Decoding the News Media Diet of Disinformation Spreaders

Anna Bertani [1,2,]*[ID]**, Valeria Mazzeo** [2] **and Riccardo Gallotti** [2,]*

1   Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 9, 38123 Povo, TN, Italy

2   CHuB, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, TN, Italy; valesdn@gmail.com

*   Correspondence: anna.bertani@unitn.it (A.B.); rgallotti@fbk.eu (R.G.)

Abstract: In the digital era, information consumption is predominantly channeled through online news media and disseminated on social media platforms. Understanding the complex dynamics of the news media environment and users' habits within the digital ecosystem is a challenging task that requires, at the same time, large databases and accurate methodological approaches. This study contributes to this expanding research landscape by employing network science methodologies and entropic measures to analyze the behavioral patterns of social media users sharing news pieces and dig into the diverse news consumption habits within different online social media user groups. Our analyses reveal that users are more inclined to share news classified as fake when they have previously posted conspiracy or junk science content and vice versa, creating a series of "misinformation hot streaks". To better understand these dynamics, we used three different measures of entropy to gain insights into the news media habits of each user, finding that the patterns of news consumption significantly differ among users when focusing on disinformation spreaders as opposed to accounts sharing reliable or low-risk content. Thanks to these entropic measures, we quantify the variety and the regularity of the news media diet, finding that those disseminating unreliable content exhibit a more varied and, at the same time, a more regular choice of web-domains. This quantitative insight into the nuances of news consumption behaviors exhibited by disinformation spreaders holds the potential to significantly inform the strategic formulation of more robust and adaptive social media moderation policies.

Keywords: misinformation; socio-technical systems; entropy; network science; computational social science

1. Introduction

In the contemporary digital landscape, the spread of misleading and false information is becoming one of the most challenging issues that our society has ever faced [1,2]. With the advent of online social media [3], the access to information has drastically increased, allowing ideas and news to circulate rapidly and to reach a global audience. This shift has prompted individuals to actively seek and consume information directly through online social media platforms. In contrast to the pre-digital era, where people relied on traditional news media outlets and printed journals for information, the digital age has multiplied and simplified access to these sources. News media journals, once accessible primarily through physical copies, are now readily available with just a few clicks, marking a fundamental change in the way individuals obtain and engage with information [4]. In this context, online social media serves as a digital space where people can look for online news articles as well as come across articles posted by other users [5]. New technologies have increased the possibilities for how people receive and send information [6], completely disrupting the reading habits of online users [7]. The increasing number of accessible online news media articles has surely contributed to the problem of information overload [8–11], which is also caused by the absence of a proper content regulation on the Internet [12] and might affect the different spheres of democracy, such

as the freedoms of thought, belief, and expression [7]. In the current scenario, the spread of misinformative and disinformative content exerts its influence across various social domains, including, but not limited to, the political sphere, international relations, and the formulation of public policies related to climate change [13]. Moreover, it has significant ramifications for numerous health-related issues, as evidenced by the dissemination of misleading narratives by the anti-vaccination movement [14,15] and, notably, the onset of the COVID-19 pandemic in early 2020 [16]. In this context, the information overload might be exacerbated during highly debated and contentious topics, such as political elections [17] or the COVID-19 pandemic [18], which have been characterized by a notable increase in the dissemination of unreliable content [19]. Based on these premises, the aim of this work is to investigate the intricate news media digital landscape by analyzing the type of news media sources that dominated the online conversation during the first year of the COVID-19 pandemic. In addition, our interest focuses on the characterization of the behavioral patterns that each user displays in the selection of the news media types and web-domains chosen to read and share online. While extensive research has been conducted on news media consumption [20], particularly concerning the transformative impact of the Internet on information acquisition through websites and online articles, a notable gap exists in quantifying the diversity of users' news media diets within the digital and social landscape. Consequently, our interest goes beyond a mere descriptive analysis of user behavior in accessing and staying informed through online social media. We aim to introduce two measures capable of capturing the *variety* of news media consumption and the *regularity* of accessing this information, drawing inspiration form two distinct entropic measures. A deeper insight into users' news media habits might shed light on the extent to which there is a variety in choosing different news media journal or if they are more likely to stick to a few of them. Decoding which sources online users rely on and share might raise our knowledge about how misinformative and disinformative content spread online and how users are more likely to be attracted by certain web-domains rather than others over time.

2. Materials and Methods

2.1. Overview of the Dataset

We analyzed online social media data gathered by the COVID-19 Infodemic Platform [21], a comprehensive database containing messages posted on Twitter throughout the COVID-19 pandemic. Specifically, we considered about 9,157,655 messages containing at least one URL (of which we were able to classify 2,549,226) posted in 2020.

We only consider the original messages, defined as tweets, thus not including in our analysis replies, retweets, or quotes. To ensure a robust sample, we considered in our analysis only 25 countries with an average of at least 500 tweets per day collected on our platform. The countries included in our analysis are those associated with the following iso3 codes: COL, IRL, ITA, VEN, TUR, SWE, ESP, BRA, NLD, DEU, JPN, NGA, POL, CHE, AUS, AUT, CAN, CUB, ECU, GRC, MEX, PRT, ROU, SLV, ZAF. In total, we gathered information about messages posted by 211,493 users from the beginning of February until the end of 2020.

2.2. News Reliability

In our analysis, we evaluate the reliability of news domains by matching the URLs included in the textual content of messages against information manually verified from various publicly accessible databases that cover scientific and journalistic sources. We specifically utilized data from MediaBiasFactCheck [22], an organization maintaining an extensive and regularly updated database. Their methodology involves a comprehensive evaluation of the ideological leanings and factual accuracy of media and information outlets, employing both quantitative metrics and qualitative assessments, https://mediabiasfactcheck.com/methodology/ (accessed on 9 March 2024). Their evaluation is based on the web-domain itself, without considering the actual news. To

assess the reliability score of each source, they review several headlines and news stories. The classification proposed by MediaBiasFactCheck has been further integrated with a number of other publicly available sources (see [23] for more details), identifying a total of 4417 domains. In the process of developing this domain classification, we also assessed the language representation of the classified web-domains to take into consideration the intrinsic multilingual and multicultural aspects of our analysis. As demonstrated by Gallotti et al. [23], comparing web traffic statistics for the different countries and focusing on the top 50 most visited websites, the domains we classified match the top-tier websites across several countries in different native languages, suggesting the reliability of the results for a comprehensive multilingual and multicultural analysis.

