Mapping opinion landscapes: analyzing network structures in climate change debates on Twitter

Author: Javier Castillo Uviña*
Master en Física dels Sistemes Complexos i Biofísica.
Facultat de Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.†

Advisors: Dra. Luce Prignano and Dr. Emanuele Cozzo
Dra. Prignano Postdoctoral Researcher at Complexity Lab Barcelona and
Dr. Cozzo Senior Researcher, Universitat de Barcelona Institute of Complex Systems - UBICS
(Dated: September 25, 2023)

Abstract: Society is a complex system, and studying it is challenging. We need to develop mechanisms to obtain data for statistical analysis. Our goal is to collect a substantial amount of data on a subject while also understanding how human opinion is structured and how it evolves over time, providing us with additional insights. With the advent of online social platforms, we have the opportunity to study how users generate content (data) and how others interact with this content. This presents us with a perfect tool and opens up an entire universe of possibilities.

In this project, we aim to characterize opinions on the issue of climate change using Twitter data [1]. The focus of this research lies in developing a method to study the structure of human opinions. To achieve this, we will analyze three Twitter discussions where users are expected to take positions on climate change-related policies, express their opinion about COP Meetings, and discuss the creation of the 2030 Agenda among Catalan and Spanish speakers. By capturing these data based on user opinions, we will conduct a study by analyzing the networks that shape these opinions and interactions. Through this analysis, we hope to uncover the underlying structure of a Twitter discussion within the framework of complex networks. Additionally, we will attempt to identify if distinct communities are formed and who the most influential accounts are in each case. We aim to assess the polarization of the network to determine if there are two clear sides [2]. On one hand, we anticipate users who support or oppose the climate change issue, and on the other hand, we aim to identify a "denialist" side that opposes the concept of climate change. Expanding our study, we can also track the evolution of public opinion. Nevertheless, we must acknowledge certain limitations. Not everyone uses Twitter, resulting in incomplete representation of all communities. Furthermore, such platforms do not necessarily mirror real-world interactions, yet they do offer a robust reflection of public opinion beyond the screens.

The results have been less encouraging in terms of polarization. For most of the cases studied, we observed a dominant structure that does not exhibit clear polarization for or against the idea of climate change. However, we did identify another noteworthy type of structure worth discussing.

In conclusion, the hypothesis that the "denialist" side possesses significant enough support to disrupt other types of structures has not proven accurate. Nonetheless, we discovered that these debates are often steered by highly influential *hubs*. Analyzing these *hubs* can provide us with a solid understanding of the current opinion landscape.

I. INTRODUCTION

Since the appearance of social platforms, our capacity to communicate and disseminate information or opinions has evolved. There are pros and cons surrounding this concept of social platform. On one hand, we overcome geographical barriers, the information travels very quickly. Additionally, a lot of data is created and stored. Nowadays, we can virtually live study the society on a new scale. On the other hand, people sometimes lose their sense of reality. Besides, there is no universal regulation or supervision on what is shared. This fact implies the viralization of fake information or extremist opinions. This generates a hate speech that tries to undermine the importance of relevant issues, such as climate change in our case. It can affect to the real world and how society evolves.

As a society, we are a complex system. Studying a social platform data is a powerful tool to gather a lot of information. From a social system, we can understand how information diffuses, what tendencies exist and social and political polarization. We will focus on the latter.

On social platforms, users act as individuals with their own thoughts and expressions. They have the power to write brief messages and share them with others. Within this social platforms, a web of relations exist between users, statuses and punctual interactions.

The two kinds of relations in these social studies are:

- **States**: "Static" relations between two nodes. On Instagram, a *follower*, or a *friend* on Facebook. This relations usually remains over time, unless the users decide to change their state.

* jcastiuv7@alumnes.ub.edu
† master.complex.biophys@ub.edu

- **Events or interactions**: These are more dynamic actions. A *like*, a *comment*, *sharing* and *retweeting* are the actions we are refering to. These are specific actions at a given time, and may or may not repeat among the same profiles.

