DE FACTO
Observatory
of Information

DIGITAL MAPPING OF THE FRENCH MEDIA ECOSYSTEM

Maxime Crépel, Benjamin Ooghe-Tabanou, Kelly Christensen, Béatrice Mazoyer, Guillaume Plique, Jean-Philippe Cointet, Sylvain Parasie, Dominique Cardon, Katharina Tittel, Antoine Machut

Sciences Po médialab - DEFACTO - 02/23/24 –final version

SciencesPo AFP CLEMI Le centre pour l'éducation aux médias et à l'information WIKI

Co-funded by the European Commission under the Connecting Europe Facility (CEF) - Telecommunications Sector (Agreement No. INEA / CEF / ICT / A2020 / 2394372)

DIGITAL MAPPING OF THE FRENCH MEDIA ECOSYTEM

1. Introduction

The mapping of the French media ecosystem aims to produce a technical infrastructure to monitor the media landscape's news information and circulation dynamics. We produced a database comprising a large sample of major authoritative news media and alternative sources. Using a global and structural approach, we analyzed the relationships between the French media sources regarding their authority, ideological position, and audience. We also used De Facto's enriched database of fact checks, the statistical and topological properties of Twitter accounts regularly sharing misinformation to measure their activity and visualize their position in the media ecosystem. First, we present the protocol and the methodology used to produce the dataset, then we describe the general structure and the properties of the French media ecosystem with a specific focus on disinformation.

2. Methodology

a. Protocol and dataset

To select the media sources, we first extracted a list of the 10,000 most shared domain names on twitter[1] over a period of 12 months, from the 01/01/2022 to the 12/31/2022. We defined the most shared domain names by sorting them with the production of a normalized "twitter score" composed by three quantitative criteria defining their visibility, the number of users who shared them on twitter and the regularity of their presence on twitter:

- Sum of tweets: sum of all tweets including a link towards the domain name over the period.
- Number of accounts: number of unique Twitter user accounts who shared (meaning who posted, retweeted or quoted) at least one tweet including a link towards the domain name over the period.
- Median activity: median total of tweets including a link towards the domain name each month.

We have chosen a broad definition of media, considering all French websites whose content covers news in various fields as politics, culture, sport, technology, science, etc. (see annex 1 for the list of qualitative criteria). We double-blind coded a randomized sample of 500 domain names and calculated the intercoder measure[2] to control the robustness of our definition of media. Then, we manually selected a list of 747 media sources corresponding to our definition in the list of the first 3,000 domain names sorted form the highest "twitter score" to the lowest.

Using médialab's web crawling tool Hyphe[3], we crawled during the summer of 2023 these 747 websites with a depth of 2 clicks from the homepage and aggregated all hyperlinks between the 2,210,949 collected webpages. We also collected for each website a set of metadata related to the topology, the notoriety, the administrative status, and the diffusion of misinformation (see annex 2).

Aggregating all links between the websites two by two, we produced a network of the 732 websites included in the core

[1] The list is based on a data collection of French-language tweets including a hyperlink, representing an average of 3 million tweets per day.

[2] For Fleiss' kappa coefficient alpha = 0.73 and for Krippendorff's alpha coefficient alpha = 0.73

[3] Jacomy M., Girard P., Ooghe-Tabanou B., et al, (2016) "Hyphe, a curation-oriented approach to web crawling for the social sciences.", in International AAAI Conference on Web and Social Media. Association for the Advancement of Artificial Intelligence. https://hyphe.medialab.sciences-po.fr/

component[4] connected by 27,556 hyperlinks. We visualized the network using the spatialization algorithm ForceAtlas2[5] within the Gephi[6] software. We used the Louvain algorithm[7] implemented in Gephi to detect the main clusters separated by modularity classes in the network. We qualitatively annotated the clusters to interpret the main media categories which appear in those clusters. Finally, we projected the metadata on the network to describe the characteristics of the media ecosystem.

The database, the network file and high-resolution images are available on the Sciences Po research data repository: https://doi.org/10.21410/7E4/VMMY7L

b. Limitations of web mapping techniques

The web exploration and mapping methods enable the analysis of the relational structure among actors. These techniques are based on the hypertext links extraction from the crawled pages of websites. Two websites are considered connected if at least one hypertext link points to one another within the corpus. However, these methods have certain limitations which need to be explicated.

Firstly, data collection through a web crawler is not exhaustive and does not guarantee the retrieval of all hypertext links within a website. The number of links collected by the crawler depends on the definition of the crawl depth and the architecture of the website. Considering technical capabilities and the extent of media websites which can be very large, we chose a common depth of 2 for all crawls, resulting in the collection of between a few hundred web pages and up to more than 50000 web pages for some of the biggest medias, providing a snapshot of the main pages of each media website along with the articles from a few weeks before the date of the crawl.
Other technical issues can also limit the capacity of the crawler to collect hypertext links on some websites, such as server inaccessibility during extraction, paywalls, popup windows or technical infrastructures (such as full Javascript websites) that may prevent the crawling process for certain websites. Such limitations imply that the topological analysis of node positions, which depends on their connectivity to the rest of the actors in the network, as well as the analysis of incoming and outgoing hypertext citation links, may vary depending on these technical constraints.