The URLs within the tweets were automatically scrutinized using this domain list, and each URL was categorized based on its source type, such as political, satire, mainstream media, science, conspiracy/junk science, clickbait, fake/hoax. Naturally, a fraction of domains were not present in our list: the most frequent cases being ((i) shortened URLs for which the original link was not possible to reconstruct) and (ii) all the web-domains that have not been classified by external experts.

Building on the approach outlined in [23], we classified news sources as reliable (belonging to the science or mainstream media categories), low-risk (satire, clickbait, political, other), or unreliable (fake or hoax, conspiracy or junk science), as shown in Figure 1. The distinction between web-domains, which represent the actual URLs pointing to specific news media sources in the messages, and the categorized types within each domain are pivotal in our analysis, which wants to study, at the same time, the macro and microscopic perspectives of the phenomenon of news "consumption" considering both the aggregation in the news categories and the individual domains.

### 2.3. Building the Information Networks

Each user is characterized by a sequence X of web-domains and news media categories shared during the time considered (see Figure 2). Based on these sequences, we reconstructed the undirected weighted networks where the nodes are web-domains and news media types, and the weight assigned to the edge is the frequency of co-occurrence, aggregated by the total number of users. An undirected network is a type of network where connections between nodes do not have a specific direction, meaning interactions are reciprocal between nodes. In our context, it implies that the network is built based on the subsequent posting of web-domains and news types without specifically tracking which one was mentioned first and which one second. Moreover, we quantified the number of times users iteratively shared the same web-domain and news media type, inspecting the role of *self-loops* in the network, i.e., connections where a node is linked to itself, in depth to gain a better understanding of their appeal towards certain web-domain or news media categories.

2.4. User Categories

Given a sequence of different web-domains and news media categories (see again Figure 2), we studied the news media diet of each user based on the variety of web-domains and news categories shared. We are interested in exploring whether there are any differences in the selection of specific domains and news media categories based on the type of users considered. For this reason, we distinguish users into 5 categories: reliable low-risk users (not sharing any news coming from fake or conspiracy news sources) and high-risk users (sharing at least one fake or conspiracy news piece), classified further based on their increasing engagement with unreliable sources (more than 1, 5, 10, or 100 tweets in total). In the Supplementary Materials, the same analysis is re-proposed considering more levels of engagement with high-risk sources.

News Category	Level	Definition
Science	**Reliable**	Domains providing content validated via scientific scrutiny.
Mainstream Media (MSM)	**Reliable**	Domains providing content that is generally subjected to professional fact checking and abides by the rules of media accountability.
Satire	**Low-Risk**	Domains providing content that is intentionally and explicitly aiming at providing a distorted representation of events as a form of humour and/or social critique.
Political	**Low-Risk**	Domains providing content that presents a partisan representation and interpretation of facts to support a political position over rival ones.
Clickbait	**Low-Risk**	Domains providing content that generally distorts or intentionally misrepresents information to capture attention.
Other	**Low-Risk**	Domains pointing to general content that cannot be easily classified, such as videos on YouTube.
Fake/Hoax	**High-Risk**	Domains providing manipulative and fabricated content with the purpose of misleading public opinion on socially relevant issues and provoking inflammatory responses.
Conspiracy/Junk science	**High-Risk**	Domains providing systematically manipulative and fabricated content with the purpose of legitimizing implausible conceptualizations of facts and knowledge through argumentative methods that coarsely mimic those of scientific reasoning but without any sound logical or factual basis, targeting individuals or social groups as covert instigators or perpetrators of harmful actions.

Figure 1. Classification of the news media categories based on their level of reliability based on the work of Gallotti et al. [23].
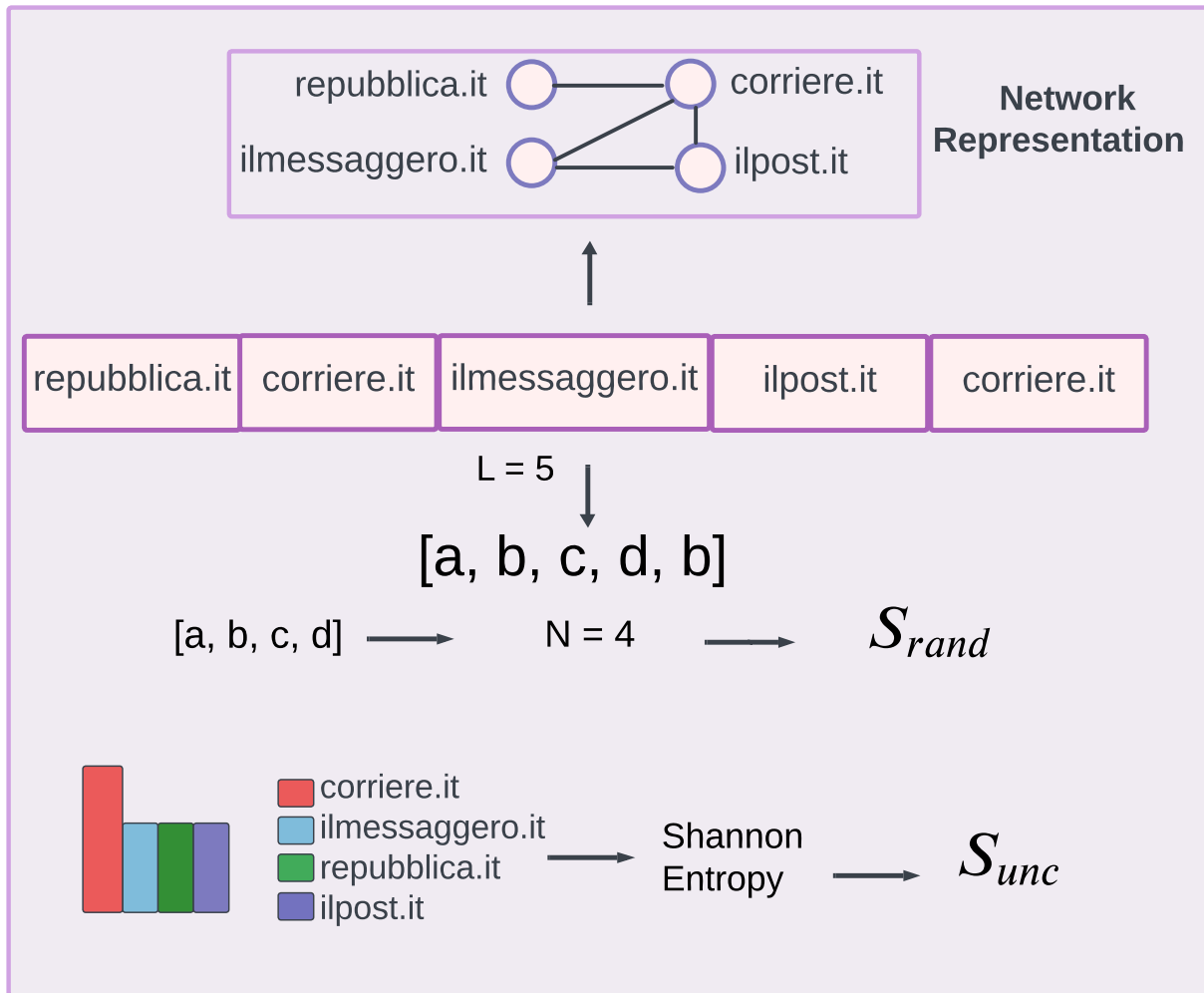
Figure 2. The News Media Diet. Network and list of the sequence of web-domains posted by a random user on Twitter. We defined the length of each sequence with the term L. Each sequence is decoded in a list of identified letters. On these lists, we calculate the entropy value to obtain information about the variety of each sequence. In particular, we performed both random entropy S_{rand} and Shannon entropy S_{unc} calculations for each sequence.

