procedure to compare each hoax to only one legitimate article created on the same day. They also considered other types of articles, such as wrongly flagged articles and failed hoaxes. Another potential differentiating factor is the method of extraction for the plain text, markup content, and links for each page, which might contribute to not obtaining exactly the same results.

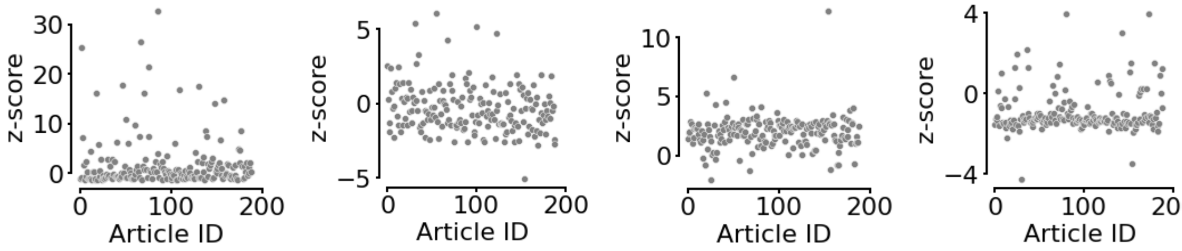
Figure 3b shows that hoaxes tend to have a similar density of wiki-links when compared to non-hoaxes. This is important, since to quantify online attention toward a topic we compute the volume of traffic to the wiki-link neighbors of an article. Thus, in the following analysis on traffic, we can safely exclude potential confounding factors due to different linking patterns between hoaxes and non-hoaxes.

Figures 3c and 3d show the distributions of the ratio of plain to markup text and of external link density, respectively. Aside from a few outliers, hoaxes almost always contain more plain text than markup text, compared with non-hoaxes. This is also consistent with the findings of (Kumar et al., 2016), who observed that, on average, 58% of a legitimate article, 71% of a successful hoax, and 92% of a failed hoax is just plain text.

In summary, hoaxes tend to have more plain text than legitimate articles and fewer links to external web pages outside of Wikipedia. This means that non-hoax articles, in general, contain more references to links residing outside Wikipedia. Such behavior is expected as a hoax's author would need to put a significant effort into crafting external resources at which the hoax can point.

4.2 Traffic Analysis

Recall that the cohort of a hoax is defined as all the non-hoax articles created on the same day it



(a) Plain text length (b) Wiki-link density (c) Plain to markup text ratio (d) External link density

Figure 3: Modified z-scores for all hoaxes in our sample relative to non-hoax articles in their cohorts for the four appearance features we considered. Hoaxes tend to have similar or slightly smaller count of plain text words (however with several higher-count outliers), lower external link density, higher plain to markup text ratio, and similar wiki-link density.

was created. To understand the nature of the relationship between the creation of hoaxes and the attention their respective topics receive, we first seek to quantify the relative volume change before and after this creation day. Here, a *topic* is defined as all of the (non-hoax) neighbors linked within the contents of an article i.e., its (non-hoax) out-links. Traffic to Wikipedia is known to fluctuate following circadian and weekly patterns, and is likely to depend on a host of additional, unknown factors, such as the relative popularity of Wikipedia over the years, the total number and geographic distribution of web users (Yasseri et al., 2012), etc. To account for these potential confounding factors, (Ciampaglia et al., 2015) proposed to quantify the volume change in a way that controls for the circadian rhythm and the fluctuating nature of traffic on the Web (Ciampaglia et al., 2015; Thompson et al., 1997). They have shown that studying traffic over a 14-day observation window, 7 days before and after the creation day, considers both short spikes in attention and weekly changes in traffic. The relative volume change is defined as:

$$\frac{\Delta V}{V} = \frac{V^{(b)} - V^{(a)}}{V^{(b)} + V^{(a)}} \tag{2}$$

where $V^{(b)}$ and $V^{(a)}$ are respectively the median traffic to neighbors in the 7 days before and after the creation of the article. According to eq. (2), $\Delta V/V > 0$ when the majority of traffic occurs before an article is created, i.e., attention toward the topic of the articles precedes its creation. When $\Delta V/V < 0$, attention tends instead to follow the creation of the hoax. Note that our traffic data covers a period spanning from December 2007 to August 2016. Since not all hoaxes in our dataset fell within that time frame, $\Delta V/V$ was calculated only for the 83 hoaxes (and their cohorts) whose creation dates fell within that period.

Figure 4 shows the distribution of the $\Delta V'/V'$ values for each cohort, the cohort mean, and the value of $\Delta V/V$ of the corresponding hoax, for a manually selected sample of hoaxes collected from our data.

Having defined a way to quantify whether traffic to a given article preceded or followed its creation, we want to determine whether hoaxes tend to have a greater $\Delta V/V$ than legitimate articles in general. Unfortunately, we know very little about the distribution of $\Delta V/V$ over multiple pages, and how it has changed over the course of the history of Wikipedia. However, if hoaxes do not differ from legitimate articles, then on average the difference the $\Delta V/V$ of a hoax and that of its cohorts should be zero. Therefore, we define:

$$D = \frac{\Delta V}{V} - \mathsf{E}\left[\frac{\Delta V'}{V'}\right] = \frac{\Delta V}{V} - \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta V'_i}{V'_i} \tag{3}$$

where $\mathsf{E}\left[\frac{\Delta V'}{V'}\right]$ indicates the expected $\Delta V'/V'$ of legitimate articles. Thus, when $D > 0$ a hoax accumulates more attention preceding its creation, compared to its cohort.

To test whether $D > 0$ holds in general, we estimate the mean value of D in our sample of hoaxes, and used bootstrapping to compute the confidence interval of the mean. To perform bootstrapping, we resampled the original list of D values $10,000$ times with replacement.

In general, we observe a trend in which hoaxes tend to have greater $\Delta V/V$ than their cohort: $D > 0$ in 75 out of 83 of the hoaxes in our data. The histogram in fig. 5 (left) shows the distribution of the differences, and shows that the mean is approxi-

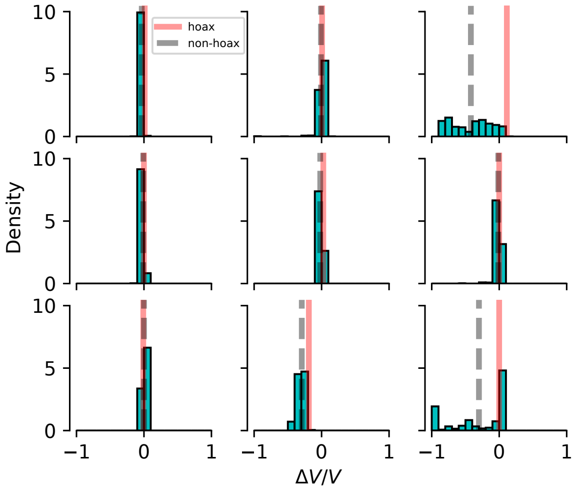


Figure 4: For a manual sample of hoaxes, the distribution of the $\Delta V'/V'$ values for each cohort (turquoise blue histograms) in comparison to the $\Delta V/V$ of the respective hoax (black dashed line). The $\Delta V/V$ of hoaxes tend to, in general, be higher than the mean of their cohorts (red solid line).

mately equal to 0.123, with a bootstrapped 95% confidence interval of $(0.1227, 0.1234)$.

According to the Central Limit Theorem (CLT), the distribution of sample means approximates the normal distribution with the increase of sample size n, regardless of the original distribution of data (Feller, 1991). Figure 5 (right) shows the distribution of the means for each of the $10,000$ resampled vectors. It is worth noting that all of the means returned were positive, implying a greater $\Delta V/V$ for the hoax.

5 Discussion and future work

Our study analyzes the role of online attention in the supply of disinformation on Wikipedia (Q1). From an operational point of view, we study the relationship between the creation of hoaxes and the traffic preceding each hoax's creation day d (Q2). To do so, we collected the view count of the out-link neighbors of the hoaxes and their cohorts for $d \pm 7$ days. Following prior work (Ciampaglia et al., 2015), to assess the allocation of attention during that period, we calculated the relative traffic volume change, which accounts for potential confounding factors due to traffic fluctuations. We observe that 90% of hoaxes have a higher $\Delta V/V$ than their respective cohort and confirmed it by means of resampling. This indicates that, on average, hoaxes tend to have more traffic accumulated before their creation than after. In summary, our observed D indicates that the generation of hoaxes in Wikipedia is associated with prior consumption of information, in the form of online attention, providing an answer to our original research question (Q1).

This study has some limitations that need to be acknowledged. First of all, our results are based on a list of only 83 hoaxes. Even though we originally collected a dataset that was twice the size of this one, we were limited by the fact that not all hoaxes were covered in our traffic dataset. Future work should extend our results to larger available samples of hoaxes (e.g., NPP-based) to ensure consistent results with prior work.

Additional limitations stem from our operational definition of the topic of a new article (hoax or non-hoax). In this work, we relied on outgoing hyperlinks (out-links) and neglected incoming hyperlinks (or in-links), owing to our lack of access to data on hyperlinks to hoaxes. This data is present but not publicly accessible in the Wikipedia database, presumably due internal regulations within the Wikipedia project. In the future, we would like to extend our analysis to include in-links as well.

Future work should also consider a more advanced definition of an article's topic that does not rely solely on hyperlinks, as they provide a very rough notion of topic. Links to very generic entries like locations or dates (e.g., 'United States of America' or '1492') typically convey little information about an article's topic .

Third, our traffic dataset is based on an older definition of pagecounts, which is affected by known limitations, including the presence of automated crawlers, which are known to generate large amounts of hits in a short period of time. To mitigate for the presence of this type of outliers, in our definitions of traffic volume we rely on the median instead of the mean, which is more robust to outliers. However, in the future we would like to include a more recent traffic dataset that is not affected by this and other biases (Wikimedia Foundation, Inc., 2022c).

In conclusion, our study sheds light on an important factor affecting the supply of disinformation on the Web. Future work should extend our results to venues other than Wikipedia, for example social media platforms like Facebook or Twitter. In addition, other types of media (like video, audio, etc.) should be considered — hoaxes do not

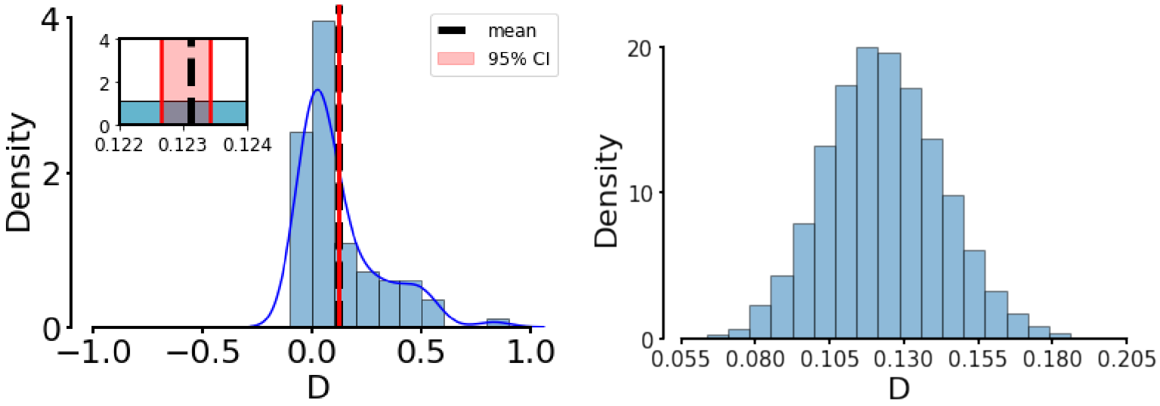


Figure 5: Left: Histogram of the relative traffic change differences D (see Equation (3)). The black dashed line is the sample mean, and the red area the 95% bootstrapped CI. The blue solid line is a kernel density estimate. The inset shows the sample mean relative to the confidence interval. Right: The sampling distribution of means obtained by bootstrapping $10,000$ samples with replacement.

only come in the form of textual articles, and attention is an effective incentive for people to keep spreading more disinformation, regardless of its medium. Future work should also consider studying the role of attention in versions of Wikipedia other than English. We expect similar trends to ones observed here to apply to non-English language editions as well. However, the signal may be weaker owing to lower traffic volume of non-English language editions. A comparative analysis of the role of attention in the supply of disinformation across cultures could shed more light about these type of threats to the Web as a whole.

References

5j9. 2022. Github inc., – wikitextparser. https://github.com/5j9/wikitextparser. Last accessed: 09-March-2022.

B. Thomas Adler, Luca de Alfaro, and Ian Pye. 2010. Detecting wikipedia vandalism using wikitrust - lab report for PAN at CLEF 2010. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, volume 1176 of *CEUR Workshop Proceedings*, page n.p., Aachen, Germany. CEUR-WS.org.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Marco Amoruso, Daniele Anello, Vincenzo Auletta, Raffaele Cerulli, Diodato Ferraioli, and Andrea Raiconi. 2020. Contrasting the spread of misinformation in online social networks. *Journal of Artificial Intelligence Research*, 69:847–879.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA. Association for Computing Machinery.

Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2015. The production of information in the attention economy. *Scientific Reports*, 5:9452.

Caitlin Dewey. 2015. The story behind Jar'Edo Wens, the longest-running hoax in Wikipedia. Last updated: 24-October-2018.

William Feller. 1991. *An Introduction to Probability Theory and Its Applications*, 2nd edition. John Wiley & Sons, Inc., New York, USA.

Ruth García-Gavilanes, Anders Mollgaard, Milena Tsvetkova, and Taha Yasseri. 2017. The memory remains: Understanding collective memory in the digital age. *Science Advances*, 3(4):e1602368.

R. Stuart Geiger and Aaron Halfaker. 2013. When the levee breaks: Without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym '13, New York, NY, USA. Association for Computing Machinery.

Michael H. Goldhaber. 1997. The attention economy and the net. *First Monday*, 2(4):n.p.

Yevgeniy Golovchenko, Cody Buntain, Gregory Eady, Megan A. Brown, and Joshua A. Tucker. 2020. Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. *The International Journal of Press/Politics*, 25(3):357–389.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1):eaau4586.

Andrew M. Guess, Brendan Nyhan, and Jason Reifler. 2020. Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, 4(5):472–480.

Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688.

Aaron Halfaker and John Riedl. 2012. Bots and cyborgs: Wikipedia's immune system. *IEEE Computer*, 45(3):79–82.

Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. 2011. Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 83–88, Portland, Oregon, USA. Association for Computational Linguistics.

Boris Iglewicz and David C. Hoaglin. 1993. How to detect and handle outliers. In Edward F. Mykytka, editor, *The ASQC Basic References in Quality Control: Statistical Techniques*, volume 16. ASQC, Milwaukee, WI, USA.

Gary King, Jennifer Pan, and Margaret E. Roberts. 2017. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501.

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 591–602, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

MediaWiki contributors. 2021. Extension:textextracts — mediawiki. `https://www.mediawiki.org/w/index.php?title=Extension:TextExtracts&oldid=4940004`. Last accessed: 9-March-2022.

MediaWiki contributors. 2022a. Api:main page — mediawiki. `https://www.mediawiki.org/w/index.php?title=API:Main_page&oldid=5019333`. Last accessed: 9-March-2022.

MediaWiki contributors. 2022b. Api:revisions — mediawiki. `https://www.mediawiki.org/w/index.php?title=API:Revisions&oldid=5037632`. Last accessed: 9-March-2022.

Mariella Moon. 2022. A Chinese Wikipedia editor spent years writing fake Russian medieval history. Last accessed: 13-Sep-2022.

Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12):2381–2403.

Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic vandalism detection in wikipedia. In *Advances in Information Retrieval*, pages 663–668, Berlin, Heidelberg. Springer Berlin Heidelberg.

Diego Sáez-Trumper. 2019. Online disinformation and the role of wikipedia. *CoRR*, abs/1910.12596.

Similarweb LTD. 2022. Top websites ranking. `https://www.similarweb.com/top-websites/`. Last accessed: Mar-09-2022.

Herbert A Simon. 1971. Designing organizations for an information-rich world. In Martin Greenberger, editor, *Computers, communications, and the public interest*, volume 72, pages 37–52. Johns Hopkins Press, Baltimore.

Koen Smets, Bart Goethals, and Brigitte Verdonk. 2008. Automatic vandalism detection in wikipedia: Towards a machine learning approach. In *Proceedings of the 2008 AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 43–48, Palo Alto, CA, USA. AAAI.

Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proc. of the International AAAI Conference on Web and Social Media*, pages 230–239, Palo Alto, CA, USA. AAAI.

K. Thompson, G.J. Miller, and R. Wilder. 1997. Wide-area internet traffic patterns and characteristics. *IEEE Network*, 11(6):10–23.

Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with History Flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 575–582, New York, NY, USA. Association for Computing Machinery.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: toward an interdisciplinary frameworkfor research and policy making. Technical Report DGI(2017)09, Council of Europe, Strasbourg, FR.

Wikimedia Foundation, Inc. 2022a. Help: Page name. https://meta.wikimedia.org/wiki/Help:Page_name. Last accessed: 09-March-2022.

Wikimedia Foundation, Inc. 2022b. Page view statistics for Wikimedia projects. https://dumps.wikimedia.org/other/pagecounts-raw/. Last accessed: 09-March-2022.

Wikimedia Foundation, Inc. 2022c. Research:page_view. https://meta.wikimedia.org/wiki/Research:Page_view. Last accessed: 13-September -2022.

Wikimedia Foundation, Inc. 2022d. Wikimedia Statistics – English Wikipedia. https://stats.wikimedia.org/#/en.wikipedia.org. Last accessed 09-March-2022.

Wikipedia contributors. 2021. Wikipedia:vandalism does not matter. https://en.wikipedia.org/wiki/Wikipedia:Vandalism_does_not_matter. Last accessed: Mar-09-2022.

Wikipedia contributors. 2022a. Wikipedia:list of hoaxes on wikipedia. https://en.wikipedia.org/wiki/Wikipedia:List_of_hoaxes_on_Wikipedia. Last accessed: Mar-09-2022.

Wikipedia contributors. 2022b. Wikipedia:new pages patrol. https://en.wikipedia.org/wiki/Wikipedia:New_pages_patrol. Last accessed: Mar-09-2022.

Wikipedia contributors. 2022c. Wikipedia:size of wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. Last accessed: Mar-09-2022.

Wikitech contributors. 2021. Portal:toolforge — wikitech. https://wikitech.wikimedia.org/w/index.php?title=Portal:Toolforge&oldid=1932575. Last accessed: 9-March-2022.

Taha Yasseri and Filippo Menczer. 2021. Can crowdsourcing rescue the social marketplace of ideas? Technical report, arXiv.

Taha Yasseri, Robert Sumi, and János Kertész. 2012. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PLOS ONE*, 7(1):1–8.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 218–226, New York, NY, USA. Association for Computing Machinery.

Ahmad Zareie and Rizos Sakellariou. 2021. Minimizing the spread of misinformation in online social networks: A survey. *Journal of Network and Computer Applications*, 186:103094.

“Gettr-ing” User Insights from the Social Network Gettr

Filipo Sharevski
DePaul University
243 S Wabash Ave
Chicago, IL
fsharevs@depaul.edu

Amy Devine
DePaul University
243 S Wabash Ave
Chicago, IL
adevine@depaul.edu

Emma Pieroni
DePaul University
243 S Wabash Ave
Chicago, IL
epieroni@depaul.edu

Peter Jachim
DePaul University
243 S Wabash Ave
Chicago, IL
pjachim@depaul.edu

Abstract

As yet another alternative social network, Gettr positions itself as the “marketplace of ideas” where users should expect the truth to emerge without any administrative censorship. To understand how this marketplace actually works, we interviewed 124 Gettr users and analyzed their views on how truth is meted out on the platform. While the right-leaning users joined Gettr as a result of a perceived freedom of speech infringement by the mainstream platforms, notably Twitter, the left-leaning users followed them in numbers as to “keep up with the misinformation.” The absence of restrictive actions like muting, removing, or labeling was pointed both as an enabler of free speech in practice (right-leaning) and as an enabler of bullying and harassment (left-leaning). We also found that almost every tenth user in our sample had a bad experience on a mainstream social media platform that forced them to seek refuge on Gettr.

1 Introduction

Promising a reprieve from banning and content moderation, fringe social networks are offering alternative social media experience to users disenchanted with participation on mainstream places like Twitter or Facebook. One such place is Gettr (a portmanteau of the words ‘Get Together’), brandishing an image of an alternative platform founded on “the principles of free speech, independent thought and rejecting political censorship and ‘cancel culture”’ (Gettr, 2022). This image is not new, as other alternative social networks like Parler offer(ed) users to “express openly, without fear of being deplatformed for their views” (Parler, 2022), Gab “champions free speech” (Gab, 2022), and 4chan allows “anyone to post comments and share images” (4chan, 2022). Therefore, a question arises about what novelty in particular Gettr brings for the fringe communities online.

Alternative social media platforms attract research curiosity with their lax moderation policies, palpable toxicity, and discourse ridden with polarizing and conspiracy narratives. 4chan, with its notorious, politically incorrect /pol board, receives considerable attention in analyzing trends of self-consciously offensive culture and meme virality on social media (Colley and Moore, 2022; Mittos et al., 2020; Zannettou et al., 2018b; Hine et al., 2017). Parler, infamous for providing “just enough” networking cohesion for the violent mob attack on the United States Capitol on 6 January 2021 (Munn, 2021), was empirically analyzed to reveal the patterns of amplification of its political pundits and the deliberate user experience design that inhibits a user’s ability to search for alternative political narratives (Peironi et al., 2021; Aliapoulios et al., 2021). Gab, branded as the “free speech” alternative to Twitter, was found to attract alt-right users, conspiracy theorists, and other trolls that disseminate hate speech on the platform much higher than Twitter, but lower than 4chan’s /pol board (Zannettou et al., 2018a; Lima et al., 2018). And an early look at Gettr (Paudel et al., 2021), showed yet another outlet for toxicity akin to Gab and 4chan, although yet to achieve the level of engagement and activity characteristic for the online fringe communities.

These content-focused looks further inspire an important line of inquiry following online extremism (Phadke and Mitra, 2021; Gaudette et al., 2021), ideological radicalization (Youngblood, 2020), hate speech (Mathew et al., 2020; Kennedy et al., 2020), and false information (Bleakley, 2021). However, the content-focus looks fail to shed a light on these platforms from the vantage point of the users themselves i.e. no analysis so far gathered users' insights from participation on any of these alternative platforms. Doing such a user study has nothing to do with legitimizing the platforms' existence or mission, but instead provides a meaningful context to the predominantly content-focused and data-driven investigations so far. Falling short of understanding the incentive structure for migrating to, and interacting on, these platforms, leaves a rather uninformed research conditions to conjectures about the future trajectories of the fringe communities online. Our study, to the best of our knowledge, is the first one that does a user experience assessment of the Gettr platform in response to this gap.

2 Content-focused Look on Gettr

Content-focused looks on Gettr are scarce to find as the platform maintains a highly restricted access to its content. There is, however, a GoGettr API, a third party client for scraping data that was created by the Stanford Internet Observatory (R. Miles McCain et al., 2022), but it does not allow for an up-to-date representative data retrieval. So far, the only paper that analyzes data on Gettr (Paudel et al., 2021) was submitted immediately before Gettr changed the indexing structure of the platform's posts, a limitation for future content-focused looks that was also acknowledged by the GoGettr API team, confirming no explanation or remediation offered by Gettr so far.

With a very limited data available up to August 9th 2021 thorugh the GoGettr API, we collected 6.8 million posts, 373,725 users, and 18,274,986 unique follower/followee relationships to have at least some rough understanding of how the content in structured on Gettr, following the preliminary platform scoping in (Paudel et al., 2021). The summary of the dataset is shown in Table 1.

	Count
Posts	6,814,244
Users	73,725
Followings	18,274,986

Table 1: Dataset Overview

Analyzing the timestamps of the posts as the closest alternative to an order indexing, we built a third-order univariate regression model (Seabold and Perktold, 2010) that captured 96.3% of the variability in the total number of posts over that period of time (r^2), with an adjusted r^2 of 95.9%. The model helped us to uncover a weekly cyclical pattern usage/participation on the platform, highlighted in Figure 1, showing a decrease in platform usage on Fridays, Saturdays, and Sundays.

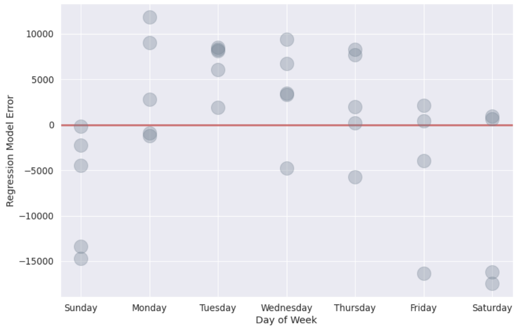


Figure 1: Residuals by day of the week.

Next to when users do post on Gettr, we look how users interact between each other. Due to the limitations of the Gettr API to distinguish between different types of follows, we decided to take a look at how often users mention one another, and how frequently two users mention one another. These relationships constitute some form of a "friendship," as defined in (Huberman et al., 2009), because they show a more deliberate effort between two people to support each other in a manner that is proactive. Previous work on Twitter examining these "friendships" has found a stronger positive correlation with a user's engagement on the platform than other relationships with other users like declared follower/followee relationships (Huberman et al., 2009).

In our case, due to the relative newness of the platform in the dataset that we obtained, we limited the number of mentions between two users to be at least one for them to be considered friends. This is simply because users have not had a chance to mention each other too much. We found 1,872 friendship relationships within our dataset, meaning that approximately 0.55% of users are in a friendship. These friendships naturally formed 592 disconnected subsets, of which the majority (500) had just 2 users. A summary of the friendship set is shown in Table 2. A corresponding graph of all friendships on the platform is plotted in Figure 2.