The network visualization methods also raise some limitations. The tension between actors due to the links connecting them generally leads authoritative websites to position themselves at the center and less connected actors on the periphery of the network. Similarly, when applying a community detection algorithm, some nodes at the border between two clusters may sometimes be associated with a neighboring cluster because they maintain numerous links with the actors in that subset.

[4] A set of 15 medias are isolated from the core component due to the absence of links collected from and to their websites.

[5] Jacomy M, Venturini T, Heymann S, Bastian M (2014) ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9(6): e98679. https://doi.org/10.1371/journal.pone.0098679

[6] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. From AAAI. https://gephi.org/

[7] Blondel V., Guillaume J.L., Lambiotte R., Lefebvre E., (2008), "Fast unfolding of communities in large networks", Statistical Mechanics: Theory and Experiment (10), P1000.

3. Mapping of the French media ecosystem

a. Topology and thematic analysis

The media ecosystem analysis (figure 1) reveals that the most connected media websites (emitting and receiving the most hypertext links) occupy a central position within the network. Thus, national media websites such as BFMTV, FranceTVInfo.fr, Le Figaro, Le Monde, 20minutes, Libération, etc. logically position themselves as significant players in the media landscape. We also observe the presence of two prominent local press websites (PQR), Ouest France and Le Parisien, which also hold central positions in the network. Considering the web audience data published by ACPM[8] in 2023, these central websites are also the ones with the largest audiences in 2022 (e.g., Lefigaro.fr has an average of 141 million unique visitors per month, Bfmtv.com of 140 million visitors, Francetvinfo.fr of 130 million visitors, Lemonde.fr of 128 million visitors, etc.).