2.5. Entropy

Entropy is a key measure to quantify the uncertainty of sources of information and it has played a central role in the field of communication theory [24]. We employed three different measures of entropy to explore to what extent there is a diversity in the choice of web-domains and news media categories, considering different types of users.

First is the random entropy $S_{rand} = \log_2 N_i$, where N_i is the number of different characters in the sequence considered, in this case representing distinctive web-domains or news media types shared by user i. This measure describes the size of the ensemble of different news shared by a user and is useful to illustrate the predictability of the user's news media diet if each web-domain or news media category is shared with equal probability. Second, the temporal-uncorrelated entropy, computed as the Shannon entropy

$$S_{unc} = -\sum_{j=1}^{N_i} p(x_j) \cdot \log_2(p(x_j))$$

where $p(x_i)$ is the historical probability that a web-domain x_i was shared by user i. This second measure thus allows us to characterize the heterogeneity of the news media diet. Lastly, actual entropy, S, depends not only on the frequency of web-domains chosen but

also on the order in which each web-domain is shared, capturing the temporal order present in a user's news media diet [25]. By definition [26], these three measures should respect the following inequality

$$S \leq S_{unc} \leq S_{rand}$$

In our data, this inequality did not hold in some cases for the actual entropy S estimated using the Lempel–Ziv algorithm, which is an accurate estimator for the real entropy of strings sufficiently long [27,28]. However, this is not always the case in our data. For this reason, we decided to take into account only users who had at least three different web-domains and news media categories in order to exclude flawed entropy estimates from our analysis.

To complete our analysis, we also calculated the number of repetitions of web-domains and fact types over the total number of potential couples over the length of each sequence in order to explore the weight of these repeated sequences of web-domains and news media categories shared by each user.

More specifically, given a sequence X of length L, where $X = [x_1, x_2, \ldots, x_n]$, we counted the number of repeated pairs $x_j = x_{j+1}$ over the total number of potential pairs $L - 1$ to gain insights into the behavioral patterns in the sharing of information, considering different type of users.

3. Results

We consider the sequence of messages containing a URL and the corresponding news media category for more than 211,493 user accounts which have posted at least two different original tweets during 2020 (Figure 2). Each sequence has length L and has a specific number N of unique domains posted. In this work, we decided to remove all users having posted only a unique web-domain during the period considered. These user accounts represent 24% of our sample, leaving 160,228 unique users. In a second moment, as discussed above, we also decided to remove all users who shared only three different types of web-domains or news media categories in order to preserve the inequality of the three entropy measures [26], leaving almost 18,000 unique users.

We first characterize the sequences. Figure 3 shows the probability distribution function of the sequence length L of the (panel A) web-domains, considering the entire dataset (Total), the dataset containing only those users who have two or more unique domains shared ($N > 1$), and the different types of users we introduced in the Methods Section. Panel B shows the distribution of the number of tweets associated with each news media category in our dataset. The results confirm that the mainstream media and "other" categories are the main news media categories shared. Panel C reveals the distribution of the number of tweets associated with the different news types shared by one particular user who had shown a great activity during the time considered.

This user was classified as *reliable/low risk* since he/she did not post any URL associated with conspiracy and fake news sources.