	Count
Friendship Subsets	592
Basic Friendship (2 users)	500
Total Friendships	1,872
Friendships / Total Users	0.55%

Table 2: Friendships Overview

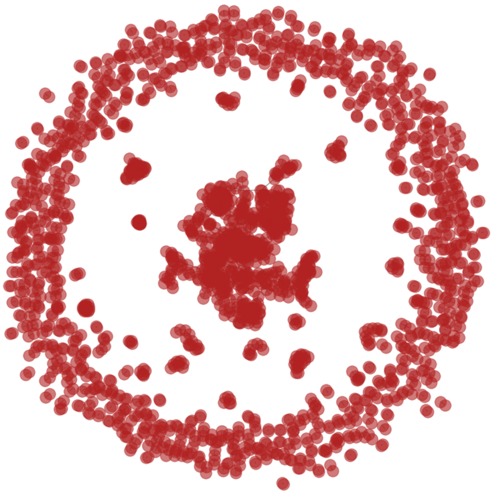


Figure 2: All friendships on the platform - users are represented in red dots, and their friendships are represented with black lines between them, or an overlap. The ring of friendships are smaller for subsets of two or three user friendships on the fringes, and the larger for the friendship groups in the center

As we examined the friendship subgraphs, we noted that while most friendship subgraphs only consisted of a couple of users, there were a couple that were substantially larger. The second largest friendship subgraph, depicted in Figure 3, consisted of a large number of prominent republicans (Matt Gaetz, Rudy Giuliani, Rick Santorum, Mike Pompeo, Sean Spicer), republican organizations (CPAC, Arizona GOP, the Heritage Foundation), and right-leaning political news organizations (Epoch Times, Steve Bannon's War Room, Newsmax, National Pulse). Together, this shows a public sphere of influence, where people in that subgraph mention one another, and there are clear ideological similarities in the narratives that they present to their audiences.

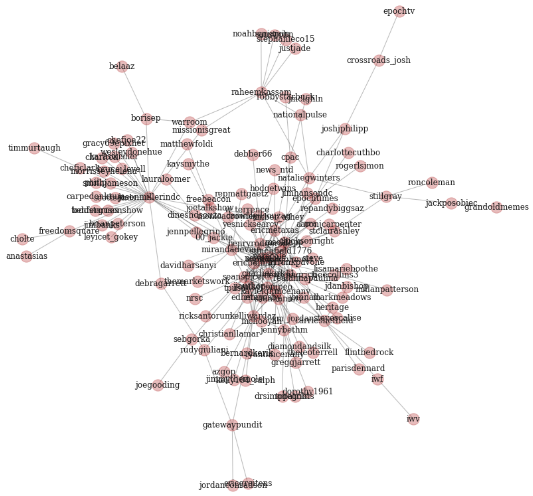


Figure 3: This is a friendship subgraph that shows a large number of prominent right-thought leaders, including prominent republicans, republican organizations, and right-leaning news organizations

3 Gettr: User Insights

A content-focused look is certainly useful in uncovering the macroscopic trends that shape the participation and the appeal of Gettr. However, a look into Gettr from a user perspective provides us with a context of platform participation on a more microscopic, granular level, an analysis absent from the research on alt-platforms so far. To our knowledge, this is a first study that brings the individual voices of the Gettr users to the fore and analyzes them in the context of the their participation in the online fringe communities (Munn, 2019). Such an analysis uncovers the "pipeline" through which users normalize and acclimate to the discourse on Gettr. To this objective we conducted a user study, approved by our Institutional Review Board (IRB) before any research activities began, where Gettr users were invited to expound on:

A. What is the value proposition they see in Gettr;

B. How they participate and express themselves on Gettr;

C. What kind of information they usually get from Gettr and how they consume it; and

D. Social media and Gettr experiences.

We set to sample a population that was 18 years or above old, a Gettr account holder, from the United States through the Amazon Mechanical Turk and Prolific. Both reputation and attention checks were included to prevent input from bots and poor responses. The user study asked open-ended questions and it took around 20-30 minutes to complete. Participants were compensated with the standard participation rate of $27.64/hr. The study was anonymous and allowed users to skip any question they were uncomfortable answering. We also we collected participants' political leanings, race/ethnicity, level of education, gender identity, and age.

After the consolidation and consistency checks, a total of 124 participants completed the study. To ensure consistency in the analysis and validity of the results, each of the open-ended responses in the survey was coded independently by three researchers. The codebook was simple and included a coding on the justification quotes from the participants on each of the four research questions above. The Fleiss's kappa $\kappa = 0.981$ for the 95% confidence, indicating an "excellent" inter-coder agreement overall.

The distribution of participants per their self-reported political leanings was: 29 left-leaning (23.38%), 42 moderate (33.87%), 40 right-leaning (32.25%), and 13 apolitical (10.5%). In respect to race/ethnicity, 91 identified as White (73.38%), 29 as Black or African American (9.67%), 13 as Asian (10.48%), 7 as Latinx (5.64%), and 1 as Native Hawaiian or Pacific Islander (0.83%). Education-wise, 22 of the participants had a high-school level (17.74%) , 84 some college or 2/4-year college (67.74%), and 19 had a gradate level of education (14.52%). Gender-wise, 47 of the participants were female (37.9%), 72 were male (58.06%), and 6 identified as non-binary (4.04%). Age-wise, 25 were in the 18 - 24 bracket (20.17%), 36 in the 25 - 34 (29.03%), 35 in the 35 - 44 (28.23%), 16 in the 45 - 54 (12.9%), and 12 in the 55 - 64 bracket (9.67%). The distribution of the sample is balanced on the political leanings, gender identity, and age, while skewed towards white and college-level educated participants.

3.1 Gettr's Value Proposition

Our results reveal several reasons why people joined Gettr. The most prevalent one is the users' objection to a perceived censorship and freedom of speech infringement by the mainstream platforms, notably Twitter [**P** denotes a participant and the number their index in the sample]:

[P44] *Gettr doesn't censor free thought and having been on Twitter for many years, I have seen the decline in the quality of the site. From suspending accounts to slapping false "fact check" labels on posts, Twitter has gone from a quality platform to a Gestapo site. Many of my friends have gone over to Gettr and love it, as do I. We aren't free if we can't express thoughts without fear of being banned.* [right-leaning]

[P38] *I don't like the politics of Twitter, and I would love to see an alternative take its place. I am tired of the censorship of conservative voices. I was further annoyed that Parler wasn't allowed to exist (although I am disappointed in Parler's lack of effort to find web hosting.)* [right-leaning]

[P7] *I decided to participate in Gettr because radical feminists are being banned from Twitter left and right just for stating basic facts. I have not been banned but friends of mine have and it's only a matter of time until I am as well.* [left-leaning]

[P52] *The main reason was free speech. A lot of posts on the other social platforms get censored if the discussion gets controversial or it does not fit the narrative of the general populace. It is always important to hear the viewpoints of everyone even if they do not align with yours and hopefully find a common platform where everyone has a voice.* [moderate]

[P32] *I was tired of censorship on other social media platforms. I want to hear different perspectives.* [apolitical]

Interestingly, the political leanings across the entire spectrum considerably factored in joining Gettr. The right-leaning and moderate participants were drawn to the freedom of speech value proposition of the platform, while the left-leaning joined Gettr in numbers mostly out of curiosity to see what the "opposition is saying:"

[P35] *Curiosity. I know that it prides itself on not censoring it's subscriber base, so wanted to see how racist//sexist/transphobic the posts were. I think knowing what the "other" side thinks, so as to better inform yourself of the issues and opinions and to know how to combat those ideas.*

[P20] *I wanted to see what the Trump supporters were up to in their new echo chamber.*

[P48] *Although I am quite liberal, I was interested to see what kind of information/misinformation was shared on Gettr. I think there's value in trying to understand the opposing views and trying to learn why they believe the things that they do.*

[P89] *I don't believe or trust anything on that site nor the users and politics that it promotes. I do though want to know what my enemy is up too. My main reason is to observe out of curiosity.*

Participants with some high school or high school graduate level were predominantly right-leaning and moderate, indicating that "*a lot of people [they] follow on twitter have been banned, so [they] joined Gettr to hopefully get uncensored and non biased information again.*" **[P71]**. The some/or 2/4-year college graduate participants reflect the aforementioned free-speech/spy-on-conservatives dichotomy as they comprised the largest part of our sample. The post-graduate educated participants, balanced on the political spectrum, provided justifications for what precisely pushed them towards Gettr, for example, "*Twitter recently they banned Babylon Bee, a satire account for their jokes, which a clear violation of first amendment where's Gettr believes in our constitutional rights.*" [**P53**]. Gender-wise, the conformity to the observed dichotomy of the Gettr's appeal is also preserved:

[P75] *To try something new for a change to share my ideas about free speech, since like using other platforms such as Twitter to express said thoughts, was a lot more difficult and hate-arousing.* [male, right-leaning]

[P12] *Curiosity to see what the far-right bubble had to say about things* [male, left-leaning]

[P37] *It seems to be a better source of information than Twitter* [female, right-leaning]

[P89] *I wanted to see what conservatives are posting since i feel like most liberals like myself aren't aware* [female, left-leaning]

[P101] *Some people I follow on twitter created accounts there, so I wanted to see their updates* [non-binary, moderate]

[P28] *I screenshot people saying weird like far-right things and make fun of them with my friends* [non-binary, left-leaning]

Age-wise, the participants in the [18 - 24] bracket were mostly joining Gettr for making connections and "*trying new things, not a fan of the modern social media giants and their censorship*" [**P90**]. The [25 - 34] participants added more context on to the "trying something new" premise, stating that "*other sites are stale; You deal with the same lack of trust of them and want to get away from all the negativity*" **[P108]**. The [35 - 44] participant got even more concrete and stated they joined Gettr "*so I can talk and converse that have same political beliefs as me and not be ostracized by everyone*" [**P59**]. The [45 - 54] participants contextualized this stance by doubling down on the "*free speech, against the massive push for cancel culture on social media*" **[P106]**. The [55 - 64] participants added the differentiation niche with Twitter, as they were "*was interested to see if indeed Gettr would be better than Twitter*". **[P83]**.

Participants also pointed out that they came to Gettr because "*some friends of their friends did*" [**P81**] and several noted that "*many of the conservative talk show personalities that they listen to are on Gettr and they recommended it*" [**P122**]. One of the participants found the platform design and features appealing: "*I enjoy the user interface of Gettr as well as it's more approachable community. As a moderate conservative, I fit in well with almost everyone I have encountered on the platform. Gettr also gives the option to link it to an individual's Twitter.*" [**P99**]. The informative value of the content on Gettr was also appealing as "*it seems to be a better source of information than Twitter*" [**P37**] where users can "*stay up to date with news*" [**P54**].

3.2 Participation and Expression on Gettr

Around 27.5% of all the participants indicated they participate through writing original posts and commenting/liking on other's user posts. Either commenting (29%) or liking (21.8%) on other's user posts was how roughly half the participants spent their time on Gettr. The remaining 21.5% of the participants indicated that they are "*mostly just browsing and exploring the platform*" [**P29**]. When asked what motivates a user to participate on Gettr, the right-leaning and apolitical participants cited the appeal of a "personally involving discourse" while the moderates and left-leaning cited leaned more towards a "constructive discourse":

[P27] *If I have a strong opinion on it or get emotionally worked up.* [right-leaning]

[P29] *I do not fully agree with their conservative agenda, but being on the site and involved in the commentary helps me see the "other side" as it was.* [moderate]

[P86] *I am interested in learning more about peoples opinions on many issues, especially political issues that are different than my own.* [left-leaning]

[P43] *If I see interesting topics that resonate with me.* [apolitical]

We queried the participants about how Gettr's self-proclaimed "free speech" image facilitates their expression on the platform. The right-leaning participants stated that Gettr: "*doesn't ban people for saying things that big government and the elites might not like*" [**P22**], "*seeks to eliminate the 'cancel culture'*" [**P33**], "*Allows free exchange of information that is factual, despite what Twitter says*" [**P44**], and "*allows a person who believes in Donald Trump to be able to express those views and not be censored*" [**P59**]. The moderates explicitly highlighted a comparison to Twitter as being "*strict about removing content that they deem is not correct - even if I don't agree with what someone is saying, I would prefer to see the content instead of having it removed*" [**P97**]. The left-leaning participants didn't miss to point out that "*Gettr prides itself on freedom of speech, but obviously it doesn't condone bullying, harassment, threatening behavior, etc; So far concerning 'ideas', Gettr is pretty lax and unobtrusive* [**P35**]. They also pointed out the UX support for freedom of speech, as "*there is no muting or removing*" on the platform [**P36**].

In regards to the way of expression, we asked the participants if they use textual content only, emojis, and/or memes (or combination of). The right leaning participants preferred textual expression as not to " *'hijack the post' with multimedia* [**P38**] or to "*voice different perspectives that the news does not do a good enough job of covering*" [**P23**]. The moderates opted for textual expression because "*it's faster and easier and allows you to be as clear and specific as possible*" [**P91**]. Although many liberal-leaning participants come to Gettr to "*lurk and spy on conservatives*" [**P14**], those who actively participate "*always try to be respectful and still get their point across*" [**P34**] as well as use "*specific wording to balance their views, but not agitate, and shut down discussions* [**P48**].