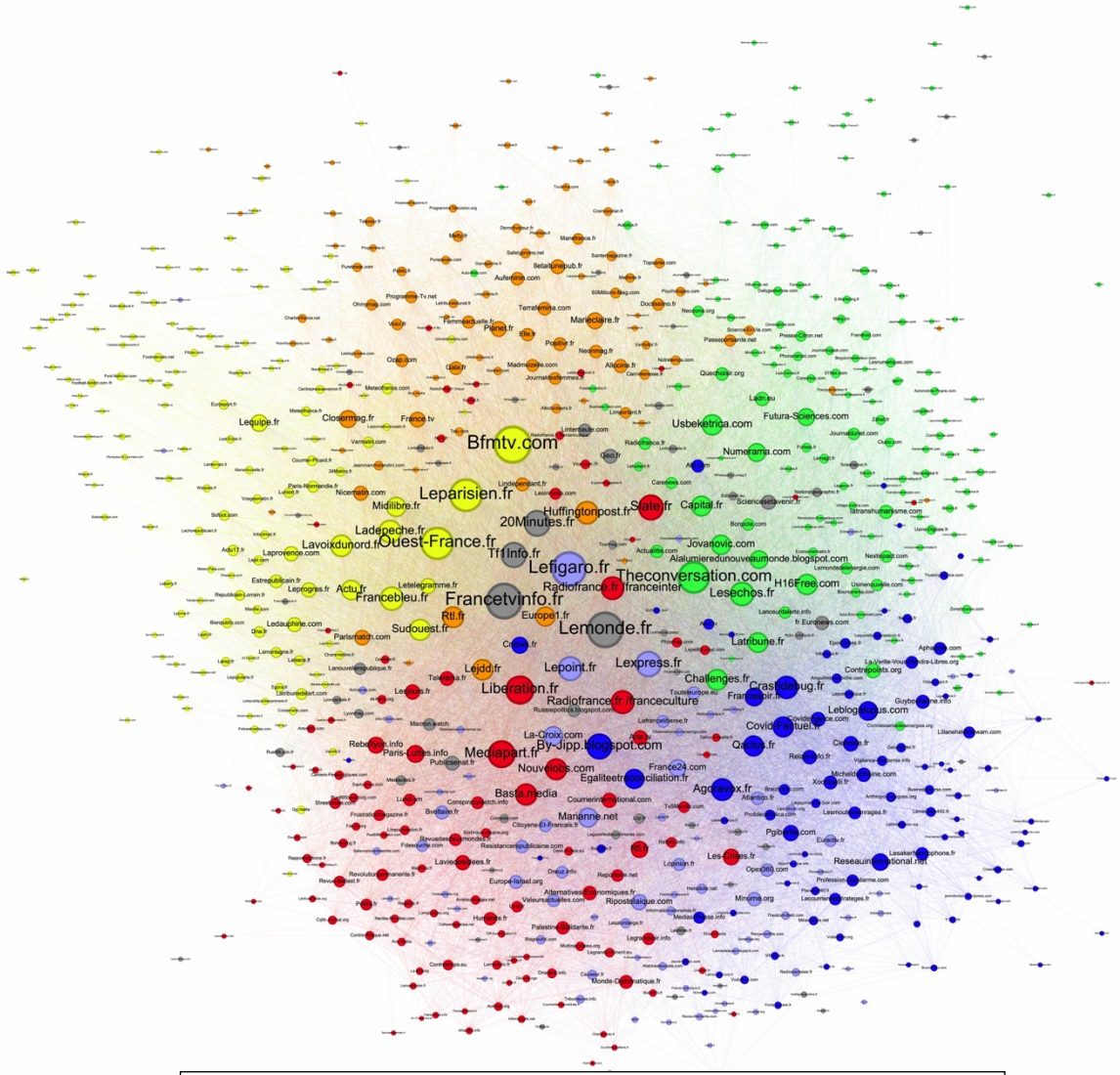

Figure 1: Network of French media websites, node size is relative to the degree, colors correspond to the clusters detected by the Louvain algorithm.

[8] https://www.acpm.fr/Les-chiffres

The French media ecosystem is structured into seven main clusters. These clusters represent areas detected by the Louvain algorithm due to the high density of links between websites. If we consider the main websites within each cluster of the network, we can easily interpret the topics and sub-topics featured in those clusters.

By separating the clusters and applying a contraction algorithm to avoid overlapping effects, we obtain a clearer view of the different topics while preserving the overall topology of the network (figure 2):

- **The cluster #1** in yellow at the top-left of the network is composed of media outlets primarily covering local news (PQR), as well as media specialized in sports and weather forecasts.
- **The cluster #2** in orange at the top-center of the network is composed of media outlets specialized in so-called “celebrity magazine" and “lifestyle” press, as well as health press, including media associated with active forums in the French web landscape.