To give an overview of the news media environment as a whole, we reconstructed the undirected network of interactions among the web-domains shared by each user, where the nodes are represented by the web-domains and the weight is given by the number of times those couples of web-domains have appeared in the sequence of posting. The network has been built at a coarser granularity having media categories as nodes in order to describe the inter-relations of web-domains of the different categories posted. In Figure 4, panel A shows that it is more likely to have posted a message containing a mainstream media source and news with a strong political bias, while panel B displays a heatmap quantifying the mean value of the relative value of the weight among the different news media categories.
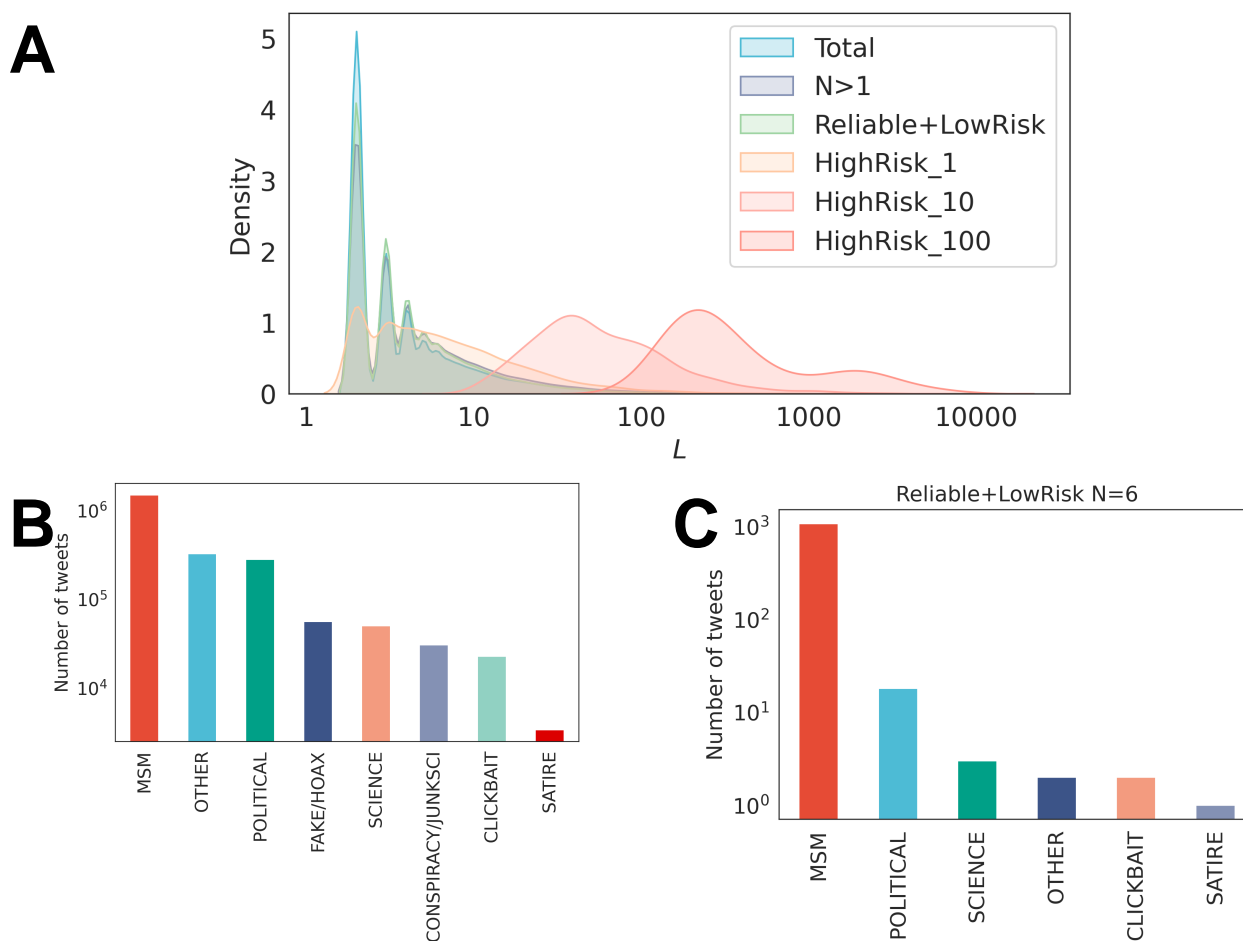
Figure 3. Overview of the dataset. (**A**) Distribution of the length sequence of web-domains, respectively, for the entire dataset, for users having posted more than 2 distinctive web-domains (N > 1) and users defined according the categories of reliability. (**B**) Distribution of the number of messages classified in one of the eight categories of news media types. (**C**) Distribution of the news media categories used by a particularly active user who has posted reliable and low-risk content, with 6 distinctive categories of news.

It is expected that links connecting the larger categories (MSM, political, and other) would have larger weights and would appear dominantly in the network we built. To compensate the effect size of the web-domains distributed across the different news categories, we compare our findings with a null model that takes into account the relations normalized by the proportions of news that belong to each category, assuming a random sequence of news shared. Differently from the previous analysis, we found that, after compensating for the category dimensions, a strong tendency to alternate between fake and conspiracy news emerges among users, as shown by panel C and by the corresponding values presented in panel D. Differently, the relationship between the category science and the other categories is more equally distributed, as shown in panel D.

In light of this strong relation between fake and conspiracy news, we refer to the users who have posted fake or conspiracy news as *high risk*, as already introduced in the Methods Section.

Moving from this first result, we decided to gain better insight into the intra-relations among domains and news media classes, also called *self-loops*. The analysis displayed in Figure 5) reveals that the distribution of mainstream media is much broader with respect to the other classes when considering the web-domains chosen. Panel A highlights that users sharing MSM articles are more likely to be further sharing other articles from the same news category than users sharing other categories of media. Differently, the category "other",

encompassing general content that is not easily classified, such as videos on Youtube or posts on Instagram, shows that the percentage of domains chosen is limited to a few within this category. Conversely, panel B shows that users sharing MSM have a lower tendency to repeatedly share the same particular domain, while repetitions of the same domain are more frequent for sequences of conspiracy theory or clickbait content.
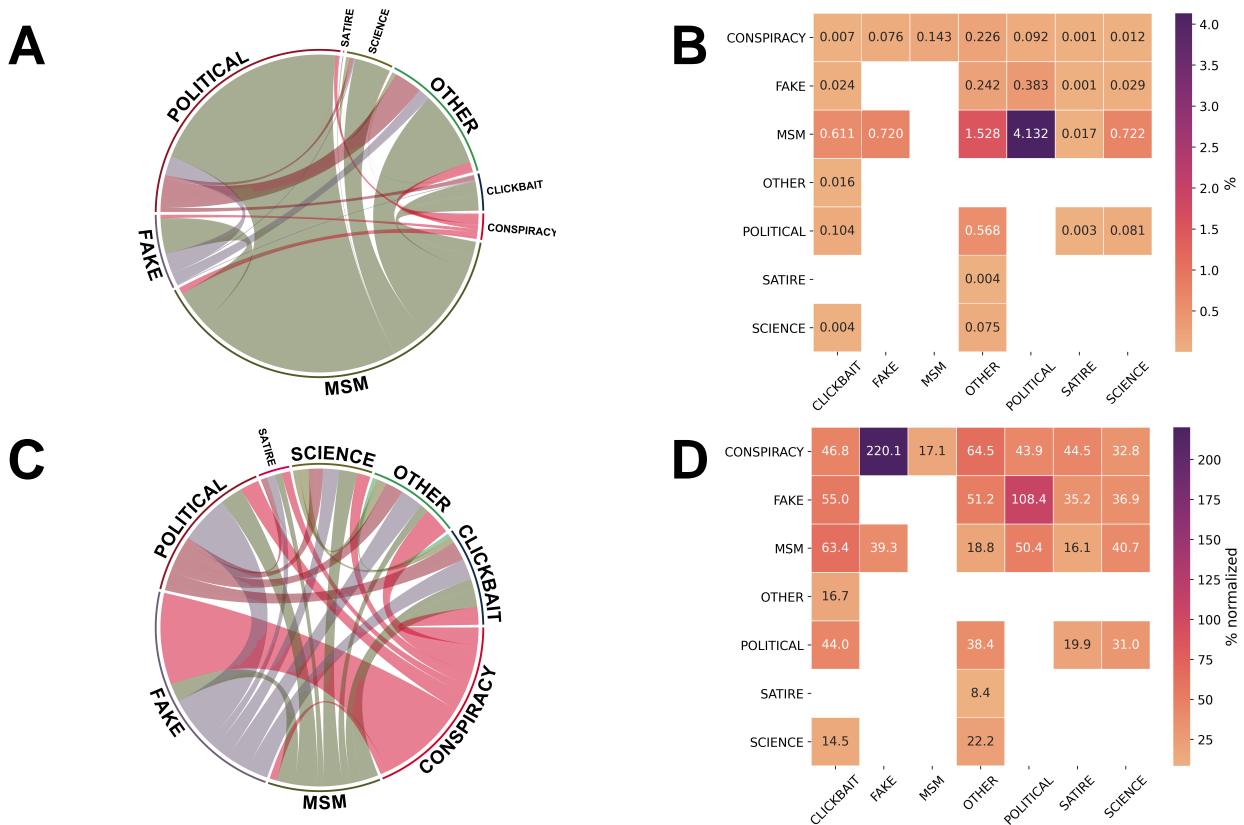
Figure 4. The news media environment. (**A**) Weighted networks of the interactions among different news media types (also known as intra-relations) for 25 countries in 2020. (**B**) Heatmap showing the corresponding value of the inter-relations among the eight categories. (**C**) Weighted networks normalized and compared with a null model accounting for the proportion of the number of news pieces belonging to each category. The chances of sharing conspiracy and fake news by the same users is much higher than the strong relation observed between mainstream media and political news (**A**). (**D**) Heatmap showing the corresponding normalized value of the inter-relations among the eight categories.