Complex networks are formed from nodes that represent any particle of our system and links that represent the relationships between these nodes. In this project we will call indistinctly network or graph to the complex networks. This formalism provide a powerful framework for studying and analyzing various interconnected systems, including social interactions. The complex network formalism is beneficial due to its possibility of: representing interactions, observing the structures and patterns that form when modeling public opinion based on interactions between people, identifying influential nodes, giving a visual representation of a complex system and exploring the temporal evolution of public opinion.
The main objective of this project is to follow a guideline for studying of a Twitter discussion: from searching data to unravel the discussion's structure. This involves a temporal analysis to determine when it is relevant to download data, data downloading methods, methods and metrics for studying complex network structures, creating the complex network from the graph and analyzing the discussion bases on the discovered metrics. Another concrete objective of this project is to provide an overiew of how public opinion on climate change evolves among Catalan and Spanish speakers. We have divided this climate change discussion into four topics: Economy, Activism, Politics and Climate. We will focus on the political arena with the goal of identifying a denialist community.

II. FROM PLATFORM TO DATA: EXTRACTING INSIGHTS FROM TWITTER

As of today, Twitter stands out as the most efficient platform for propagating opinions. Its current format of 280 characters for *tweets* provides a means to deliver concrete and concise information.
A close relationship exist between Twitter and traditional media. Viral content on the platform can sometimes transform into news, while news itself becomes content for discussions on Twitter. Both platforms amplify each other's presence.
On Twitter, the static connections are formed through following other profiles; this relation need not be mutual. The occasional interactions include *likes*, *comments*, *sharing*, *retweeting* and *saving* a tweet. While private conversations are possible, we do not have access to this information. To construct our network, we will represent users who tweet as individual nodes, with interactions between them depicted as edges or links. We will not consider the static relationships between users.

To gather data for mapping, Twitter provides its own API.

A. Twitter Data

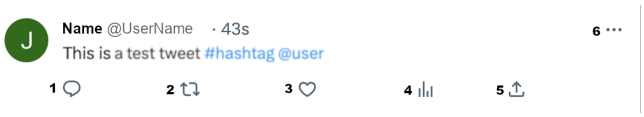


FIG. 1. Dummy tweet with a hashtag (#), a mention (@) and the other interactions numbered from 1 to 6.

In FIG 1, we can observe the components of a tweet:

- The green bubble represents the default picture of the user, which can be customuzed. The letter inside is assigned according to the chosen screen name.
- The "Name" is what we want to display to other users, also refered to as the *Screen Name*.
- @UserName must be unique as it serves as the account identifier.
- The total text size, including hashtags and mentions, must not exceed 280 characters.
- A hashtag is a word or phrase without spaces, preceded by #, it is used to quickly locate a topic and identify the tweet's subject.
- The number "4" indicates how many users have seen this tweet.

These are the interactions we can have with a tweet:

- Mentioning another user using the format @user. We can also mention ourselves.
- Using "1" to add a comment to this tweet.
- "2" has two uses: first, it is for showing this tweet to our followers (a *retweet*); second, it is for showing the tweet content but adding a comment (a *quote*).
- "3" stands for the *like*. This button indicates agreement with the user's tweet and signals our appreciation.
- Button "5" is used to share the tweet by copying the URL or sending it as a direct message. It also provides the option to save this tweet for quick access in our bookmarks.
- "6" offers various options, including unfollowing this user, blocking them, and reporting to Twitter administrators.

We will delve deeper into the concepts of node source and node target shortly. Despite the direction of information flow, we consistently regard the node source as the user initiating interaction with the node target. For example, the node source can comment on a tweet by the node target, or retweet a tweet from the node target. This dynamic interaction is illustrated in FIG 2, where the arrow's direction indicates attention flow. Node A receives information from both Node B and Node C, and in turn, directs attention towards them. Furthermore, as a result of Twitter dynamics, a link forms between Nodes A and C, with the arrow direction influenced by attention flow. Therefore, when a node, such as A, takes action on another node's tweet, like B's, A functions as the source node, and B as the target node.

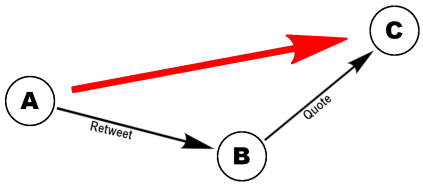


FIG. 2. We can observe a dummy graph. Node B quotes a tweet from Node C. Node A retweets this quotation, which becomes a tweet itself. As a result, a new link appears between Nodes A and C (indicated by the red arrow).