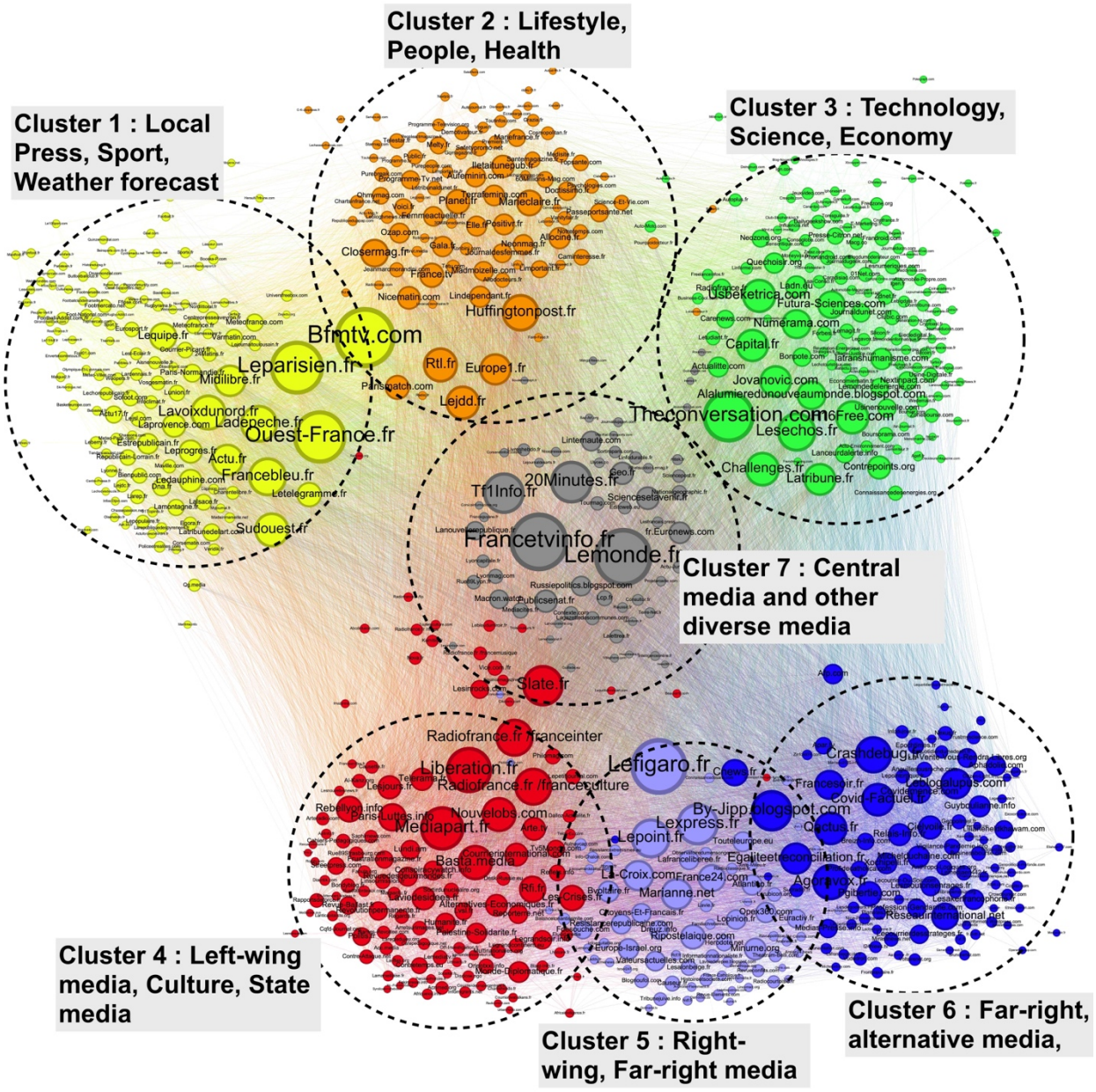

Figure 2: Exploded view of the network of French media websites, node size is relative to the degree, color corresponds to the clusters detected by the Louvain algorithm.

- **The cluster #3** in green at the top-right of the network is composed of media outlets specializing in the fields of technology, science, and economics.
- **The cluster #4** in red at the bottom-left is mainly composed of left-wing news media, media specialized in culture, and a few public media outlets.
- **The cluster #5** in blue at the bottom center is primarily composed of center and right-wing news media and far-right news media.
- **The cluster #6** in dark blue is composed of news media associated with the far-right and websites known as "alternative media" such as “reinformation“ or “counter-information”.
- **The cluster #7** in gray at the center of the network is an aggregation of small clusters composed of mainstream media and other diverse media which the community detection algorithm identifies as strongly linked to each other but do not constitute a major cluster within the media space.

b. Political news media analysis

The media primarily covering political news, social issues, and economic news (56% of the medias) are predominantly concentrated in the lower part of the network, while media specialized in other themes tend to appear in the upper part of the network (figure 3).

If we filter the network to keep only political news websites, we can again spatialize the network based on hyperlinks between these media outlets. We observe the same concentration effect of media with the largest audiences at the center of the network, while less visible media are distributed at the periphery.

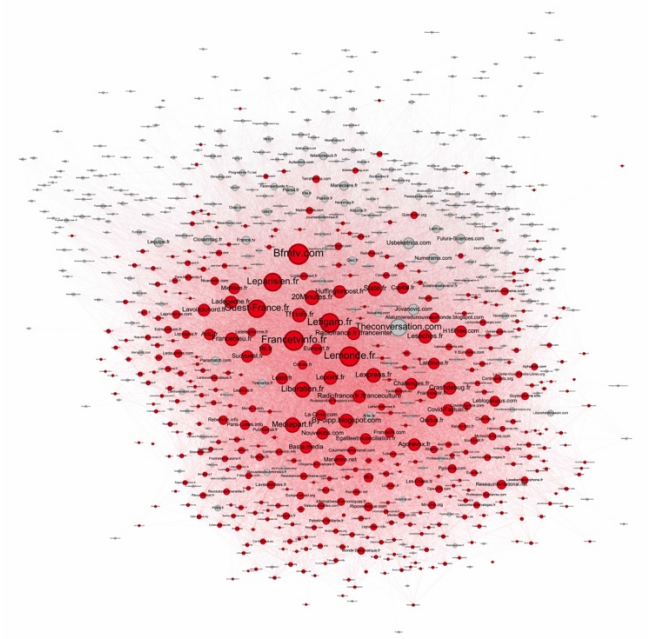

Figure 3: Network of French media websites; node size is relative to the degree. Nodes focusing on political, economic, and societal topics are highlighted in red.

By applying the Louvain algorithm again to detect clusters on this subset, we obtain a network that polarizes news media around three main clusters. This new network, composed of media focused on political, economic and social news, offers several layers of interpretation (figure 4). Firstly, there is a distinction between mainstream media with high visibility at the center of the network and other media at the periphery. Secondly, a clear distinction is noted between local press in the upper part of the network and national press in the center of the network. Thirdly, there is also an axis from left to right linked to the political orientations of the media.

The left part of the network is mostly populated by media with an editorial line politically oriented to the left of the political spectrum. This is the case, for example, for media outlets such as L’Humanité, Libération, Mediapart, Bastamedia, etc.