Beyond this analysis about the general patterns of news sharing in online media, our focus extends to a more detailed exploration of users' news consumption habits. In this regard, we classified each user into distinct groups: those not sharing fake and/or conspiracy news and those showing different levels of unreliability (*high risk*, 1, 5, 10, and 100) based on their increasing engagement with high-risk news (see Methods). In this context, we assigned to each user the three measures of entropy to better understand the news media diet and to gain insights into the potential differences between different kinds of users.
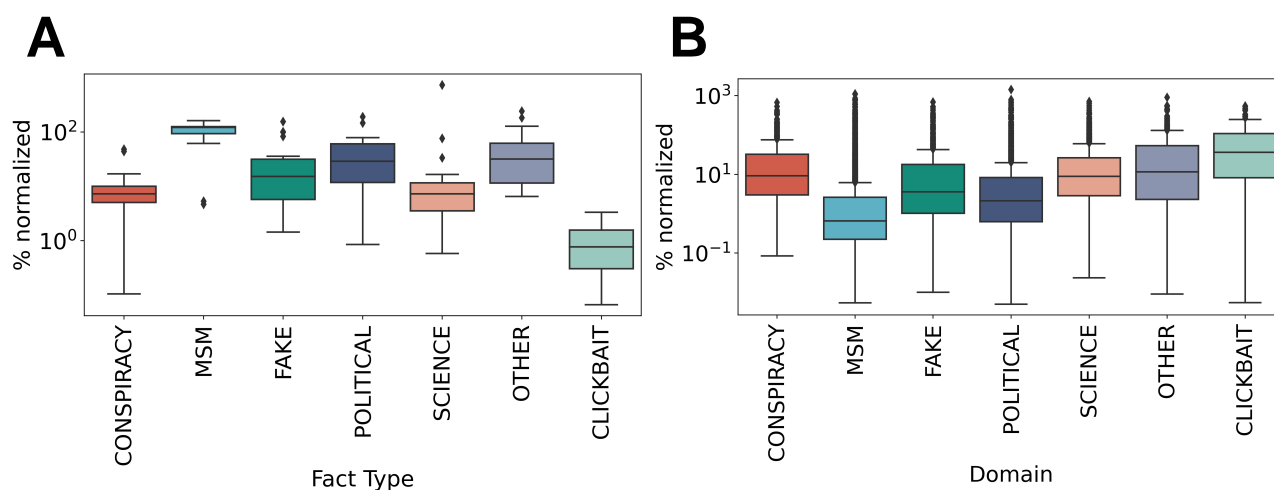
Figure 5. Self-loops of web-domains and types of news shared. (**A**) The percentage distribution of the number of self-loops for different users, pointing to the same web-domains for all the categories considered. (**B**) The percentage distribution of the number of self-loops of different users pointing to the same category of news, regardless of the different web-domains shared.

Figure 6 shows the distribution of the random entropy S_{rand} (panel A), the compression rate between the uncorrelated entropy S_{unc} and the random one S_{rand} (panel B), and the compression rate between S and S_{rand} for different types of users (panel C). To validate our results, we calculated the proportion of web-domains and news media types classified as high risk relative to the total within their respective classes. This allowed us to extract a control sample of random web-domains with a similar size to that of high risk domains. We also tested that our results are not influenced by any potential selection biases associated with the definition of our user categories by randomly reshuffling the news categories in our tweet ensemble and re-assigning users in these shuffled categories. This process resulted in a constant value of entropy across all categories and confirms absence of bias. These mean entropy values obtained from reshuffled samples (indicated by the dashed lines) are consistently lower than those of the four high-risk groups, especially for the S_{rand}, which we interpret as a measure of variety in the news media diet. Users sharing reliable or low-risk content seems to have a more restricted news media diet not only when considering the news media categories chosen but also when looking at the number of domains shared. Differently, users classified as high risk show higher values of random entropy as their engagement with these types of content increases, indicating that the consumption and sharing of high-risk content can be associated with wider exploration of a larger number of different news categories and domains shared.