B. API: Data Retrieval

The Twitter API provides access to data offered by the company itself. Due to changes in ownership, there have been several alterations in policies. Different levels of access existed; we utilized the old academic access level. This allowed us to search for any time period since the launch of Twitter. There were no limits on how many tweets were published each day, and there was a monthly download limit of 10 million tweets' content [3].
Unfortunately, gaining API access for further research is currently not feasible for most academic researchers due to the high costs associated with this access. Consequently, conducting actual studies based on Twitter data poses challenges at present.
Various tools can be employed for processing Twitter data. We are using Twarc2 [4]. This tool is equipped with queries to search for and collect requested information. A *query* is constructed using keywords, usernames, *hashtags* and more.
It is crucial to have a clear focus when searching on Twitter due to its diverse dynamics. For information diffusion, the main interaction to consider is the *retweet*. Identifying the principal statement or keywords suffices, as most users use similar wording. After explaining how the queries are constructed, we will show the queries we have designed to make a temporal analysis of the use of these important words. They are also the same as the ones used to build the graphs that we will study.

- **Logical operators**: This queries have logic on them. NOT, OR and AND are the most common logic operators used. When we are doing a search we have to use a combination of words an this logic help us to be precise. For example an empty space between words means an AND, "happy pretty" this query will capture a tweet as follows:

  "Pretty flowers make me happy".

  The OR operator is used for searches which we want to capture any of the different terms. The NOT operator is for avoiding tweets with the word after the NOT.

- **Exact sentences**: If we want to capture the tweets with an exact phrase we have to put a \" before and after, for example \"`black cat`\" will capture:

  "I want to adopt a black cat".

  But will not capture:

  "The cat is black".

- **Extra notes**: The Twitter API does not differentiate between upper and lower case. It does not read special characters either. There exists more filters: language, it is possible to set the period time we want to focus... It is also important to having into account and it is the use of plural, "rose" and "roses" are not the same search.

The queries used in this article are:

```
"es": [
    "(((energias␣(renovables␣OR␣limpias
       ))␣OR␣(las␣renovables)␣OR(
       energia␣(renovable␣OR␣limpia))␣
       OR␣(transicion␣energetica))lang
       :es)",
    "((ecologista␣OR␣ecologistas␣OR␣
       ecologismo␣OR␣ambientalismo␣
       ORambientalistas␣OR␣
       ambientalista)␣lang:es)",
    "(((agenda␣2030)␣OR␣agenda2030␣OR␣(
       acuerdo␣de␣paris)␣
       ORacuerdodeparis␣OR␣
       parisagreement␣OR␣cop21␣OR␣
       cop26)␣lang:es)",
    "(((calentamiento␣global)␣OR␣((
       emergencia␣OR␣crisis)␣(
       climaticaOR␣sequia␣OR␣sequias))
       )␣lang:es)"
```

```
    ],
"ca": [
    "(((energies␣(renovables␣OR␣netes))
       ␣OR␣(les␣renovables)␣OR␣(
       energia␣(renovable␣OR␣neta))␣OR
       ␣(transicio␣energetica))␣lang:
       ca)",
    "((ecologista␣OR␣ecologistes␣OR␣
       ecologisme␣OR␣ambientalisme␣OR␣
       ambientalistes␣OR␣ambientalista
       )␣lang:ca)",
    "(((agenda␣2030)␣OR␣agenda2030␣OR␣(
       acord␣de␣paris)␣OR␣acorddeparis
       ␣OR␣parisagreement␣OR␣cop21␣OR␣
       cop26)␣lang:ca)",
    "(((escalfament␣global)␣OR␣((
       emergencia␣OR␣crisi)␣(climatica
       ␣OR␣sequia␣OR␣sequies)))␣lang:
       ca)"
    ]
```