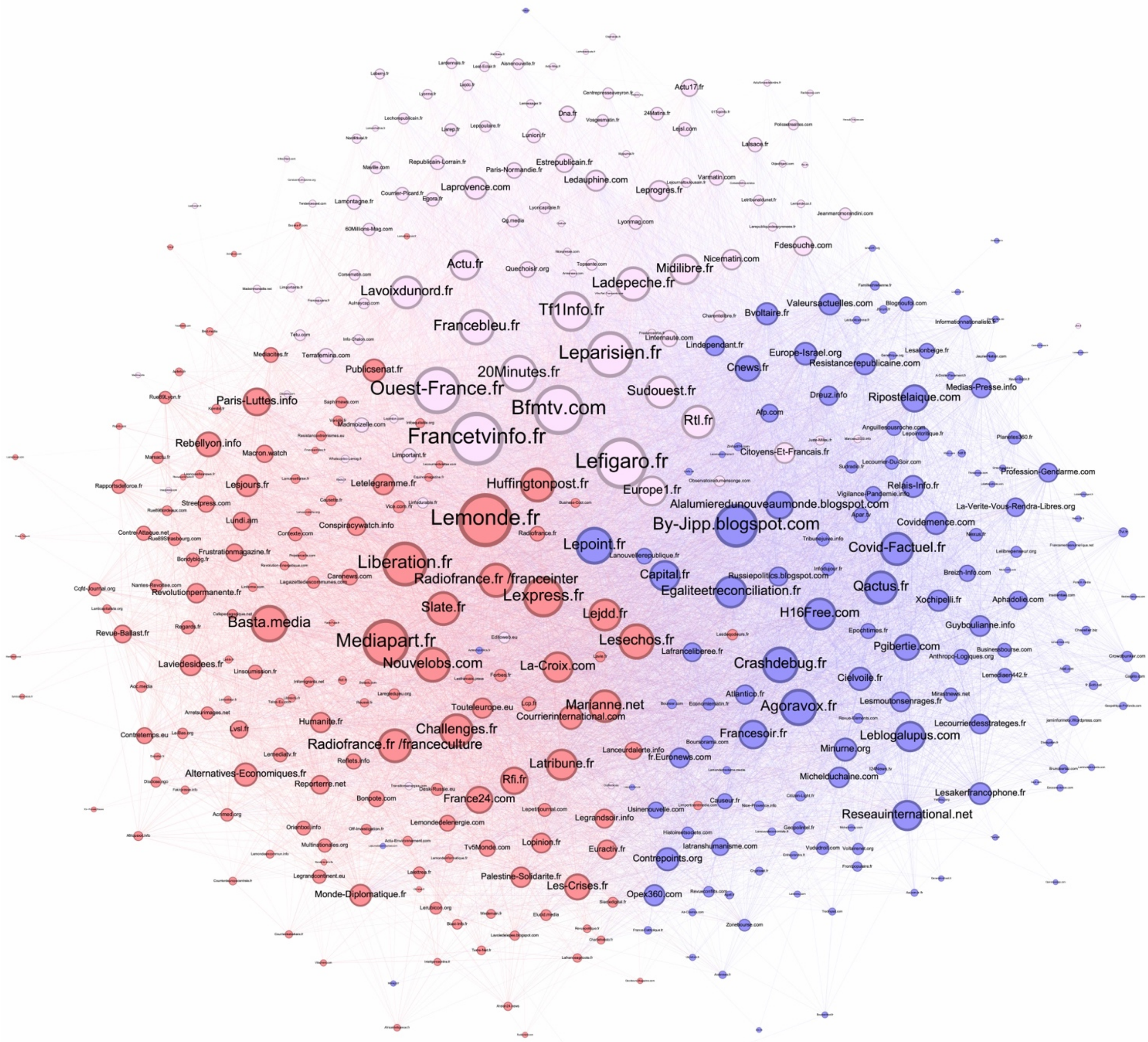

Figure 4: Network of French media websites covering political news, node size is relative to the degree, color corresponds to the clusters detected by the Louvain algorithm.

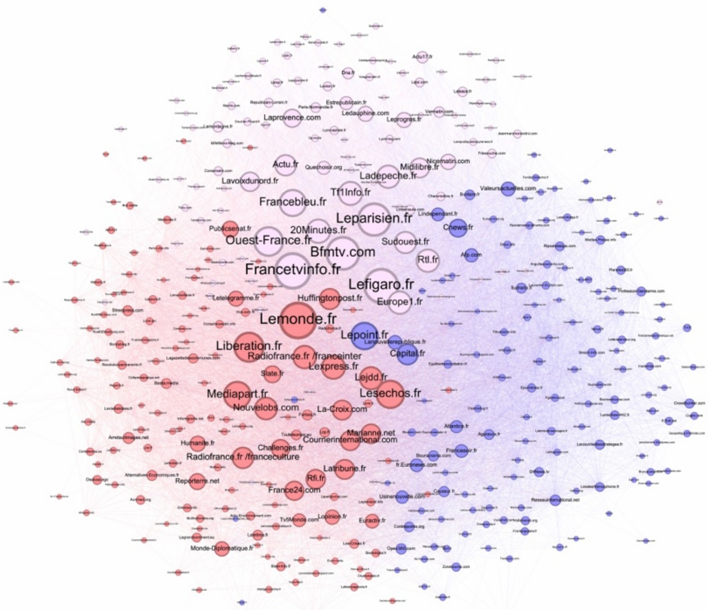

Figure 5: Network of French media websites covering political news, node size is relative to the in degree, color corresponds to the clusters detected by the Louvain algorithm.