The right-leaning participants utilized emojis mostly to accentuate a point, e.g. "*use emojis that correspond with the emotion of my reaction to a post, for example, if I am angry about what is said in a post, I will use the anger emoji* **[P47]**. The moderates utilized emojis *just to show appreciation for other's posts* **[P120]**. The left-leaning participants avoided using emojis, arguing that "*emojis give people a way out, and if I engage, I am generally looking for answers or at least understanding of why certain views are held*" **[P34]**. When it comes to memes, the left-leaning and apolitical participants strongly avoided using them, while the moderates used them "*usually as joke/humor*" **[P106]**. Fun was also the most cited reason for using memes by the right-leaning participants, as "*memes are funny in a political way; I post memes that poke fun at liberals.*" **[P122]**.

3.3 Information Consumption on Gettr

Regarding information consumption, we first asked the participants if there is any information they get exclusively on Gettr and nowhere else on social media. Participants singled out:

[P22] *News involving COVID vaccines and statements*

[P53] *Babylon Bee, James O'Keefe, and Project Veritas content and commentary*

[P59] *The MAGA movement information and the ways to de-certify the election*

[P19] *Opinions on minor Republican primary candidates*

We also asked if the participants have compared information between other social networks and Gettr. From the mainstream platforms, expectedly, Twitter was the most sought after place for information comparison across the political spectrum. The left-leaning and apolitical participants mostly avoided the alternative platforms (less then 8% in both groups have looked outside of Twitter, Facebook, Instagram or Reddit), while the moderates had a preference for comparison with 4chan. The right-leaning participants were equally interested in comparing Gettr information with Parler, Truth Social, Gab, and 4chan (23.35%) while maintaining the main focus on the mainstream social networks (76.65%).

We noticed that male participants have a much stronger preference for comparing information between Gettr on one side, and Reddit, Truth Social, Gab, and 4chan on the other side (35.78%). The female and non-binary participants mostly turned to the mainstream social networks for information comparison (only 14% and none, respectively, had looked at any alternative community). The participants with some/high school degree mostly preferred Twitter and Reddit, but also turned to the alternative platforms. The some/college graduates participants had equal preference within the mainstream and alternative platform groups, though much in favor of the Twitter, Facebook, and Reddit (72.64%). The post-graduate participants mostly sought comparison of information on the mainstream platforms (74.45%). Age-wise, the trend remains similar to the above, with the less interest for the alternative social networks among the young and more senior participants (only 7% and 4.5%, respectively, had looked at any of these communities).

3.4 Social Media and Gettr Experiences

We asked the participants if they had had a bad experience on Gettr or another another social network. We did so because our content analysis revealed that 620 unique users mentioned that they were subject of personal harassment in their profile descriptions (a similar trend was observed on Parler where the "banning" was used as a badge of honor (Peironi et al., 2021)). Citing personal harassment, 10.48% of the participants reported a bad experience on Twitter, Facebook, Instagram, and Reddit.

[P123] *I've gotten some pretty nasty remarks from people that I don't know on Twitter over content that I did not consider controversial* [right-leaning]

[P29] *I've been shamed as a bigot for doing nothing wrong on Twitter before. I'm not even conservative!* [moderate]

[P14] *Plenty. Discourse with Conservatives on platforms such as Facebook quickly devolves into name-calling and personal attacks* [left-leaning]

Our content analysis revealed that 496 unique users mentioned that they got banned from the mainstream social networks into their profile descriptions: 13 of those users explicitly used the term "Facebook Jail" badge of honor in their profiles, 32 people self-described themselves as "shadow-banned", citing bans and content moderation, 11.29% of the participants complained about bad experience on the mainstream platforms:

[P21] *I got put in "Facebook Jail" a few times for information and thoughts I posted about COVID, which turned me off on using Facebook* [right-leaning]

[P30] *Yeah, Facebook and Twitter would put stupid warnings on posts about stuff I posted, but I was actually right* [moderate]

[P85] *I was banned on Facebook for speaking the truth about the #metoo* [left-leaning]

We also asked the participants about what would make them consider leaving Gettr. Interestingly, the dichotomy we observed throughout the analysis so far is somewhat reversed: the left-leaning participants would not leave Gettr for anything while the moderates and right-leaning participants would seek other platforms if censorship/bans crept in on Gettr or users start abandoning it. Drawing on the experience of Parler, some participants expresses worries about Gettr being "*shut down by Apple or Google*" **[P44]**. An introduction of "*overwhelmingly biased recommendation algorithm like Twitter*" **[P68]**, "*changes in the user interface*" **[P79]** and introduction of "*content moderation*" **[P97]** were also mentioned several times besides the main points of free-speech/opposition credo:

[P4] *Censorship is the only thing that would cause me to leave. If Gettr ever gets to be like Twitter or Facebook by censoring important stories I will leave to find something better* [right-leaning]

[P27] *If Gettr started banning people and limiting freedom of expression I would probably leave it* [moderate]

[P14] *Nothing, as I only use Gettr for conflict, and to understand the Conservative zeitgeist* [left-leaning]

4 Discussion

4.1 User Experience on Gettr

The deep insights from Gettr reveal that users on the platform 'get *asymmetrically* together' usually during the work week. Perhaps the weekends are left for the other platforms, as virtually everyone we interviewed confirmed they regularly use other social networks, most notably Twitter and Reddit. The asymmetric nature of the discourse, characteristic for the fringe platforms, is also confirmed with our findings that suggest a similar centrality towards right-thought "celebrities" as is the case on Parler (Peironi et al., 2021) and Gab (Zannettou et al., 2018a).

We found evidence that people do bring their identity to the fore on Gettr and virtually in all cases the participation in the discourse encompasses expression of their political attitudes. The ones being moderate or leaning right on the political spectrum cited, in no uncertain terms, their disenchantment of Twitter's banning and moderation as the reason why they joined Gettr. Within our sample, more than 10% of the people were actually being banned from the mainstream platforms, a rather large number for a relatively small number of users reviewed. The ones leaning left, followed in considerable numbers because they saw a value in "seeing what misinformation the opposition is spreading around."

When it comes to (mis)information, the self-reported topics on Gettr were related to the COVID-19 vaccines and statements, commentary of right-thought pundits and narratives like Project Veritas, MAGA movement information, and republican candidates debates. Interestingly, very few of the participants directly mentioned the topics we identified in the cluster as ones they are interested in or discuss on Gettr. Overall, none of the participants were ready to leave the platform but indicated that it is a possibility if Gettr decides to implement "censorship". The moderate and right-leaning participants were also concerned about Gettr being shut down by Apple or Google like Parler was and the left-leaning participants were concerned about where else they can find a similar source of "conservative zeitgeist."

From our results it appears that so far users are satisfied with the value proposition of Gettr. We even got far-left participants that found a refuge on Gettr after being banned on Twitter, suggesting that Gettr—at least for now—works to maintain an image they believe is a "free-speech keeper" (Sharevski et al., 2021). Some of the participants expressed that by using this posture Gettr does not condone bullying, harassment, and threatening behaviour, which is an issue that Gettr might face soon or later since the "marketplace of ideas" metaphor does not have a static meaning in the broader legal interpretation of the First Amendment online (Schroeder, 2018).

4.2 Limitations and Future Work

In regards our user study, we were limited in obtaining a larger sample as it was difficult to reach the Gettr user population and many Gettr users declined to participate. Future research, if Gettr is still present in the social network space, could attempt to replicate this study with a larger user population in order to provide a basis for a more insightful quantitative analysis of the user experience. We also took a very cautious phenomenological approach in surveying the user experience on Gettr. The sample citations included each of the responses to the four research questions were selected by a unanimous agreement between the three independent coders/researchers, however, this does not preclude the analysis to be affected by some degree of selection bias. We acknowledge that other group of coders/researchers could select citations different than the once we selected, though we believe the end result of the analysis will remain effectively the same.

In our future work, we plan to expose Gettr users to particular content or findings about the platform and obtain their opinions as to capture how Gettr maintains (or not) the principles of free speech and independent thought. Another line of inquire following our study is the diffusion of information between Twitter, Reddit, and other alt-platforms that aims to track how the right-thought "celebrities" manage to maintain their active presence across the social network sphere as we did not delve deeper to analyze the content of the most popular accounts on Gettr.

5 Conclusion

Gettr, at the current stage, appears more as a pathway rather than a fully formed "pipeline" through which the disenchantment of the so-called Big-Tech social networks shapes the alternative agenda. With a simple user interface, it does allow acclimation for users perceived as outcasts from the mainstream social media discourse on both sides of the political spectrum and not just on the right. However, we had to uncover this evidence only through user interviews as the Gettr administrators inhibited any effort to objectively determine the larger discourse trends on the platform. We hope our attempt to look deep inside the platform is the first step towards uncovering the "truth" about the inner workings of Gettr, which, eventually will emerge.

References

4chan. 2022. 4chan: What is 4chan?

Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. An early look at the parler online social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 943–951. AAAI.

Paul Bleakley. 2021. Panic, pizza and mainstreaming the alt-right: A social media analysis of pizzagate and the rise of the qanon conspiracy. *Current Sociology*, 0(0):00113921211034896.

Thomas Colley and Martin Moore. 2022. The challenges of studying 4chan and the alt-right: 'come on in the water's fine'. *New Media & Society*, 24(1):5–30.

Gab. 2022. Gab: A social network that champions free speech, individual liberty and the free flow of information online.

Tiana Gaudette, Ryan Scrivens, Garth Davies, and Richard Frank. 2021. Upvoting extremism: Collective identity formation and the extreme right on reddit. *New Media & Society*, 23(12):3491–3508.

Gettr. 2022. Gettr: The marketplace of ideas.

Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):92–101.

Bernardo Huberman, Daniel M. Romero, and Fang Wu. 2009. Social networks that matter: Twitter under the microscope. *First Monday*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*.

Lucas Lima, Julio C.S. Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 515–522.

Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2020. "and we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):452–463.

Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online. *First Monday*.

Luke Munn. 2021. More than a mob: Parler as preparatory media for the u.s. capitol storming. *First Monday*, 26(3).

Parler. 2022. Parler: The premier global free speech platform.

Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. An early look at the gettr social network. *arXiv preprint arXiv:2108.05876*.

Emma Peironi, Peter Jachim, Nathaniel Jachim, and Filipo Sharevski. 2021. Parlermonium: A data-driven ux design evaluation of the parler platform. In *Critical Thinking in the Age of Misinformation CHI 2021*.

Shruti Phadke and Tanushree Mitra. 2021. Educators, solicitors, flamers, motivators, sympathizers: Characterizing roles in online extremist movements. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

R. Miles McCain, Konrad Iturbe, Ashwin Ramaswami, and Alex Turk. 2022. GoGettr.

Jared Schroeder. 2018. Toward a discursive marketplace of ideas: Reimaging the marketplace metaphor in the era of social media, fake news, and artificial intelligence. *First Amendment Studies*, 52(1-2):38–60.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Filipo Sharevski, Peter Jachim, Emma Pieroni, and Nate Jachim. 2021. Voxpop: An experimental social media platform for calibrated (mis)information discourse. In *New Security Paradigms Workshop*, NSPW '21, page 88–107, New York, NY, USA. Association for Computing Machinery.

Mason Youngblood. 2020. Extremist ideology as a complex contagion: the spread of far-right radicalization in the united states between 2005 and 2017. *Humanities and Social Sciences Communications*, 7(1):1–10.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018a. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1007–1014, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018b. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA. Association for Computing Machinery.

Fact-checking Multidimensional Statistic Claims in French

Oana Balalau, Simon Ebel, Théo Galizzi,
Ioana Manolescu, Quentin Massonnat
Inria and Institut Polytechnique de Paris
1 rue Estienne Honoré d'Orves,
91120 Palaiseau, France
firstname.lastname@inria.fr

Antoine Deiana, Emilie Gautreau, Antoine Krempf,
Thomas Pontillon, Gérald Roux, Joanna Yakin,
FranceInfo, Radio France
116 Av. du Président Kennedy,
75016 Paris, France
firstname.lastname@radiofrance.com

Abstract

To strengthen public trust and counter disinformation, *computational fact-checking*, leveraging digital data sources, attracts interest from the journalists and the computer science community. A particular class of interesting data sources comprises *statistics*, that is, numerical data compiled mostly by governments, administrations, and international organizations. Statistics are often *multidimensional datasets*, where multiple dimensions characterize one value, and the dimensions may be organized in hierarchies.

This paper describes StatCheck, a statistic fact-checking system jointly developed by the authors, which are either computer science researchers or fact-checking journalists working for a French-language media with a daily audience of more than 15 millions (aud, 2022). The technical novelty of StatCheck is twofold: (i) we focus on multidimensional, complex-structure statistics, which have received little attention so far, despite their practical importance; and (ii) novel statistical claim extraction modules for French, an area where few resources exist. We validate the efficiency and quality of our system on large statistic datasets (hundreds of millions of facts), including the complete INSEE (French) and Eurostat (European Union) datasets, as well as French presidential election debates.