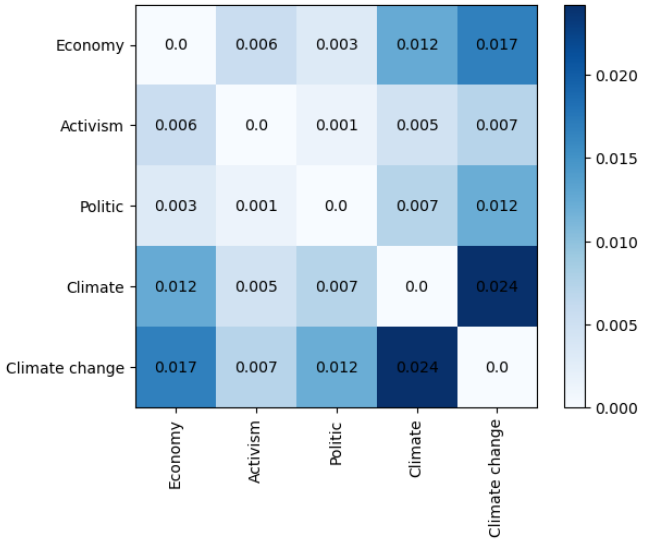
Listing 1. "es" and "ca" refer to the language, Spanish and Catalan respectively. Each language contains a list of four queries enclosed in quotation marks. The queries are: Economy, Activism, Politic and Climate respectively.

C. API: Data Collection

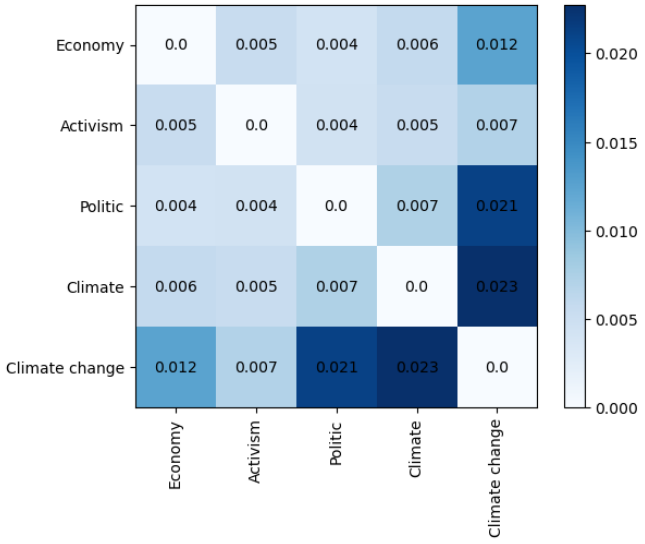
Now that we understand how to search for our data, let's delve into the various levels of data collection available through the API. Depending on the target endpoint, the API provides different formats of information. The following two endpoints are particularly significant and are the ones we have utilized for this study:

- **Counts**: This endpoint is unlimited. We can count as many tweets as needed. It is employed to uncover the temporal evolution of a query. While it provides the count of tweets per day, it doesn't offer the content or interactions. However, when a topic exhibits distinct behavior, we can use the following endpoint to delve into the details.
- **Search**: There is a download limit, necessitating precise query formulation. After executing the search, we retrieve all tweets that match the query, along with metadata related to the users and interactions between users who have also tweeted.

With all the data collected, we can construct the desired graph. It is possible to filter the interactions we want to focus on, among other actions. Our focus lies on the retweet network, while we also intend to illustrate the all interactions network. Each query will initiate a search, and from this search, we will create a graph. This graph will be composed of nodes, representing all users who have tweeted content that fulfills the query's criteria and have interacted with other nodes of the graph. The isolated nodes are no computed in the graph. Likewise, the interactions that appear are only those that exist between two nodes in the network. This is why we have to be meticulous when forming the query to collect all the tweets related to our issue.