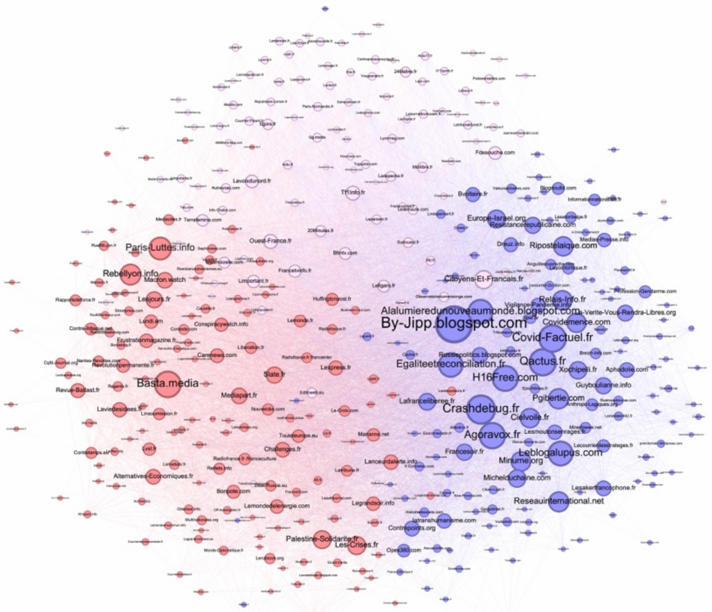

Figure 6: Network of French media websites covering political news, node size is relative to the out degree, color corresponds to the clusters detected by the Louvain algorithm.

The right part of the network is mainly composed of right-wing and far-right media, such as Valeurs Actuelles, CNews, Egalité et Réconciliation, RiposteLaïque, as well as numerous alternative medias websites such as Qactus.fr, Covid-factuel.fr. At the center of the network, we see mainstream media with different political orientations, such as Le Monde, Le Figaro, FranceTVInfo, etc. The analysis of the topology allows us to observe central positions and identify websites which are the most cited and authoritative within the French web media landscape.

The websites receiving the highest number of incoming links are central in the network and constitute the most visible websites in terms of audience and activity on Twitter (figure 5). In contrast, websites emitting the most outgoing links to other media are in the lower part and on the left and right periphery of the network (figure 6). The network structure seems to be characterized by two major properties: a strong hierarchical structure in terms of topology, audience, and visibility on social networks, and a form of asymmetry between mainstream professional medias and alternative medias, creating a separation between the center and periphery.

C. Disinformation and fact-checking mapping

Based on the three fact-checking databases available in our corpus (De Facto's fact-checks database[9], Facebook's Condor database[10], and Science Feedback's database of fact-checked articles[11]), we projected onto the network the number of fact-checked articles for each media in the database and projected the results onto the network. We observe that the two areas most impacted by the reporting of articles are the mainstream media in the center of the network and media on the right side of the network, mainly composed of far-right and alternative media. Some media outlets on the left side of the network also appear to be producing articles that have been flagged for fact-checking. However, when we focus on the fact-checking results, we note that media outlets whose information has been categorized as accurate are located in the left and central part of the network. The sources publishing articles based on information fact-checked as false are mainly concentrated in the right-side area at the network's periphery. This is the same area we previously highlighted as having lower visibility regarding audience and citation links from mainstream media.

References

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. From AAAI.

Blondel V., Guillaume J.L., Lambiotte R., Lefebvre E., (2008), "Fast unfolding of communities in large networks", Statistical Mechanics: Theory and Experiment (10), P1000.

Jacomy M., Girard P., Ooghe-Tabanou B., et al, (2016) "Hyphe, a curation-oriented approach to web crawling for the social sciences.", in International AAAI Conference on Web and Social Media. Association for the Advancement of Artificial Intelligence.

[9] https://defacto-observatoire.fr/download/Comprendre/Le-consortium-DE-FACTO-publie-un-format-pivot-pour-les-articles-de-fact-checking/WebHome/defacto-fact-checking-pivot-format-1.0.pdf?rev=1.1