1 Introduction

Professional journalism work has always involved verifying information with the help of trusted sources. In recent years, the proliferation of media in which public figures make statements, in particular online, has led to an explosion in the amount of content that may need to be verified to distinguish accurate from inaccurate, and even potentially dangerous, information.

To help journalists deal with the deluge of information, computational fact-checking (Cazalens et al., 2018; Nakov et al., 2021) emerges as a growing, multidisciplinary field. The main tasks of a fact-checking system are: *identifying the claims* made in an input document, *finding the relevant evidence* from a reference corpus, and (optionally) *producing an automated verdict* (is the claim true or false?). A reference corpus can be a knowledge graph (Ciampaglia et al., 2015), Web sources such as Wikipedia (Nie et al., 2019; Yoneda et al., 2018), or relational tables (Chen et al., 2020; Herzig et al., 2020; Jo et al., 2019; Karagiannis et al., 2020).

For fact-checks to be convincing, professional journalists prefer reference sources of high quality, carefully built by specialists. These include **statistics** produced and shared by governmental and international organizations, such as INSEE, the French national statistics institute [1] and Eurostat, the equivalent European Union office [2]. Technically speaking, such statistics are *multidimensional tables*, where a *fact* is a number, characterized by one or more a *dimensions*, such as a geographical unit, time interval, and other categories such as "Education level", etc. Unfortunately, such data sources are significantly more complex than relational tables, making their usage challenging. Consequently, despite the interest in such sources, only a few works have used them for automatic fact-checking (Cao et al., 2018; Duc Cao et al., 2019).

In our collaboration between computer scientists and fact-checking journalists, we have developed, deployed, and continue to be extending StatCheck, a fact-checking system specialized in the French media arena. StatCheck builds

[1] https://www.insee.fr

[2] https://https://ec.europa.eu/eurostat

upon the open-source code base of (Cao et al., 2018; Duc Cao et al., 2019). We significantly improved its data ingestion speed and more than doubled its statistic corpus by adding Eurostat data. Different from (Chen et al., 2020; Herzig et al., 2020; Jo et al., 2019; Karagiannis et al., 2020; Ciampaglia et al., 2015; Nie et al., 2019; Yoneda et al., 2018; Aly et al., 2021), StatCheck also includes a claim detection step, which saves journalists' time by focusing their attention on the claims worth checking; our claim detection module significantly outperforms the only one we know of for French (Duc Cao et al., 2019).

Outline. Below, we start by presenting a set of functional requirements derived from the journalist authors' experience in Section 2. Next, we describe the actual organization of statistic databases, and the StatCheck architecture, in Section 3. Then, we explain how this architecture is instantiated over two different sources, INSEE and Eurostat, whose size and organization significantly vary, in Section 4; we ingest and index all the data to support efficient search over it (Section 5). Finally, our claim detection modules are described in Section 6, then we conclude.

2 Fact-Checking Work Routine and Requirements

The journalist authors are part of the same team, specializing in fact-finding and fact-checking in a French-speaking national media. The material they author is disseminated through both the native and online media channels of their news organization. Their work is split among the two main classes identified in (Juneja and Mitra, 2022): *short-term claim centric*, focusing on the veracity of statements made continuously by public figures, which need to be checked relatively quickly; and *long-term issue-centric*, whereas individual journalists maintain and increase their knowledge of application topics, such as "law enforcement", "education and research", "defense", etc.

The short-term, claim-centric work raises several requirements. First, journalists know whose claims might interest their audience. Thus, they need an *interesting subset (selection) of social media content* to be made available through a Web platform. Journalists specify a set of social media account handles (currently Twitter and Facebook), and need the ability to modify this set themselves, as people gravitate in and out of the public's attention. Second, whenever claims about statistic entities are made in this social media content sphere, *bringing these claims to their attention*, isolating them from the mass of social communication of the figures they follow, saves journalists time and effort. Third, as previously noted in (Cazalens et al., 2018; Saeed and Papotti, 2021), data sources relevant to a given claim must be quickly identified and as precisely as possible. This again saves journalists time to search statistic data sources that may be very large, i.e., Eurostat publishes thousands of datasets, some with millions of rows.

The long-term, issue-centric work also benefits from these functionalities, yet it is more open; journalists may peruse claims for which they have not identified relevant sources yet, but still *appreciate a recommendation of most likely check-worthy claims*. User-friendly means to filter messages considered check-worthy (Should messages about the future, such as electoral promises, be considered, or not? Is a number required in a statistical claim, or not?) are also appreciated.

Common to both kinds of work, the newsroom involved in this project has the core tenet that *any verdict or judgment must be vetted by journalists*, since publishing it engages their professional responsibility. This has a set of consequences. (i) Journalists need to analyze the facts relevant to a claim and interpret them in a nuanced way for their audience. For instance, a difference of 5% between a number stated in a claim, and the value in a reference source, may be negligible or, on the contrary, a serious attempt to mislead, depending on the context. Thus, unlike prior systems (Chen et al., 2020; Herzig et al., 2020; Jo et al., 2019; Karagiannis et al., 2020), StatCheck does not compute a "true/false" verdict, leaving this tasks to journalists. (ii) For transparency and trust, links to any fact on its original publishing site must be provided together in the fact-check.

3 Fact-checking Based on Multidimensional Statistics

A multidimensional dataset consists of a set of *facts*, each having one *value* along a set of *dimensions*. For instance, Figure 1 (top) represents a three-dimensional dataset: French departments are on the horizontal axis, education levels on the vertical axis, while years are on the third (depth) axis. In each cell, the dataset stores the number of students in the respective department, level of

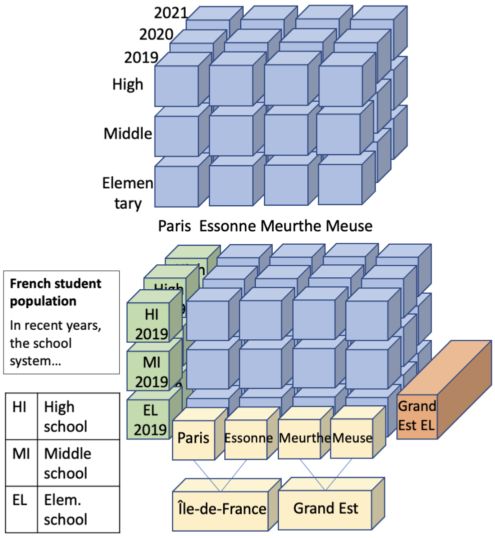


Figure 1: Multidimensional statistic data: conceptual view (top), structure of actual published dataset (bottom).

study, and year. In practice, actual Open Data statistics published by the government or international organizations are typically much more complex, as shown at the bottom of Figure 1. First, to save space, *dimension values may be encoded* into short codes, e.g., ”HI” for ”High school”, ”MI” for ”Middle school”, etc.; a decoding dictionary, associating a human-understandable term to each code, is published with, or close to the data cells. Although not shown in the figure, *dimension names* are similarly encoded. Second, *header cells*, shown in yellow and green in the figure, *may be mixed with data cells*; this requires effort to interpret them correctly. Note also that *there can be a hierarchy of headers*, e.g., a dataset at the granularity of departments may also include region names, e.g., ”Île-de-France” and ”Grand Est”, placed in the data files above, or close to, the region header cells. Third, datasets may also contain *partially aggregated results*, illustrated by the orange box holding the sum of all facts for one region (Grand Est), one education level (elementary), and the three years. Fourth, for each dataset, there may exist a separate, textual *description*, which contains a title, e.g., ”French student population”, and other comments.

Data representation in files. In practice, a multidimensional statistic dataset is published as a file, which can be CSV, a spreadsheet etc. For that, it is laid out in a bidimensional format, with some facts on each line, and as many lines as needed. If the data has more than two dimensions, which is often the case, this leads to *row header cells encoding several dimensions and their values*, such as ”HI 2019”, ”MI 2019” etc. in the figure. The file may start with the column headers (yellow), then the encoded multidimensional row header cell ”EL 2019” followed by the four cells corresponding to it, then a similar line for ”MI 2019”, a line for ”HI 2019”, followed by similar lines for 2020, then 2021 etc. Partially aggregated results are interspersed between such lines.

Challenges. To exploit such datasets for fact-checking, a set of challenges must be addressed. The useful information, e.g., ”How many elementary school students were in Île-de-France region in 2019?”, is a number in a cell. To find such information, we must **identify and store its relationships with human-understandable descriptions of its dimensions**, such as ”Education level: Elementary school”. In this example, the question is asked at a granularity (region) that is more coarse than the granularity of the data. To find the answer, we must exploit the fact that Paris and Essonne are departments in the Île-de-France region. Further, statistic claims may use similar but different language, e.g., a claim may be made about ”pupils in Île-de-France”. **Linguistic knowledge must be leveraged** to connect the claim terminology with that of the dataset. As mentioned in our requirements (Section 2), **fine-granularity answers are preferred**, that is: if the answer consists of one or a few cells only, those should be extracted from the dataset and returned, to avoid journalists' efforts to search in potentially large files. Finally, **speed at scale is important**, to enable journalists to work efficiently.

Architecture. To address these challenges, based on the requirements described in Section 2, we have devised an architecture shown in Figure 2. The modules in the lower row acquire and process reference datasets (Section 4), e.g., statistics about education in France. Those in the upper row acquire content to be fact-checked, e.g., a tweet stating: ”*More teachers are needed to educate 200K pupils in Île-de-France!*”, extract claims (Section 6), in this case ”*200K pupils in Île-de-France*”, and identify the most relevant facts for checking these claims, by searching the appropriately indexed reference datasets (Section 5).

4 Statistic Fact Database and Storage

By crawling, we acquired the complete INSEE and Eurostat statistics, and store them as follows.

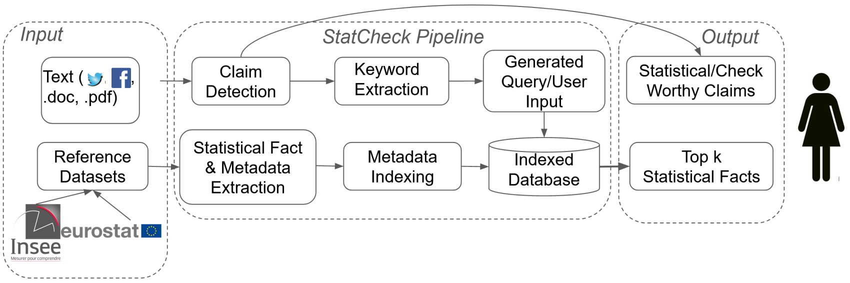


Figure 2: StatCheck architecture overview.

INSEE publishes each statistic report as an HTML page containing a description (title and comments on the data), and statistic tables in Excel or in HTML. As of May 2022, there are 60,002 Excel files (each of which may contain several tables) and 58,849 HTML tables. *The table organization varies significantly across the datasets*; nested headers are frequent. The largest table has 50.885 lines. Following (Cao et al., 2017), to capture all the elements of an INSEE dataset, we turn it in an **RDF graph** (www-rdf), where each data cell, header cell, and partial aggregate becomes an RDF node (URI). Each data cell or partial aggregate node is connected, through an RDF triple, to the cells corresponding to its closest header cells. Thus, the number of elementary school students in Paris in 2019 is connected to header cells labeled ”Paris”, respectively, ”Elementary school 2019” (where ”EL 2019” was decoded using the dictionary). Finally, each header cell is connected through an RDF triple to its parent header cell. This allows us to easily find out that the elementary school students in Paris in 2019 are also to be counted as being in the Île-de-France region. We also create an RDF node per dataset, which is connected to all its header cells and to the textual title and comments (each modeled as an RDF literal). The INSEE corpus lead to **7,362,538,629 RDF triples**, including **22,366,376 header cells**. We store them in the Jena Fuseki server with the TDB2 persistent back-end (www-tdb2).

Eurostat publishes 6,803 statistic tables, ranging from 2 lines to 37 million lines, and 580 dictionaries that, together, decode 243,083 statistical concepts codes into natural-language descriptions, all of which we acquired in StatCheck' database. Together, the Eurostat data files total **414.908.786 lines**. In Eurostat, *dimension hierarchies are described in the dictionaries*; we store these in memory. The statistic tables are simple-structure TSV files, thus, storing each of them as a table in a relational database was an option. However, their number is relatively high, and storing a file in a database inevitably increases its storage footprint. Therefore, to keep the data more compact, in view also of future extensions of our platform with more statistics from the World Health Organization, World Development Index etc., we store them as plain files, complemented by specialized indexes, as we explain below.

5 Statistic Search

Given a keyword query $Q = \{k_1, k_2, \ldots, k_n\}$, such as ”middle school pupils in Île-de-France in 2020”, the task we consider here is to find:

- the most relevant facts from our complete INSEE and Eurostat corpus;
- or, if a concrete fact is not found, but some datasets as a whole appear related to the query, return those datasets.

There may be several fact- (or cell-) level as well as dataset-level answers; we return a ranked list based on their relevance.