(a) Catalan

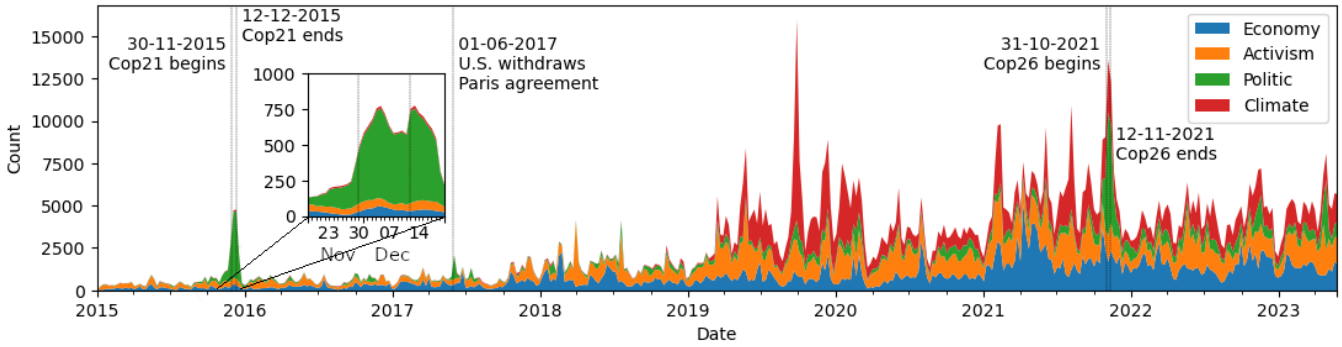


(b) Spanish

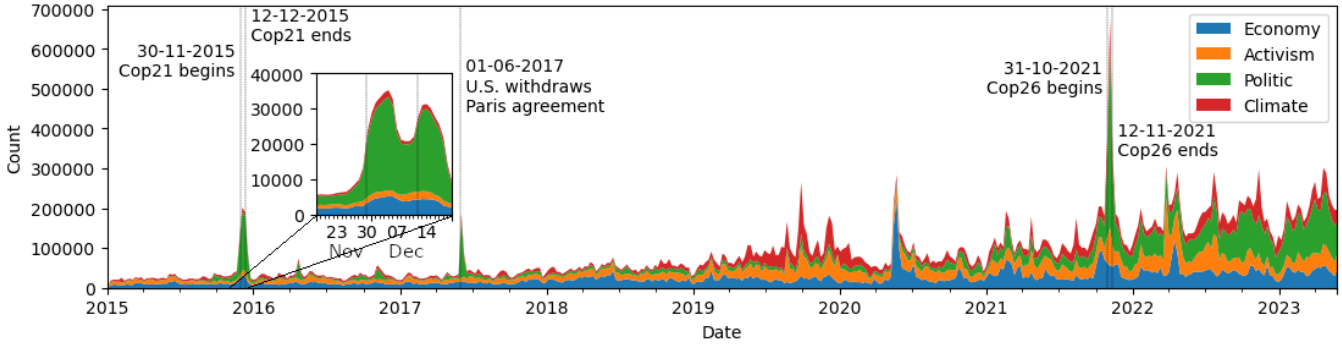
FIG. 3. Average tweet overlap for the entire time period depicted in FIG 4. The terms: Economy, Activism, Politic, Climate and Climate Change corresponds to those in the queries 1 and 2

III. TEMPORAL ANALYSIS

Our objective is to provide an overview of the climate debate landscape among Catalan and Spanish-speaking Twitter users, enabling us to compare their behaviors.



(a) Catalan



(b) Spanish

FIG. 4. Here appears the tweets counts week by week from 2015-01-01 to 2023-05-25. The different colors correspond to the different queries we designed to separate the topics involved with the climate change. We show three specific dates with a black-dotted line because they will be our study case. There is a zoom between 2015-11-20 and 2015-12-20 with a resolution day by day in order to see this double peak we can not see in the week resolution.

Through research, we identified a query that captures a significant portion of tweets related to user opinions on climate change. To gain a comprehensive perspective, we categorized these tweets into four major topics: Economy, Activism, Politics, and Climate. This categorization facilitates the relation of climate change opinions with each specific topic.

In FIG 4, the evolution of tweet counts over time is illustrated. Notably, we observe an immediate response from Twitter to real-world events, as demonstrated by the dotted lines; platform activity aligns closely with real-time occurrences.

While can neglect the tweet overlap, where the same tweet could be counted for different queries, this approximation is justified given that the average temporal overlap is less than 1%, as evidenced by FIG 3. The overlap comparison includes the overarching Climate Change topic, described by the following queries:

```
"es":["((cambio␣climatico)␣lang:es)",],
"ca":["((canvi␣climatic)␣lang:ca)",]
```

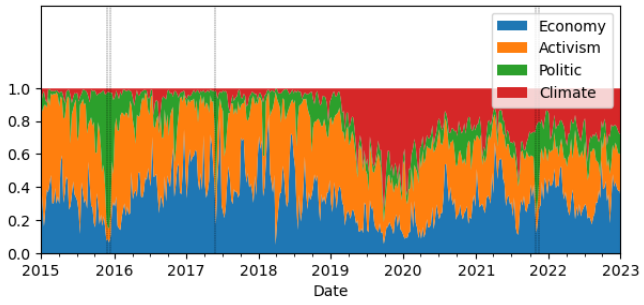
Listing 2. Climate Change queries.