[10] https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one

[11] https://science.feedback.org/

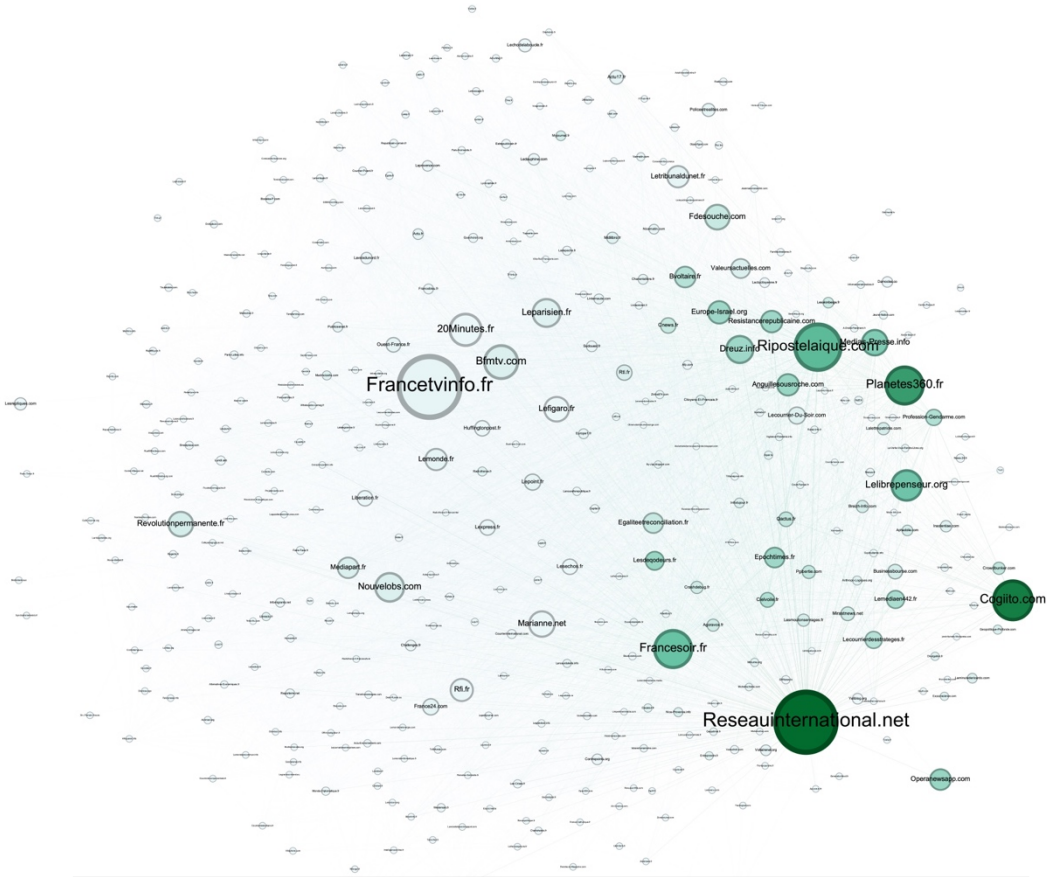

Figure 7: Network of French media websites covering political news, node size is relative to the number of factcheck cases, colors correspond to the number of factchecks categorized as "false"