We call **metadata of a statistic dataset** all the natural-language elements that are part of or associated with the dataset: its title, comments, and human-understandable versions of all its header values. We use $\mathcal{L} = \{T, C, H\}$ to denote the set of the **locations** in which a term can appear in metadata, respectively: the dataset title, a comment, or a header. The locations are important since a term appearing in a title is more significant than one appearing in a header, and we exploit this when retrieving the datasets most relevant for a query (Section 5.1). Also, locations help determine whether a dataset matches some keywords headers of different dimensions - in which case the cell(s) at the intersection of those dimensions likely have a very relevant result (Section 5.2).

5.1 Dataset Indexing and Search

We split the metadata of each dataset d into a set of tokens $T = \{t_1, \ldots, t_N\}$, and remove stop words. For each token t, we identify based on a Word2Vec model the 50 tokens t' closest semantically to t. Next, for each appearance of a token t in a location l within d, our **term-location index** I_{TL} stores: (i) the index entry (t, d, l) corresponding to the token actually found in d; and (ii) 50 entries of the form $(t', d, l, dist)$, for the 50 tokens closest to t. These extra entries enable answering queries on terms close to (but disjoint from) the dataset content. For instance, when t is "*school*", t' could be "*teacher*", "*pupil*", "*student*", etc. For fast access, I_{TL} is hosted in the Redis in-memory key-value store (www-redis). To find the datasets relevant for the query Q, we look up the query keywords in I_{TL}, and consider relevant any dataset associated with at least one keyword.

The above indexing mechanism leverages word distances. Separately, we used *geographic resources*, in particular (Eurostat, 2022) for EU locations, to make our system aware of the relationships between geographic units (cities, departments, regions) across Europe. This ensures that a dataset is considered relevant if it mentions a geographic unit that includes or is included in the query. It is important to identify geographic names in the metadata. We have adopted the FlashText algorithm (Singh, 2017), capable of finding, in a dataset metadata of size N, one of M fixed keywords in $O(N)$ time complexity. This is much faster than the $O(NM)$ cost of regular expression pattern matching used in the previous system (Cao et al., 2018) and significantly sped up indexing of the INSEE corpus[3].

Coarser-grain indexing of Eurostat statistics The large size of this corpus prevents cell- or row-level metadata indexing, as the index might outgrow the memory. Instead, we index *occurrences of statistical concept codes in datasets*, as follows. Let c be a Eurostat concept, e.g., "EL", appearing in dataset d at a location $l \in \mathcal{L}$, and d_c be the decoding of c, e.g., "*Elementary school*" for "EL". Let $T_{dc} = \{t_1, t_2, \ldots, t_N\}$ be the tokens in d_c, and for $1 \leq i \leq N$, let t_i^j, for $1 \leq j \leq 50$, be the tokens closest to t_i. For each $t_i \in T_{dc}$, we insert in the **term-dataset index** I_T, also stored in Redis:

- a (t_i, d, l) entry;
- for every t_i^j similar to t_i, an entry $(t_i^j, d, l, dist, t_i)$, where $dist$ is the distance between t_i and t_i^j.

Indexing the complete Eurostat data in this way took around 4 minutes.

Given the query $Q = \{k_1, \ldots, k_n\}$, we search I_{CL} and I_T for entries of the form (k_i, d, l) or $(k_i, d, l, dist, k_i')$. Any dataset having an entry for at least one k_i is potentially interesting; we retain the 20 highest-score ones.

Dataset ranking We rank datasets based on the relevance score introduced in (Cao et al., 2018). It is a weighted combination of the word distances between the query keywords and the datasets' metadata; the weights reflect the locations where relevant terms appear in each dataset. We have also experimented with the classic BM25 (Robertson and Zaragoza, 2009) computed over all the datasets' metadata, but the results were less good, in particular, because BM25 does not handle synonyms well. We also considered embedding the query and the metadata using Sentence-Bert (Reimers and Gurevych, 2019) and comparing these with the query embedding, but opted not to use it because, for our purposes, the term location in the metadata is important, and treating the metadata as a single text loses this information.

5.2 Data Cell Indexing and Search

Our next task is to extract results at the finest granularity level possible. Let d be one of the most interesting datasets, and $I(d)$ be the set of **all index entries for the query** Q **and** d. For our sample query Q and dataset in Figure 1, $I(d)$ contains:

- For "middle school", header (H) entries for "Middle school" (exact), as well as for "High school" and "Elementary school" (similar); a title (T) entry for "student" (similar); and a comment (C) entry for "school" (similar);
- For "pupils", H, T, and C entries for the similar words above;
- For "Île-de-France", an exact H entry, and two similar H entries for "Paris" and "Essonne";
- For "2020", exact H entries.

[3] Together with other optimizations related to batching calls to the Spacy tokenizer and pipelining the indexing with the data acquisition process, this brought the total INSEE indexing time from 29 hours to 4 minutes.

If $I(d)$ only features title (T) or comment (C) locations, then d is pertinent as a whole, and we do not search for cell-level answers.

On the contrary, if $I(d)$ has several header entries (having $l = H$), matching two or more distinct query keywords (or close terms), this means that d holds some fine-granularity results for the query. If $I(d)$ holds an entry along each dataset dimension d, these entries, together, designate exactly one cell, which we should return. Otherwise, the result is a collection of all the cells from d characterized by the dimension values designated by the entries in $I(d)$.

In our example, we should return the cells for ”MI 2019”, ”2020”, and locations ”Paris” and ”Essonne”, which belong to Île-de-France. For that:

1. If d is an INSEE dataset, $I(d)$ contains the headers of the respective row and column headers. Then, the cell is identified by asking a SPARQL (W3C, 2013) query, evaluated by Fuseki, as in (Cao et al., 2018). The query requests ”all the data cells from dataset d whose closest header cells are those from $I(d)$”.

2. If d is an Eurostat dataset, $I(d)$ only specifies that ”*some* row (column) headers match”, and more effort is needed to identify the relevant cells. A Eurostat file has at most a few dozen columns, but it may have tens of millions of rows.
   - To find the *column* referred to by an $I(d)$ entry whose key is k, we search for k in the first (header) line of d.
   - To identify the relevant *rows* efficiently, we created another index I_R on the Eurostat data files, inspired by the Adaptive Positional Map of (Alagiannis et al., 2015). I_R stores the positions, in the data file of d, of the rows containing a certain keyword k in their header. We store I_R directly as a binary file on disk.
   - Knowing the rows and column indexes, we read those row(s) from d, and extract from them the relevant data cell(s).

Using Fuseki, cell extraction takes 35ms up to 2.86s. On Eurostat, using I_R, we record 4.76μs up to 2.66s. The lower bound is higher for INSEE because we have to pass SPARQL queries across a connection to the Fuseki server.

6 Claim Detection

A claim is a statement to be validated, that is, we aim to establish if it is true or false. The validation is achieved by finding related statements, called evidence, which back up or disprove the claim. In our work, the claims are detected in an input text, while the evidence is retrieved from a set of trusted sources, our reference datasets. Our platform detects claims from text stored in *.txt*, *.odt*, *.docx* or *.pdf* files, and from the Twitter and Facebook posts of public figures. Our platform regularly retrieves the most recent updates of a predefined group of users for posts.

6.1 Statistical Claim Detection

Previous work addresses statistical claim detection in a supervised manner by predicting statistical entity-value pair from text patterns (Vlachos and Riedel, 2015). In (Duc Cao et al., 2019), the authors introduced a statistical claim detection method that given an input set of statistical entities, e.g. *chômage*, *coefficient budgétaire*) and a sentence, it retrieves all the *statistical statements* of the form $\langle$`statistical entity, numerical value, and unit, date`$\rangle$ present in the sentence. *The statistical statement, if present, represents the statistical claim to be verified.* The statistical entities and units are retrieved using exact string matching, while the date is extracted using HeidelTime (Strötgen and Gertz, 2010), a time expression parser. If the parser finds no date, the posting timestamp is used. More context about the claim to be verified is found using a Named Entity Recognition (NER) model, which returns organizations and locations. We note, however, that the organization and location are optional, while a statistical statement is not complete without one of its three elements. The initial statistical entity list is constructed from the reference datasets by taking groups of tokens from the headers of tables, we refer to (Duc Cao et al., 2019) for more details.

We improved this method to optimize both its speed and the quality of extractions. We refer to the two methods as OriginalStatClaim (Duc Cao et al., 2019) and StatClaim. We first performed a more careful match between the tokens of a sentence and our input statistical entities. Using the syntactic tree of the sentence and a lemmatizer, statistical entities are matched using their lemma and are extended to contain the entire nominal group of the matched token. Numerical values are associated with units using both lemmas matching from our set of units and syntactic analysis. Units can be a noun following a numerical value or a nominal group containing one or more units. (e.g. "*millions d'euros*"). As in the original approach, if we retrieve a statistical statement of the form ⟨`statistical entity, numerical value, and unit, date`⟩, we have found a claim to verify. In the default setting of our algorithm, a claim should contain all three elements. In addition, we filter out claims from sentences whose verb is in the future tense or the first person since these are promises about the future and not verifiable. Journalists found, however, that these may also be interesting for their long-term, issue-centric work (Juneja and Mitra, 2022). Thus, StatCheck allows them to turn the future and first-person filters on and off.

6.2 Check-worthy Claim Detection

To complement the statistical claim detection model, we developed a model that is not conditioned on a set of initial statistical entities. The model classifies a sentence as check-worthy or not, where check-worthiness is defined as *sentences containing factual claims that the general public will be interested in learning about their veracity* (Arslan et al., 2020). We leveraged the ClaimBuster dataset (Arslan et al., 2020), containing check-worthy claims in English from the U.S. Presidential debates, to train a cross-lingual language model, XLM-R (Conneau et al., 2019), which can perform zero-shot classification on French sentences after training on English data.

The ClaimBuster dataset ClaimBuster is a crowd-sourced dataset where the sentences from the 15 U.S. presidential elections debates from 1960 to 2016 have been annotated. The labels are Non-Factual Sentences (NFS), Unimportant Factual Sentences (UFS) or Check-Worthy Factual Sentences (CFS). The dataset contains $23K$ sentences, and the authors produced a subset of higher quality of $11K$ sentences for training models on classification tasks. In this smaller dataset, the NFS and UFS labels are grouped as negative labels, and the CFS labels are considered positive. We chose this higher-quality dataset to fine-tune the XLM-R model.

Fine-tuning the XLM-R model The XLM-R model is a Transformer-based masked language model trained on one hundred languages with 2.5TB of Common Crawl data. It achieves state-of-the-art results on multilingual tasks such as the XNLI benchmark (Conneau et al., 2018), while remaining competitive on monolingual tasks. We used a pretrained model with a vocabulary size of $250K$, 12 hidden layers of size 768 and 12 attention heads. We used a weighted cross-entropy loss to account for the unbalanced ratio of labels. The dataset was split into train, dev and test datasets with a ratio of $80\%/\%10\%/10\%$.

Dataset	P	R	F1 score
ClaimBuster	0.883	0.848	0.865
French tweets	0.612	0.769	0.682

Table 1: Evaluation of the fine-tuned XLM-R model.

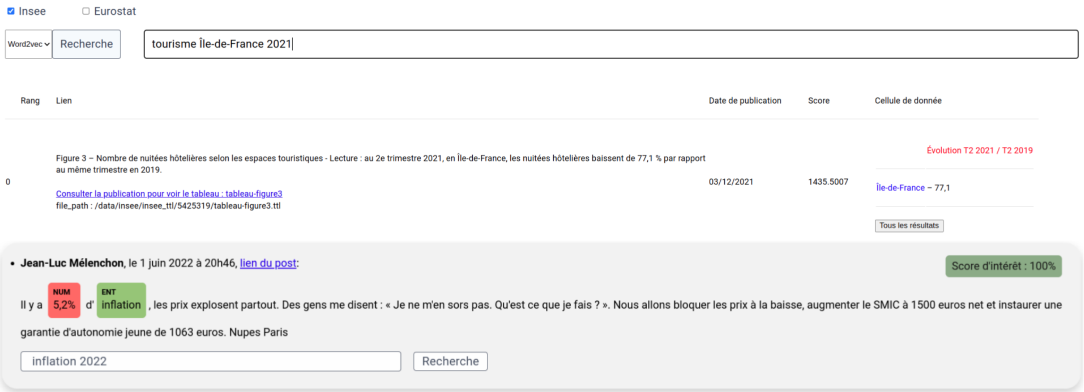
Models	P	R	F1 score
OriginalStatClaim	0.692	0.466	0.557
StatClaim	**0.833**	0.517	0.638
CheckWorthyClaim	0.701	**0.915**	**0.794**

Table 2: Model evaluation on verifiable numerical claims.

Models	P	R	F1 score
OriginalStatClaim	0.282	0.688	0.400
StatClaim	**0.333**	0.750	**0.462**
CheckWorthyClaim	0.195	**0.938**	0.323

Table 3: Model evaluation on INSEE statistical claims.