The compression rate between S_{unc} and S_{rand} appears relatively flat for all the types of users considered, but here the pattern for high-risk users is more compressible than the random expectations. This tendency is even stronger in the third panel (C), which displays the compression rate between S and S_{rand} as a measure of regularity of the news media diet. In this case, the compression rate for high-risk users is much lower than the second panel (B). This indicates that the same heavy consumers and spreaders of high-risk content that we see have a tendency to explore a larger number of news categories and domains, displaying at the same time a stronger regularity in their choice of web-domains and news media categories. To validate our results, we also performed the same analyses by differentiating our users into verified and unverified users as well as taking into consideration a wider classification of high-risk users, demonstrating that the same pattern is recognizable in such scenarios, as shown in the Supplementary Materials (Figures S1–S3).
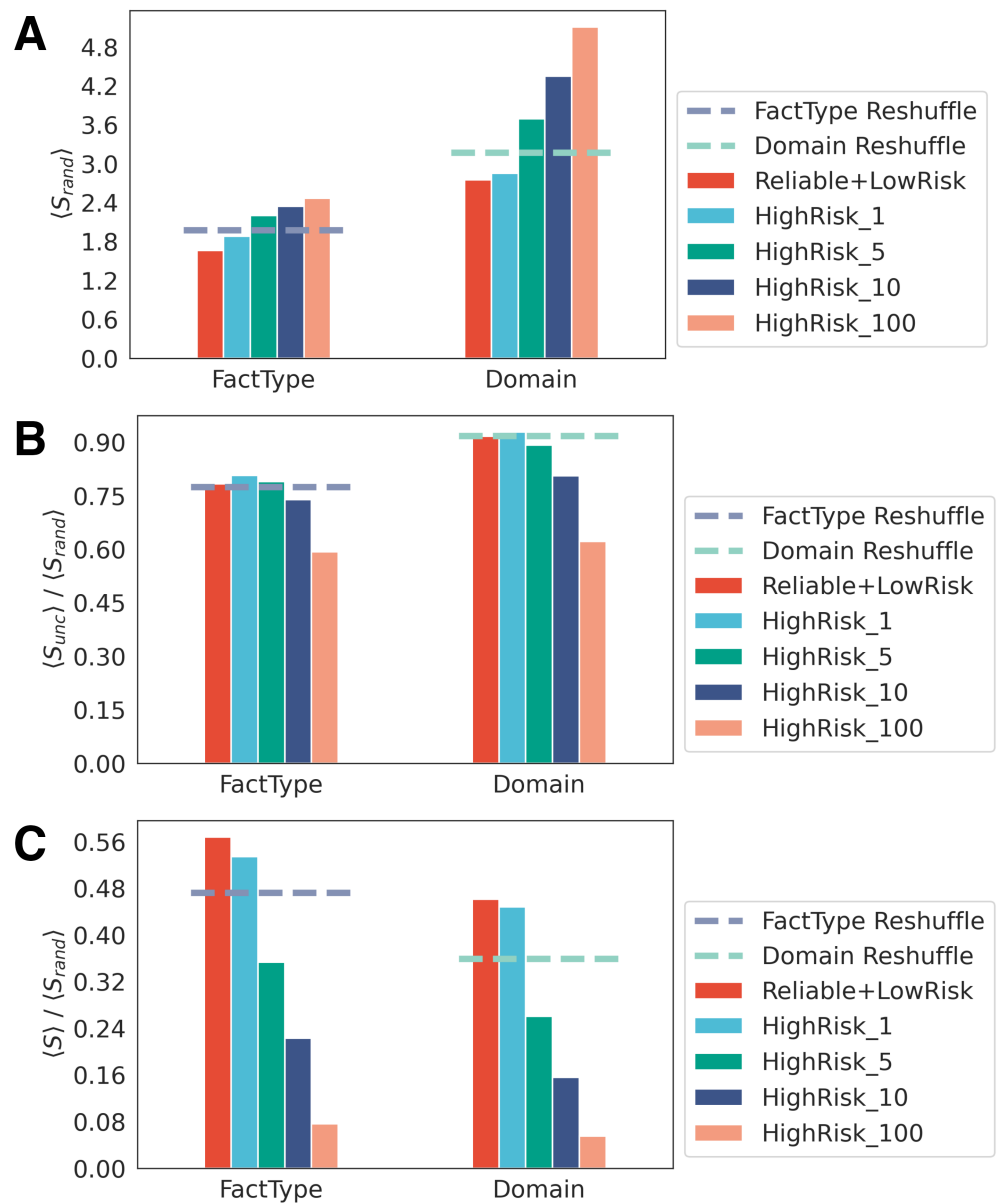
Figure 6. News media diet for different types of users. (**A**) Random entropy S_{rand} calculated for different types of accounts: users posting reliable or low-risk content and users posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. (**B**) Shannon entropy S_{unc} calculated for the domain and the type of news shared by different types of users accounts: users posting reliable or low-risk content and users posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content. (**C**) The actual entropy S calculated for the domain and news media categories shared by different type of users: those posting reliable or low-risk content and those posting different levels of high-risk (conspiracy/junk science and/or fake/hoax) content.

In order to gain deeper insights into this counter-intuitive finding, we study the relations between S_{rand} and the percentage of news media types and web-domains repeated, shedding more light on users' tendencies to either explore new web-domains and categories or stick to a few of them. Figure 7 illustrates the values of N of different news media category (or web-domains) shared against the percentage of news media repeated for reliable/low risk and high-risk users, as defined in the Methods Section. Panel A demonstrates that as the number of news media types increases, the percentage of fact types repeated grows. The last distribution characterizing the users with the highest value of S_{rand} entropy are the ones with a higher number of mainstream media and political journals repeated. The expectation