Although we anticipated a higher overlap for the Climate Change label, it is reasonable to assume that when users discuss a particular topic, their focus remains on that specific topic. Consequently, FIG 4 faithfully represents the climate concerns within the Twitter community.

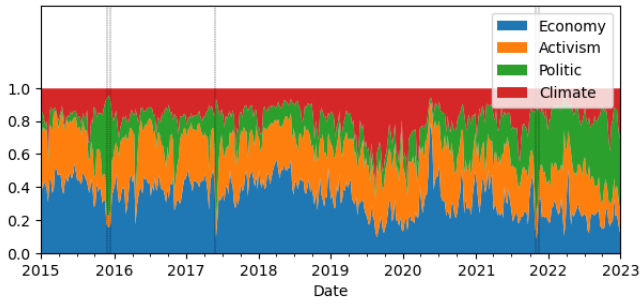
As illustrated in FIG 4, a wealth of information is extracted. Initial usage for opinion diffusion of the platform was modest until 2017. Moreover, at that time Twitter was in full decay due to limited usability. The platform's growth trajectory shifted in 2017 when the character limit was expanded from 140 to 280, enhancing the popularity of the platform [12]. Both for everyday use and for a more formal use as it is nowadays, where Twitter is a platform where great personalities and media publish relevant information. While the Catalan usage pattern exhibited a step-like progression rather than a valley, this anomaly is not reflected in the temporal analysis of climate change discussions in both Catalan and Spanish. Notably, no pre-existing patterns were observed in either case. However, a growth trend is apparent after 2017. While Climate Concerns might not have been a prominent topic prior to 2017, the platform's redesign likely amplified its presence.

We observe usage valleys at the end of each year. It is difficult to identify other patterns over the years due to the impact of COVID on Twitter usage. In the case of Catalan, there is a decrease at the beginning and end of the lockdown (May 2020 and July 2020, respectively).

However, we do not observe this behavior for the Spanish case. Another notable fact is the stabilization of Twitter usage since the beginning of 2021. Moreover, when comparing the marked first and second periods, we can observe that for tweets in Catalan, the second period is notably smaller than for tweets in Spanish. This may indicate that discussions about this issue were more developed in the American continent. Thus, a comparison between these two languages can provide insight into the geographic locations of discussions.



(a) Catalan



(b) Spanish

FIG. 5. Topic count rate per week of tweets related to climate concern. (a) corresponds to Catalan queries, and (b) corresponds to queries in Spanish. Dotted lines represent the events we will study on detail.

Additional information can be obtained from Fig. 5. The usage of the "Activism" terms is equally proportioned. Throughout this period, the average percentage of activism-related tweet counts is 0.372 for Catalan with a variance of 0.019. For the Spanish case, the average is 0.277 with a variance of 0.006. The most intriguing case is the political panorama. For Catalan users, it is rarely used, except during the dates we have pointed out. In the case of Spanish speakers, this theme becomes more frequent from mid-2022 onwards, with the indicated peaks.

The term "Climate" was not used before 2019. For Catalan, its usage suddenly increased for over a year, reaching an average value of 0.48 with a variance of 0.011 between *2019-08-15* and *2020-02-15*. Following this period, it became a recurrent topic with a non-negligible but not excessively high tweet volume. On the other hand, for the Spanish language, recurrence is more stable, but a peak in usage is also observed after 2019. Although weaker, this peak is relevant. For the same dates, the average value reaches 0.365 with a variance of 0.011.

The economic panorama varies significantly over time for both languages, and their patterns do not coincide. While they share some events, such as a valley at the end of 2015 and a peak around mid-2020, they generally differ from each other.

We must consider that there are Spanish-speaking people both in Spain and America. This fact needs to be taken into account in these studies, as it can affect the conclusions depending on the subject being studied.