Evaluation To optimize the performance, we trained the model with different hyperparameters. The best results were obtained with a learning rate of $5 \cdot 10^{-5}$, a batch size of 64, and using the AdamW optimizer. To evaluate the performance of the different models on French data, we annotated 200 randomly sampled French tweets and labeled them as check-worthy or not following the definition in (Arslan et al., 2020). Two annotators labeled each tweet; in the golden standard, a tweet is deemed check-worthy if both annotators agree on it, and not check-worthy otherwise. The Cohen Kappa score for inter-annotator agreement is 0.6, which is considered moderate to substantial agreement. The results can be found in Table 1. The performance on this test set is encouraging, however lower than on the original English dataset. This is expected given the zero-shot setting, as the tweets' format and vocabulary might differ from the ones in the training dataset.



☑ Insee ☐ Eurostat

Word2vec Recherche tourisme Île-de-France 2021

Rang	Lien	Date de publication	Score	Cellule de donnée
0	Figure 3 – Nombre de nuitées hôtelières selon les espaces touristiques - Lecture : au 2e trimestre 2021, en Île-de-France, les nuitées hôtelières baissent de 77,1 % par rapport au même trimestre en 2019. Consulter la publication pour voir le tableau : tableau-figure3 file_path : /data/insee/insee_ttl/5425319/tableau-figure3.ttl	03/12/2021	1435.5007	Évolution T2 2021 / T2 2019 Île-de-France – 77,1

- **Jean-Luc Mélenchon**, le 1 juin 2022 à 20h46, lien du post:

Score d'intérêt : 100%

Il y a NUM 5,2% d' ENT inflation , les prix explosent partout. Des gens me disent : « Je ne m'en sors pas. Qu'est ce que je fais ? ». Nous allons bloquer les prix à la baisse, augmenter le SMIC à 1500 euros net et instaurer une garantie d'autonomie jeune de 1063 euros. Nupes Paris

inflation 2022

Figure 3: Screen captures of StatCheck' GUI. Top: statistic search interface with sample query result (data cell with row header in blue and column header in red); bottom: tweet analysis interface.

6.3 Integration and Evaluation of the Claim Detection Models

We evaluate the claim detection models, *(OriginalStatClaim* (Duc Cao et al., 2019), *StatClaim* and *CheckWorthyClaim)*, on a set of 1595 tweets. Each tweet was labeled with two classes: *"Verifiable numerical claim"* (True if the tweet contains at least one numerical and verifiable claim") and *"INSEE statistical claim"* (True if the tweet contains at least one numerical, statistical claim verifiable against the INSEE dataset"). We chose these two labels as the first one gives us an indication of the tweets that can be verified if we had unlimited access to resources, while the second class identifies the tweets verifiable in the setting in which we have access to only one resource. We gathered 1595 random tweets from our scraped dataset to construct our set. Then, we automatically detected if a tweet contained a numerical value, if not, the tweet was labeled as negative for both classes. After that first step, we manually labeled the remaining 101 tweets. Two annotators labeled each tweet, and the gold standard was chosen as True if both annotators agreed. For the class *"verifiable numerical claim"*, we obtained a Kappa inter-Annotator Agreement score of 0.917 (almost perfect agreement), and *59* tweets were labeled as positive. For the class *"INSEE statistical claim"* we obtained an inter-annotator Agreement score of 0.807 (substantial agreement) and *16* tweets were labeled as positive.

Evaluation procedure For StatClaim and OriginalStatClaim, a tweet is considered positive if models return at least one extracted statistical statement. Our StatClaim was used in its default configuration: extractions with numerical values and without verbs conjugated in the future or in the first person. For CheckWorthyClaim, a tweet is considered positive if the model returns a check-worthy score > 0.9. We report the results in Table 2 and Table 3. StatClaim performs better than the original at detecting INSEE verifiable claims, and CheckWorthyClaim vastly outperforms both models on the detection of numerical claims, as they are a subset of check-worthy sentences that the model was trained to detect.

Finally, we evaluate the performance of our model directly against the journalist authors' prior manual work. For example, during the 2022 French presidential debate, the journalist team highlighted 29 of the 1954 uttered sentences and fact-checked them. The XLM-R model, on the other hand, classifies 443 of these sentences as check-worthy, and 27 of the 29 sentences chosen by the journalists are correctly classified. In other words, our model reduces by 77% the number of sentences to consider while retaining 93% of the sentences the journalists actually want to fact-check, saving the journalists considerable time without them missing too many important claims.

Default claim detection strategy. By default, StatCheck uses StatClaim for statistical claim detection. However, given the good performance of CheckWorthyClaim on numerical claims, we allow users to switch to it, even if we might not be able to verify them against the reference datasets.

7 Conclusion and Perspectives

Fact-checking journalists need automated tools to help scale up their daily work. We developed the StatCheck tool, which allows the journalist authors to focus their attention directly on check-worthy statements falling into one of two overlapping classes: those that can be checked based on statistics from two major institutions; and those that human users find interesting, even if the data to back up the checks is not present in the database. StatCheck is in daily use in the fact-checking team; Figure 3 illustrates its GUI.

Quantitative question answering based on open data is gaining interest (Ho et al., 2020, 2022). In our continuing collaboration, we will work to extend StatCheck with more multidimensional statistic datasets from national governments and international organizations.

Acknowledgements This work is partially funded by AI Chair SourcesSay project (ANR-20-CHIA-0015-01).

References

2022. Media audience survey. Online (anonymized for double-blind reviewing).

Ioannis Alagiannis, Renata Borovica-Gajic, Miguel Branco, Stratos Idreos, and Anastasia Ailamaki. 2015. Nodb: efficient query execution on raw data files. *Commun. ACM*, 58(12):112–121.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-worthy Factual Claims. In *14th International AAAI Conference on Web and Social Media*. AAAI.

Tien Duc Cao, Ioana Manolescu, and Xavier Tannier. 2017. Extracting linked data from statistic spreadsheets. In *International Workshop on Semantic Big Data*, International Workshop on Semantic Big Data, pages 1 – 5.

Tien-Duc Cao, Ioana Manolescu, and Xavier Tannier. 2018. Searching for Truth in a Database of Statistics. In *WebDB*, pages 1–6. Code available at: https://gitlab.inria.fr/cedar/excel-search.

Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A content management perspective on fact-checking. In *WWW (Companion Volume)*, pages 565–574. ACM.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact : A large-scale dataset for table-based fact verification. In *ICLR*.

Gianluca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLoS ONE*, 10(6).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Tien Duc Cao, Ioana Manolescu, and Xavier Tannier. 2019. Extracting statistical mentions from textual claims to provide trusted content. In *NLDB*. Code available at: https://gitlab.inria.fr/cedar/statstical_mentions.

Eurostat. 2022. European geographic location dictionary. https://ec.europa.eu/eurostat/estat-navtree-portlet-prod/BulkDownloadListing?sort=1&file=dic%2Ffr%2Fgeo.dic.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *ACL*, pages 4320–4333.

Vinh Thinh Ho, Koninika Pal, Niko Kleer, Klaus Berberich, and Gerhard Weikum. 2020. Entities with quantities: Extraction, search, and ranking. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 833–836. ACM.

Vinh Thinh Ho, Daria Stepanova, Dragan Milchevski, Jannik Strötgen, and Gerhard Weikum. 2022. Enhancing knowledge bases with quantity facts. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 893–901. ACM.

Saehan Jo, Immanuel Trummer, Weicheng Yu, Xuezhi Wang, Cong Yu, Daniel Liu, and Niyati Mehta. 2019. Verifying text summaries of relational data sets. In *SIGMOD*, SIGMOD '19, page 299–316.

Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. In *Proceedings of the 2022 Conference On Computer Supported Cooperative Work And Social Computing (CSCW '22). Proceedings of the ACM on Human-Computer Interaction.*

Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: Fact checking statistical claims. *Proc. VLDB Endow.*, 13(12):2965–2968.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *IJCAI*, pages 4551–4558. Survey Track.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6859–6866. AAAI Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Mohammed Saeed and Paolo Papotti. 2021. Fact-checking statistical claims with tables. IEEE Data Engineering Bulletin.

Vikash Singh. 2017. Replace or retrieve keywords in documents at scale. *CoRR*, abs/1711.00046.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Int'l. Workshop on Semantic Evaluation*, pages 321–324.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics.

W3C. 2013. SPARQL 1.1 query language. http://www.w3.org/TR/sparql11-query/.

www-rdf. 2022. Resource description format. https://www.w3.org/RDF/.

www-redis. 2022. Redis. https://redis.io/.

www-tdb2. 2022. Apache Jena TDB. https://jena.apache.org/documentation/tdb2/.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *FEVER*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

A Multidimensional Datasets Example

A.1 Insee

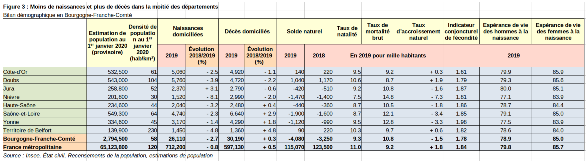


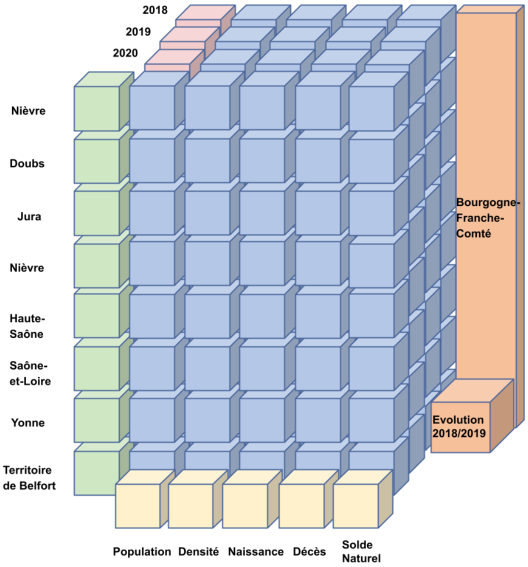
Figure 3 : Moins de naissances et plus de décès dans la moitié des départements

Bilan démographique en Bourgogne-Franche-Comté

	Estimation de population au 1er janvier 2020 (provisoire)	Densité de population au 1er janvier 2020 (hab/km²)	Naissances domiciliées		Décès domiciliés		Solde naturel		Taux de natalité	Taux de mortalité brut	Taux d'accroissement naturel	Indicateur conjoncturel de fécondité	Espérance de vie des hommes à la naissance	Espérance de vie des femmes à la naissance
			2019	Évolution 2018/2019 (%)	2019	Évolution 2018/2019 (%)	2019	2018	En 2019 pour mille habitants			2019		
Côte-d'Or	532,500	61	5,060	- 2.5	4,920	- 1.1	140	220	9.5	9.2	+ 0.3	1.61	79.9	85.9
Doubs	543,000	104	5,760	- 3.9	4,720	- 2.2	1,040	1,170	10.6	8.7	+ 1.9	1.79	79.3	85.6
Jura	258,800	52	2,370	+ 3.1	2,790	- 0.6	-420	-510	9.2	10.8	- 1.6	1.87	80.0	85.1
Nièvre	201,800	30	1,520	- 8.1	2,990	- 2.0	-1,470	-1,400	7.5	14.8	- 7.3	1.81	77.1	83.9
Haute-Saône	234,600	44	2,040	- 3.2	2,480	+ 0.4	-440	-360	8.7	10.5	- 1.8	1.86	78.7	84.4
Saône-et-Loire	549,300	64	4,740	- 2.3	6,640	+ 2.9	-1,900	-1,600	8.7	12.1	- 3.4	1.85	79.1	85.0
Yonne	334,600	45	3,170	- 1.4	4,290	+ 1.8	-1,120	-990	9.5	12.8	- 3.3	1.98	77.5	83.9
Territoire de Belfort	139,900	230	1,450	- 4.8	1,360	+ 4.8	90	220	10.3	9.7	+ 0.6	1.82	78.6	84.0
Bourgogne-Franche-Comté	**2,794,500**	**58**	**26,110**	**- 2.7**	**30,190**	**+ 0.3**	**-4,080**	**-3,250**	**9.3**	**10.8**	**- 1.5**	**1.78**	**78.9**	**85.0**
France métropolitaine	**65,123,800**	**120**	**712,200**	**- 0.8**	**597,130**	**+ 0.5**	**115,070**	**123,500**	**11.0**	**9.2**	**+ 1.8**	**1.84**	**79.8**	**85.7**

Source : Insee, État civil, Recensements de la population, estimations de population