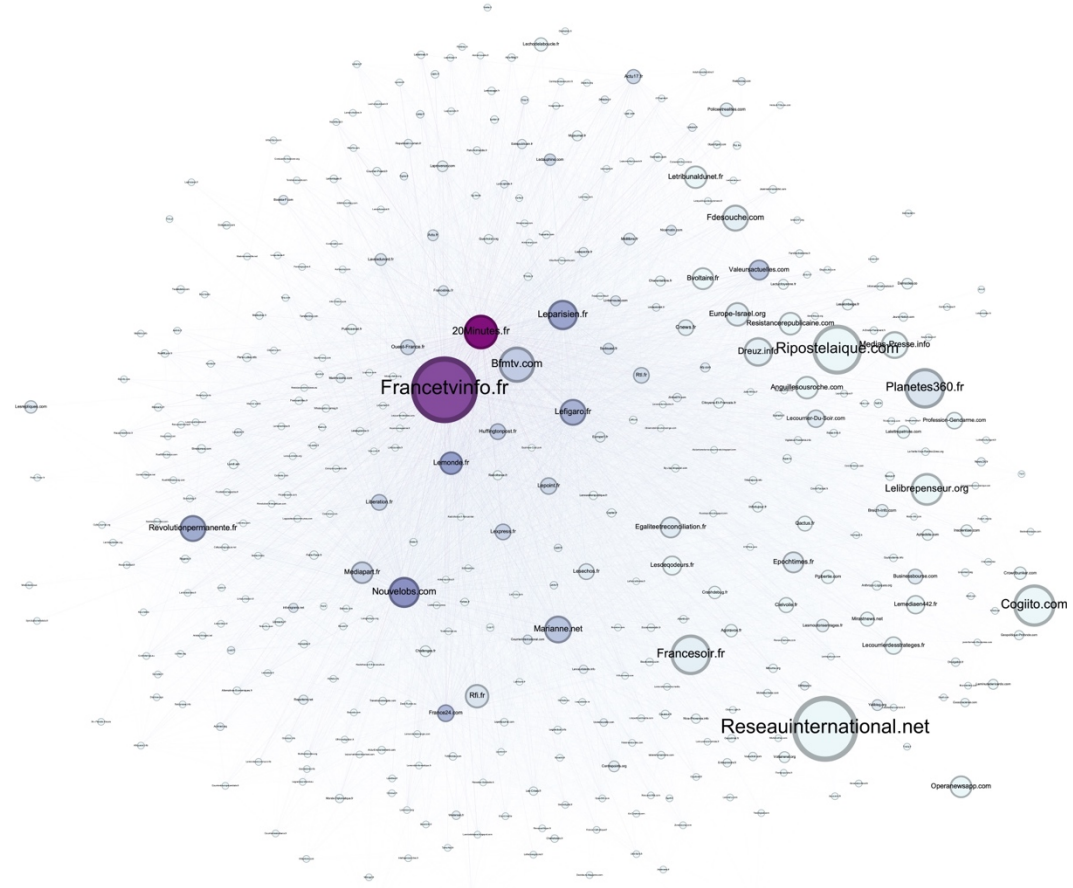

Figure 8: Network of French media websites covering political news, node size is relative to the number of factcheck cases, colors correspond to the number of factchecks categorized as "true"

Annex 1: Qualitative criteria for the media sources selection

1-Inclusion criteria

Geographic criteria:
-French news / in France
-International media with French outlets (I24news, RTFrance)
-French media focusing on foreign news
-National media
-Local media (PQR)

Content type
-General media (French or foreign news)
-Specialized media including all types of content coverage (e.g. economics, technology, sports, culture, literature, art, science, music, lifestyle, people, professionals, weather, satire, parody, etc.)
-Institutional media (e.g. LCP, Public Sénat, metropolitan media)
-Citizen media (Agoravox, etc.)
-Personal news media (blogs, personal news sites)
-Media associated to forum functions (Jeuxvideo.com, Doctissimo, Aufeminin)

Format type
-Text (print, web)
-Video (video, tv)
-Audio (radio, podcast)

Actor defintion
Self-defined media in the “about” section of the web site

2-Exclusion criteria

-Inactive website or 404 at time of selection
-Search engine website
-Social network platform
-Blog platform
-News aggregator website
-Association and NGO website (not news-oriented / e.g. 60millionsconsommateur LeMag)
-Media observatory website (not news-oriented - e.g. OJIM)
-Media archive website (Ina, bnf (retro news))
-Media company websites (which do not publish news - e.g.: francetélévision.fr is not francetvinfo.fr)
-VOD channels and programs (tf1.fr)
-Commercial website
-Institutional website
-Political party website, campaign website, political personality blogs (even if there is a news section)
-Foreign media website (not focusing on french information)
-Scientific publication and preprint website (such as cairn, hypothèse ...)

Annex 2: Database description

variable_name	defintion
id	unique ID
label	name of the media
indegree_gephi	in degree in gephi network
outdegree_gephi	out degree in gephi network
degree_gephi	sum of in degre and out degree in gephi network
modularity_class	id of the cluster defined by Louvain algortihm
poltics_eco_society	media coverage about politics, societey and economy
sum_tweet	sum of tweets cited the name domain over the period
twitter_user	sum of unique userwho shared the name domain over the period
median_activity	median score of tweets cited the name domain each month
twitter_score	agregated score of normalized sum_tweet and twitter_user and median_activity
legal_status	presence of the media in th CPPAP database (Commission paritaire des publications et agences de presse)
editor	editor of the media
legal_status_detail	detail of the legal status
departement	localisation of the media ZIP code
qualification_status	presence in the list of media qualified as general and politics information (IPG) /39 bis A / 39 bis B) CPPAP database
qualification_detail	detail of qualification status in the CPPAP database of media qualified as general and politics information (IPG) /39 bis A / 39 bis B) CPPAP database
science_factcheck_case_status	presence of factcheck cases in the science feedback database
defacto_factcheck_case_status	presence of factcheck cases in the defacto database
condor_factcheck_case_status	presence of factcheck cases in the condor database
n_science_false	number of cases classified as false in science feedback database
n_defacto_false	number of cases classified as false in defacto database
n_condor_false	number of cases classified as false in condor database

total_factcheck_false	total number of cases classified as false in the three database
n_science_true	number of cases classified as true in science feedback database
n_defacto_true	number of cases classified as true in defacto database
n_condor_true	number of cases classified as true in condor database
total_factcheck_true	total number of cases classified as true in the three database
n_science_cases	number of cases reported in science feedback database
n_defacto_cases	number of cases reported in science feedback database
n_condor_cases	number of cases reported in science feedback database
total_factcheck_cases	total number of cases in the three database
youtube_chanel_url	youtube channel of the media
twitter_account_all	twitter account of the media
facebook_account	facebook account of the media
instagram_account	instagram account of the media
tiktok_account	tiktok account of the media
twitch_account	twitch account of the media
total_known_pages	total known pages by hyphe crawler
crawled_pages	total crawled pages by hyphe crawler
prefixes_as_url	prefixes used by hyphe crawler
home_page	homepage of the web entity
start_pages	start pages used by hyphe crawler