of repeating the same fact-type assuming a homogeneous distribution is instead expected to be decreasing for growing N, as it is trivially given by the ratio $1/N$, highlighted by the dashed line in all panels. The distribution of repeated news media categories for users posting six unique news media categories therefore displays a different tendency and is significantly higher than the random expectation for larger N. However, the same behavior is not shown when looking at users posting high-risk content, as shown in panel B. Indeed, the distribution of the percentage of news media types shared is slightly higher for users with a lower S_{rand}. The last distribution is represented by people having posted all the possible news categories ($N = 8$) Besides the news media classes, we analyzed the percentage of web-domains repeated with respect to the number of unique domains shared for reliable or low-risk users (panel C) and high-risk users (panel D). In this case, the distribution of domains repeated shows to be fixed at 25 %, independent from the size of $\langle N \rangle$, both for reliable and low-risk and high-risk users. Interestingly, this result indicates that the percentage of repeated domains is the same regardless of the type of user and the number of domains or news media categories shared.
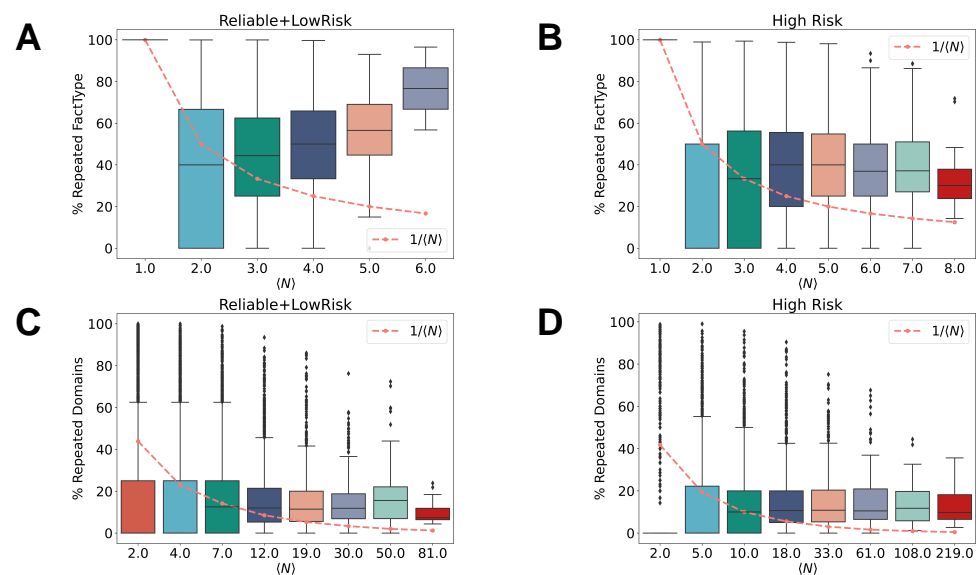
Figure 7. Fraction of repeated news media categories and web-domains for users with different N. In these boxplots, we represent the percentage of news media categories (**A**,**B**) or web-domains (**C**,**D**) repeated in subsequent tweets with respect to the total length of the sequence associated with an individual user. Each boxplot is aggregated over users with the same N. The dashed line shows the random expectation of repeated news media categories with respect to the increasing number of N unique news media categories. (**A**) Patterns of news media categories, aggregated for users classified as having posted only reliable and/or low-risk content on Twitter. There are 8 box-plots here since we have 6 distinctive categories and 2 are filtered out. (**B**) Patterns of news media categories, aggregated for users having also posted high-risk content. There are 8 box-plots here since we have all 8 distinctive categories of news media types. (**C**) Patterns of web-domains, aggregated for users classified as having posted reliable and/or low-risk content. (**D**) Patterns of web-domains, aggregated for users classified as having also posted high-risk content in their sequence of messages posted.

4. Discussion

Our research reveals that the news media digital landscape is a complex environment where users tend to be attracted to and thus repeat specific web-domains or news media categories. In particular, when we inspected the news media environment, we found that there is a strong relation between the probability of sharing mainstream media journals and political media. This relation is just apparent because if we compare our results with a null model that takes into account the effect size of each category, our analysis highlights how users sharing fake or hoax news are more likely to alternate with conspiracy and junk

science types of news, illustrating the tendency for unfortunate misinformation hot streaks. This result also suggests how easy it is to fall into this vicious circle where alternating fake and conspiracy news becomes the norm. Differently, the news categories representing the most reliable ones (mainstream media and science) do not show this strong relationship, leaving the most unreliable categories at the center of the news media digital landscape. The habit of alternating between fake news and conspiracy theories suggests a strong inclination towards reinforcing one's own opinions, often by seeking out narratives that align with pre-existing beliefs in the digital environment. This phenomenon reflects a broader trend of confirmation bias [29], wherein individuals gravitate towards information that confirms their worldview while disregarding contradictory evidence. In the digital scenario, where an abundance of information is readily accessible and algorithms personalize content recommendations based on user preferences, individuals may find themselves immersed in echo chambers where their perspectives are constantly reinforced [30]. This echo chamber effect can amplify the consumption of unreliable and misleading narratives, perpetuating the misinformation hot streaks. Furthermore, the sensational and provocative content can make fake news and conspiracy theories particularly appealing to certain audiences, evoking strong emotional responses and exploiting cognitive biases [31], such as the availability heuristic or the illusion of explanatory depth, which can lead individuals to accept them uncritically. The tendency of alternating fake and conspiracy news has raised the need for better understanding which are the characteristic features of each user, specifically when we consider different types of users. In this scenario, one might suggest that different types of users, based on their interactions with different type of content, have completely different news media habits. This is what our research shows.

In this context, we introduced two different measures to characterize and quantify the users' behavior in the way they access and retain information online, shaping their news media diet. The *variety* in the so-called news media diet is given by S_{rand}, while the *regularity* is calculated by the compression rate between S and S_{rand}. By employing these measures of entropy, we found that reliable or low-risk users have more restricted news media habits when searching for news media categories and web-domains. On the contrary, high-risk users at different levels reveal having a more varied news media diet, especially for users sharing more than 100 high-risk web-domains. Interestingly, the ratio between S_{unc} and S_{rand} demonstrates that reliable or low-risk users have less compressible sequences if compared with high-risk users. This is even more explicit when we observe the compression rate between the real entropy S and the random entropy S_{rand}, which highlights how high-risk users display a remarkable regularity in their news consumption, even if they are characterized by having a more diverse diet if we just consider the random entropy, which takes into account only the number of unique web-domains and news media categories. Getting deeper into this study, we found that the repetitions of news media categories is in general higher than expected if we look at the users characterized by a huger variety of news media categories selected. Interestingly, this effect is stronger for reliable and low-risk users. In that context, we observed that those users are usually also repeating more often political news media categories and represent particularly active users. On the contrary, the repetitions of web-domains seems independent to the variety of the news diet, whether the user is a reliable or low-risk or high-risk one, showing that there is the same probability of sharing web-domains, independently of how much the user has wider news media habits.

This work has shed light on news media behavior and the patterns of the users during highly contentious events, such as the COVID-19 pandemic, revealing how much certain types of users have more varied but at the same time regular news diets as they are more likely to share and to repeat the same news media sources. In particular, we found that this happens for users heavily sharing high-risk disinformation sources. At the same time, we also observed from an aggregated perspective how the news media digital environment is shaped according the strong relationship of fake and conspiracy news: highly debated and unprecedented topics can generate debates driven by fake and conspiracy news, which can

monopolize the attention of particular users. These users with a greater tendency to engage with this high-risk content are also those more likely to have a wider content base but are also the ones more likely to express a strong regularity in the way they share news media content online over time.