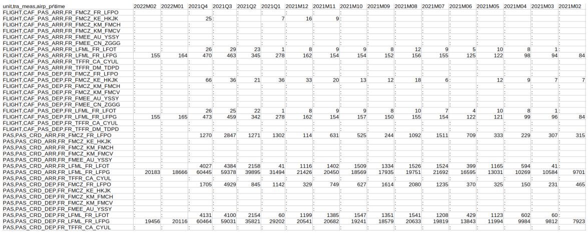
(a) An Insee table



(b) Insee multidimensional statistic dataset view

Figure 4: Example of Insee dataset

A.2 Eurostat

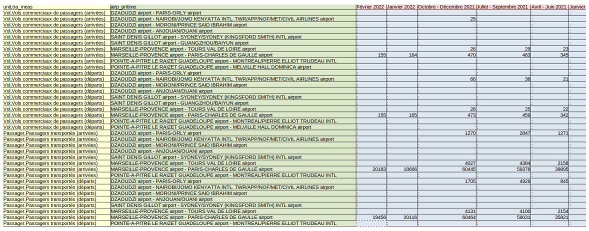


unit,tra_meas,airp_pr\time	2022M02	2022M01	2021Q4	2021Q3	2021Q2	2021Q1	2021M12	2021M11	2021M10	2021M09	2021M08	2021M07	2021M06	2021M05	2021M04	2021M03	2021M02
FLIGHT,CAF_PAS_ARR,FR_FMCZ_FR_LFPO	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_FMCZ_KE_HKJK	:	:	25	:	:	7	16	9	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_FMCZ_KM_FMCH	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_FMCZ_KM_FMCV	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_FMEE_AU_YSSY	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_FMEE_CN_ZGGG	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_LFML_FR_LFOT	:	:	26	29	23	1	8	9	9	8	12	9	5	10	8	1	:
FLIGHT,CAF_PAS_ARR,FR_LFML_FR_LFPG	155	164	470	463	345	278	162	154	154	152	156	155	125	122	98	94	84
FLIGHT,CAF_PAS_ARR,FR_TFFR_CA_CYUL	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_ARR,FR_TFFR_DM_TDPD	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_FMCZ_FR_LFPO	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_FMCZ_KE_HKJK	:	:	66	36	21	36	33	20	13	12	18	6	:	12	9	7	7
FLIGHT,CAF_PAS_DEP,FR_FMCZ_KM_FMCH	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_FMCZ_KM_FMCV	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_FMEE_AU_YSSY	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_FMEE_CN_ZGGG	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_LFML_FR_LFOT	:	:	26	25	22	1	8	9	9	8	10	7	4	10	8	1	:
FLIGHT,CAF_PAS_DEP,FR_LFML_FR_LFPG	155	165	473	459	342	278	162	154	157	150	155	154	122	121	99	96	84
FLIGHT,CAF_PAS_DEP,FR_TFFR_CA_CYUL	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
FLIGHT,CAF_PAS_DEP,FR_TFFR_DM_TDPD	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_ARR,FR_FMCZ_FR_LFPO	:	:	1270	2847	1271	1302	114	631	525	244	1092	1511	709	333	229	307	315
PAS,PAS_CRD_ARR,FR_FMCZ_KE_HKJK	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_ARR,FR_FMCZ_KM_FMCH	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_ARR,FR_FMCZ_KM_FMCV	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_ARR,FR_FMEE_AU_YSSY	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_ARR,FR_LFML_FR_LFOT	:	:	4027	4384	2158	41	1116	1402	1509	1334	1526	1524	399	1165	594	41	:
PAS,PAS_CRD_ARR,FR_LFML_FR_LFPG	20183	18666	60445	59378	39895	31494	21426	20450	18569	17935	19751	21692	16595	13031	10269	10584	9701
PAS,PAS_CRD_ARR,FR_TFFR_CA_CYUL	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_DEP,FR_FMCZ_FR_LFPO	:	:	1705	4929	845	1142	329	749	627	1614	2080	1235	370	325	150	231	465
PAS,PAS_CRD_DEP,FR_FMCZ_KE_HKJK	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_DEP,FR_FMCZ_KM_FMCH	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_DEP,FR_FMCZ_KM_FMCV	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_DEP,FR_FMEE_AU_YSSY	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
PAS,PAS_CRD_DEP,FR_LFML_FR_LFOT	:	:	4131	4100	2154	60	1199	1385	1547	1351	1541	1208	429	1123	602	60	:
PAS,PAS_CRD_DEP,FR_LFML_FR_LFPG	19456	20116	60464	59031	35821	29202	20541	20682	19241	18579	20633	19819	13843	11994	9984	9812	7923
PAS,PAS_CRD_DEP,FR_TFFR_CA_CYUL	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

(a) An example of an encoded eurostat table

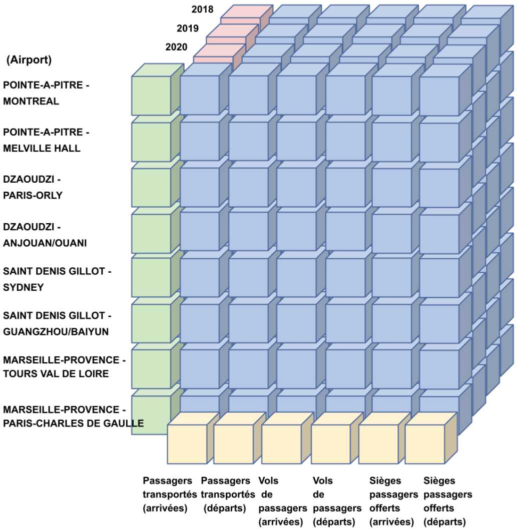
Statistical Code	Statistical Concept
FR_FMCZ_FR_FMEE	DZAOUDZI airport -SAINT DENIS GILLOT airport
FR_FMCZ_FR_LFML	DZAOUDZI airport -MARSEILLE-PROVENCE airport
FR_FMCZ_FR_LFPG	DZAOUDZI airport -PARIS-CHARLES DE GAULLE airport
FR_FMCZ_FR_LFPO	DZAOUDZI airport -PARIS-ORLY airport
FR_FMCZ_KE_HKJK	DZAOUDZI airport -NAIROBI/JOMO KENYATTA INTL. TWR/APP/NOF/MET/CIVIL AIRLINES airport
FR_FMCZ_KM_FMCH	DZAOUDZI airport -MORONI/PRINCE SAID IBRAHIM airport
FR_FMCZ_KM_FMCV	DZAOUDZI airport -ANJOUAN/OUANI airport
FR_FMCZ_MG_FMMI	DZAOUDZI airport -ANTANANARIVO/IVATO airport
FR_FMCZ_MG_FMNM	DZAOUDZI airport -MAHAJANGA/PH. TSIRANANA airport
FR_FMCZ_MG_FMNN	DZAOUDZI airport -NOSY-BE airport
FR_FMEE_AU_YSSY	SAINT DENIS GILLOT airport -SYDNEY/SYDNEY (KINGSFORD SMITH) INTL airport
FR_FMEE_CN_ZGGG	SAINT DENIS GILLOT airport -GUANGZHOU/BAIYUN airport
FR_FMEE_FR_FMCZ	SAINT DENIS GILLOT airport -DZAOUDZI airport
FR_FMEE_FR_FMEP	SAINT DENIS GILLOT airport -SAINT PIERRE PIERREFONDS airport
FR_FMEE_FR_LFBD	SAINT DENIS GILLOT airport -BORDEAUX-MERIGNAC airport
FR_FMEE_FR_LFBO	SAINT DENIS GILLOT airport -TOULOUSE/BLAGNAC airport

(b) An subsample of a dictionnary used to decode the encoded table (a)



unit,tra_meas	airp_pr\time	Février 2022	Janvier 2022	Octobre - Décembre 2021	Jullet - Septembre 2021	Avril - Juin 2021	Janvier
Vol,Vols commerciaux de passagers (arrivées)	DZAOUDZI airport - PARIS-ORLY airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	DZAOUDZI airport - NAIROBI/JOMO KENYATTA INTL. TWR/APP/NOF/MET/CIVIL AIRLINES airport	:	:	25	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	DZAOUDZI airport - MORONI/PRINCE SAID IBRAHIM airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	DZAOUDZI airport - ANJOUAN/OUANI airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	SAINT DENIS GILLOT airport - SYDNEY/SYDNEY (KINGSFORD SMITH) INTL airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	SAINT DENIS GILLOT airport - GUANGZHOU/BAIYUN airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	MARSEILLE-PROVENCE airport - TOURS VAL DE LOIRE airport	:	:	26	29	23	
Vol,Vols commerciaux de passagers (arrivées)	MARSEILLE-PROVENCE airport - PARIS-CHARLES DE GAULLE airport	155	164	470	463	345	
Vol,Vols commerciaux de passagers (arrivées)	POINTE-A-PITRE LE RAIZET GUADELOUPE airport - MONTREAL/PIERRE ELLIOT TRUDEAU INTL		:	:	:	:	:
Vol,Vols commerciaux de passagers (arrivées)	POINTE-A-PITRE LE RAIZET GUADELOUPE airport - MELVILLE HALL DOMINICA airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	DZAOUDZI airport - PARIS-ORLY airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	DZAOUDZI airport - NAIROBI/JOMO KENYATTA INTL. TWR/APP/NOF/MET/CIVIL AIRLINES airport	:	:	66	36	21	
Vol,Vols commerciaux de passagers (départs)	DZAOUDZI airport - MORONI/PRINCE SAID IBRAHIM airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	DZAOUDZI airport - ANJOUAN/OUANI airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	SAINT DENIS GILLOT airport - SYDNEY/SYDNEY (KINGSFORD SMITH) INTL airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	SAINT DENIS GILLOT airport - GUANGZHOU/BAIYUN airport	:	:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	MARSEILLE-PROVENCE airport - TOURS VAL DE LOIRE airport	:	:	26	25	22	
Vol,Vols commerciaux de passagers (départs)	MARSEILLE-PROVENCE airport - PARIS-CHARLES DE GAULLE airport	155	165	473	459	342	
Vol,Vols commerciaux de passagers (départs)	POINTE-A-PITRE LE RAIZET GUADELOUPE airport - MONTREAL/PIERRE ELLIOT TRUDEAU INTL		:	:	:	:	:
Vol,Vols commerciaux de passagers (départs)	POINTE-A-PITRE LE RAIZET GUADELOUPE airport - MELVILLE HALL DOMINICA airport	:	:	:	:	:	:
Passager,Passagers transportés (arrivées)	DZAOUDZI airport - PARIS-ORLY airport	:	:	1270	2847	1271	
Passager,Passagers transportés (arrivées)	DZAOUDZI airport - NAIROBI/JOMO KENYATTA INTL. TWR/APP/NOF/MET/CIVIL AIRLINES airport	:	:	:	:	:	:
Passager,Passagers transportés (arrivées)	DZAOUDZI airport - MORONI/PRINCE SAID IBRAHIM airport	:	:	:	:	:	:
Passager,Passagers transportés (arrivées)	DZAOUDZI airport - ANJOUAN/OUANI airport	:	:	:	:	:	:
Passager,Passagers transportés (arrivées)	SAINT DENIS GILLOT airport - SYDNEY/SYDNEY (KINGSFORD SMITH) INTL airport	:	:	:	:	:	:
Passager,Passagers transportés (arrivées)	MARSEILLE-PROVENCE airport - TOURS VAL DE LOIRE airport	:	:	4027	4384	2158	
Passager,Passagers transportés (arrivées)	MARSEILLE-PROVENCE airport - PARIS-CHARLES DE GAULLE airport	20183	18666	60445	59378	39895	
Passager,Passagers transportés (arrivées)	POINTE-A-PITRE LE RAIZET GUADELOUPE airport - MONTREAL/PIERRE ELLIOT TRUDEAU INTL		:	:	:	:	:
Passager,Passagers transportés (départs)	DZAOUDZI airport - PARIS-ORLY airport	:	:	1705	4929	845	
Passager,Passagers transportés (départs)	DZAOUDZI airport - NAIROBI/JOMO KENYATTA INTL. TWR/APP/NOF/MET/CIVIL AIRLINES airport	:	:	:	:	:	:
Passager,Passagers transportés (départs)	DZAOUDZI airport - MORONI/PRINCE SAID IBRAHIM airport	:	:	:	:	:	:
Passager,Passagers transportés (départs)	DZAOUDZI airport - ANJOUAN/OUANI airport	:	:	:	:	:	:
Passager,Passagers transportés (départs)	SAINT DENIS GILLOT airport - SYDNEY/SYDNEY (KINGSFORD SMITH) INTL airport	:	:	:	:	:	:
Passager,Passagers transportés (départs)	MARSEILLE-PROVENCE airport - TOURS VAL DE LOIRE airport	:	:	4131	4100	2154	
Passager,Passagers transportés (départs)	MARSEILLE-PROVENCE airport - PARIS-CHARLES DE GAULLE airport	19456	20116	60464	59031	35821	
Passager,Passagers transportés (départs)	POINTE-A-PITRE LE RAIZET GUADELOUPE airport - MONTREAL/PIERRE ELLIOT TRUDEAU INTL		:	:	:	:	:

(c) The decoded Eurostat table


2018
2019
2020
(Airport)
POINTE-A-PITRE - MONTREAL
POINTE-A-PITRE - MELVILLE HALL
DZAOUDZI - PARIS-ORLY
DZAOUDZI - ANJOUAN/OUANI
SAINT DENIS GILLOT - SYDNEY
SAINT DENIS GILLOT - GUANGZHOU/BAIYUN
MARSEILLE-PROVENCE - TOURS VAL DE LOIRE
MARSEILLE-PROVENCE - PARIS-CHARLES DE GAULLE
Passagers transportés (arrivées)
Passagers transportés (départs)
Vols de passagers (arrivées)
Vols de passagers (départs)
Sièges passagers offerts (arrivées)
Sièges passagers offerts (départs)

(d) Eurostat multidimensional statistic dataset view

Figure 5: Example of Eurostat dataset