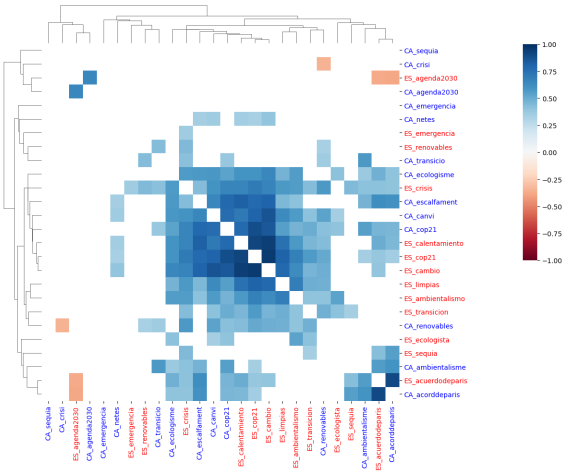
We find the peaks in the political sphere more interesting to study for several reasons. Firstly, there is a higher volume of denialism concerning political issues, which is precisely what we intend to investigate. Furthermore, we observe a similar pattern in both languages, facilitating comparison. Finally, these three peaks are temporally localized, possibly indicating that they represent specific discussions, aligning with the Twitter phenomenon we aim to characterize.

Regarding the political field, the first peak we identified occurred at the end of 2015. Despite the low Twitter usage at that time, the data amount is comparable to current usage. This peak corresponds to the United Nations Climate Change Conference in 2015 (COP21), held from November 30th to December 12th. The conference was highly controversial due to its negotiation of the Paris Agreement. The Sustainable Development Goals (SDGs) outlined in the *2030 Agenda* were also established on September 25th. Despite the low Twitter usage, these events sparked extensive debates in both languages. The collected data spans a wider period than these specific dates, starting on *2015-11-14* and ending on *2015-12-18*.

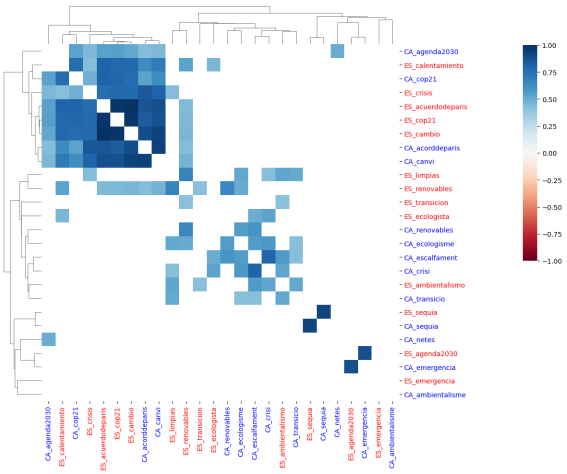
The second significant event occurred on June 01, 2017. This event was impactful within the United States (US) and had a relatively smaller impact on the rest of the world, though it appears substantial at first glance. This spike corresponds to the US withdrawal from the Paris Agreement. The decision by the Donald Trump administration to withdraw the US from the climate change mitigation agreement led to debates. This event's date range spans from *2017-05-27* to *2017-06-18*.

The last period selected for study corresponds to COP26[1]. This event generated significant discussion. We hypothesize that this may be due to the previous COP not being held due to the SARS-CoV-2 pandemic. Additionally, during lockdowns, there was increased attention on the climate issue due to noticeable differences in environmental pollution levels. This anticipation suggests that climate-related discussions were expected to be prominent on Twitter in 2021. The data collection

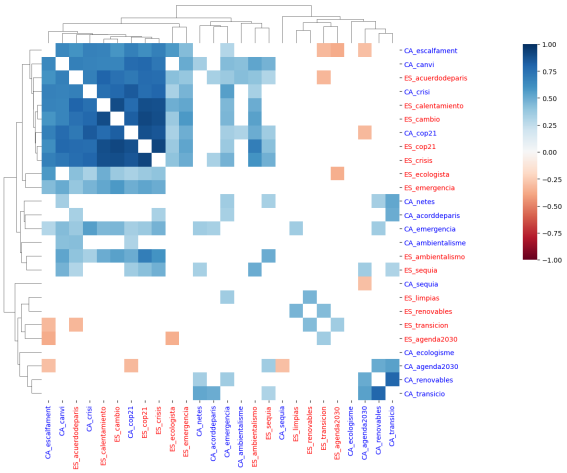
[1] The Paris Agreement was reaffirmed; it was stated that the average temperature had already risen by 1.1ºC. The Glasgow Pact was signed to double financial contributions for supporting developing countries in a sustainable manner. See more in [7].