Our findings are based on an aggregated sample of countries having at least 500 tweets per day, showing an aggregated and global tendency. For the purpose of comparison, we decided to pick some countries with lower and higher values of press freedom in order to observe if certain trends change by considering only specific countries. As shown by comparing Supplementary Figures S5 and S6 and Supplementary Figures S7 and S8 (see Supplementary Materials for more details), this is not the case; we observe a consistent pattern across countries, indicating a significant presence of misinformation and conspiracy narratives within their respective media landscapes. The process of collecting data and integrating different sources of user data provides us with the opportunity of analyzing complex human behavior phenomena. However, it is important to acknowledge that our study does not claim to be complete. Indeed, we decided to analyze only Twitter data, which might contain population biases. Twitter is indeed an online social network mainly used by well-educated males (63% of Twitter users [32]) between the ages of 18 and 34 (56% of Twitter users, according to Statista [33]). Further research might explore the news media environment and diet of users in other online contexts, highlighting differences and similarities among several social media platforms. These findings underscore the imperative for a more nuanced comprehension and proactive analysis of the ramifications of online social systems, alongside a discerning evaluation of the offline risks engendered by the dissemination of streaks of misinformative content within the digital sphere. This is certainly a topic that merits additional analyses and holds significant potential for enhancing our comprehension of the news media habits of users during contentious debates.

References

1. Lazer, D.M.; Baum, M.A.; Benkler, Y.; Berinsky, A.J.; Greenhill, K.M.; Menczer, F.; Metzger, M.J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. The science of fake news. *Science* **2018**, *359*, 1094–1096. [CrossRef]
2. Zarocostas, J. How to fight an infodemic. *Lancet* **2020**, *395*, 676. [CrossRef]
3. Osatuyi, B. Information sharing on social media sites. *Comput. Hum. Behav.* **2013**, *29*, 2622–2631. [CrossRef]
4. Westlund, O.; Färdigh, M.A. Accessing the news in an age of mobile media: Tracing displacing and complementary effects of mobile news on newspapers and online news. *Mob. Media Commun.* **2015**, *3*, 53–74. [CrossRef]
5. Nielsen, R.K.; Schrøder, K.C. The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison. *Digit. J.* **2014**, *2*, 472–489. [CrossRef]

6. Westerman, D.; Spence, P.R.; Van Der Heide, B. Social media as information source: Recency of updates and credibility of information. *J. Comput.-Mediat. Commun.* **2014**, *19*, 171–183. [CrossRef]
7. Mazzeo, V.; Rapisarda, A. Investigating fake and reliable news sources using complex networks analysis. *Front. Phys.* **2022**, *10*, 886544. [CrossRef]
8. Renjith, R. The effect of information overload in digital media news content. *Commun. Media Stud.* **2017**, *6*, 73–85.
9. Rodriguez, M.G.; Gummadi, K.; Schoelkopf, B. Quantifying information overload in social media and its impact on social contagions. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 170–179.
10. Koroleva, K.; Krasnova, H.; Günther, O. 'Stop Spamming Me!'-Exploring Information Overload on Facebook. 2010. Available online: https://aisel.aisnet.org/amcis2010/447 (accessed on 9 March 2024).
11. Roetzel, P.G. Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Bus. Res.* **2019**, *12*, 479–522. [CrossRef]
12. Rodrigues, U.M.; Xu, J. Regulation of COVID-19 fake news infodemic in China and India. *Media Int. Aust.* **2020**, *177*, 125–131. [CrossRef]
13. Salaverría, R.; Buslón, N.; López-Pan, F.; León, B.; López-Goñi, I.; Erviti, M.C. Desinformación en tiempos de pandemia: Tipología de los bulos sobre la COVID-19. *Prof. Inf.* **2020**, *29*. . [CrossRef]
14. Marietta, M.; Barker, D.C. *One Nation, Two Realities: Dueling Facts in American Democracy*; Oxford University Press: Oxford, UK, 2019.
15. Roozenbeek, J.; Van der Linden, S. Fake News Game Confers Psychological Resistance against Online Misinformation. *Palgrave Commun.* **2019**, *5*, 65. [CrossRef]
16. Brennen, J.S.; Simon, F.M.; Howard, P.N.; Nielsen, R.K. Types, Sources, and Claims of COVID-19 Misinformation. Ph.D. Thesis, University of Oxford, Oxford, UK, 2020.
17. Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science* **2019**, *363*, 374–378. [CrossRef] [PubMed]
18. Shahi, G.K.; Dirkson, A.; Majchrzak, T.A. An exploratory study of COVID-19 misinformation on Twitter. *Online Soc. Netw. Media* **2021**, *22*, 100104. [CrossRef]
19. Caldarelli, G.; De Nicola, R.; Petrocchi, M.; Pratelli, M.; Saracco, F. Flow of online misinformation during the peak of the COVID-19 pandemic in Italy. *EPJ Data Sci.* **2021**, *10*, 34. [CrossRef]
20. Dutta-Bergman, M.J. Complementarity in consumption of news types across traditional and new media. *J. Broadcast. Electron. Media* **2004**, *48*, 41–60. [CrossRef]
21. CoMuNe. COVID-19 Infodemics Observatory. 2020. Available online: https://covid19obs.fbk.eu/#/ (accessed on 13 January 2024).
22. Media Bias/Fact Check. 2020. Available online: https://mediabiasfactcheck.com/ (accessed on 13 January 2024).
23. Gallotti, R.; Valle, F.; Castaldo, N.; Sacco, P.; De Domenico, M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat. Hum. Behav.* **2020**, *4*, 1285–1293. [CrossRef]
24. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
25. Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343. [CrossRef]
26. Song, C.; Qu, Z.; Blumm, N.; Barabási, A.L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021. [CrossRef]
27. Wyner, A.D.; Ziv, J. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inf. Theory* **1989**, *35*, 1250–1258. [CrossRef]
28. Gao, Y.; Kontoyiannis, I.; Bienenstock, E. Lempel-Ziv and CTW entropy estimators for spike trains. In Proceedings of the Estimation of Entropy Workshop, Neural Information Processing Systems Conference (NIPS), Vancouver, BC, Canada, 8–13 December 2003.
29. Ling, R. Confirmation bias in the era of mobile news consumption: The social and psychological dimensions. *Digit. J.* **2020**, *8*, 596–604. [CrossRef]
30. Törnberg, P. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS ONE* **2018**, *13*, e0203958. [CrossRef]
31. Boyer, M.M. Aroused argumentation: How the news exacerbates motivated reasoning. *Int. J. Press/Politics* **2023**, *28*, 92–115. [CrossRef]
32. Distribution of Twitter Users Worldwide as of January 2023, by Gender (Statista, 2023). Available online: https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/ (accessed on 25 January 2024).
33. Distribution of Twitter Users Worldwide as of April 2021, by Age Group (Statista, 2021). Available online: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/ (accessed on 25 January 2024).