(a) COP21 period time.



(b) US withdrawal period time.



(c) COP26 period time.

FIG. 6. Correlation matrix for all subqueries in different selected time periods. Labels in red are for Spanish and blue are for Catalan tweets.

period is also broader than the dotted line dates shown in Fig. 4. The start date is *2021-10-18*, and the end date is *2021-12-01*.

In Fig. 6, we observe the correlations between each term present in the queries 1. In Fig. 6(a), we can observe that all terms (subqueries) related to politics are correlated among themselves and across different languages. However, there is an exception: the term *2030 Agenda* in Spanish is negatively correlated with COP21 and Paris Agreement tweets. Users seem to transition from discussing the *2030 Agenda* to focusing on COP21. A similar separation between the *2030 Agenda* and other political terms is observed in the US withdrawal case, but in Catalan tweets, the correlation is higher. The figures in 6 do not have confusion, except for COP26, since it didn't exist yet. The correlations become stronger in this last period due to a larger amount of data, which allows for clearer positive or negative tendencies to emerge. Climate change often shows a consistently positive correlation; discussions about politics are intertwined with concerns about the climate issue.

IV. ANALYSIS OF POINTED EVENTS

A. Network Formalism

In network science, the most commonly used formalism for analysis is the adjacency matrix. This matrix provides a square representation where columns and rows correspond to nodes, and each cell indicates the presence or absence of an edge, representing an interconnection between two nodes. In our case, user interactions can be asymmetric, so the adjacency matrix is asymmetric as well. However, using an adjacency matrix demands substantial computational memory, as space is allocated even for unlinked nodes. To address this, we will use another metric that only stores edges: the edge list. This list consists of connected nodes forming edges $[(node_i, node_j), (node_k, node_l)]$. Given our focus on time evolution, we will include a parameter for creation date. Additionally, as interactions between users are asymmetric, edges need not be bidirectional. It's important to differentiate the left node (source) from the right node (target) in an edge.

1. *Properties of Our Graphs*

- **Directed Graph**: The links between nodes have directionality, with the source on the left and the target on the right. In cases of bidirectional links, both edges must be collected separately.
- **Weighted Graph**: Each link carries a value reflecting the strength of the relationship. Multiple retweets or a combination of comments and likes contribute to the weight of a link. The weight is

determined by the sum of all interactions from one node to another.

2. Basic Characterization

- **Indegree and Outdegree**: In the context of directed graphs, indegree counts how many times a node appears on the right side of edges, indicating the number of users interacting with the given node. Outdegree is determined by the number of times the node appears on the left side of edges, representing how many users the node has interacted with.

- **Clustering Coefficient**: This measure gauges the tendency of nodes to form groups. It calculates the fraction of neighboring nodes that are interconnected. The coefficient is obtained by dividing the number of triangles a node belongs to by the number of triplets it forms. By selecting specific interactions for analysis, such as retweets, comments, or mentions, different dynamics can be discovered, like affinity or opposing opinions.

- **Modularity**: This metric quantifies the degree of separation between different groups of nodes compared to the expected connections in a random network. Essentially, modularity measures the extent to which a network can be partitioned into distinct communities that have more internal links and fewer links between communities than would be expected in a random arrangement. It serves as an indicator of the presence and strength of modular organization within a network, allowing us to identify groups of nodes that exhibit cohesive interactions among themselves while being relatively less connected to nodes outside the group.

These metric analyses provide insights into the graph's structure. For instance, the clustering coefficient reveals graph cohesion.

3. Community Detection

Communities refer to groups of nodes with strong internal connections and limited links to nodes outside the group. Detecting communities is intricate due to the absence of a universal method or algorithm. Selection depends on case-specific characteristics. Sometimes, complete network partitioning is required, while in other cases, nodes may belong to multiple communities.

A widely used approach, and the one we will employ, is **modularity maximization** [8]. This method optimizes a quality function that compares the actual subgraph edges to the expected edges in a random distribution, without considering community structure. Maximizing this function also maximizes connections within groups and minimizes connections between groups.

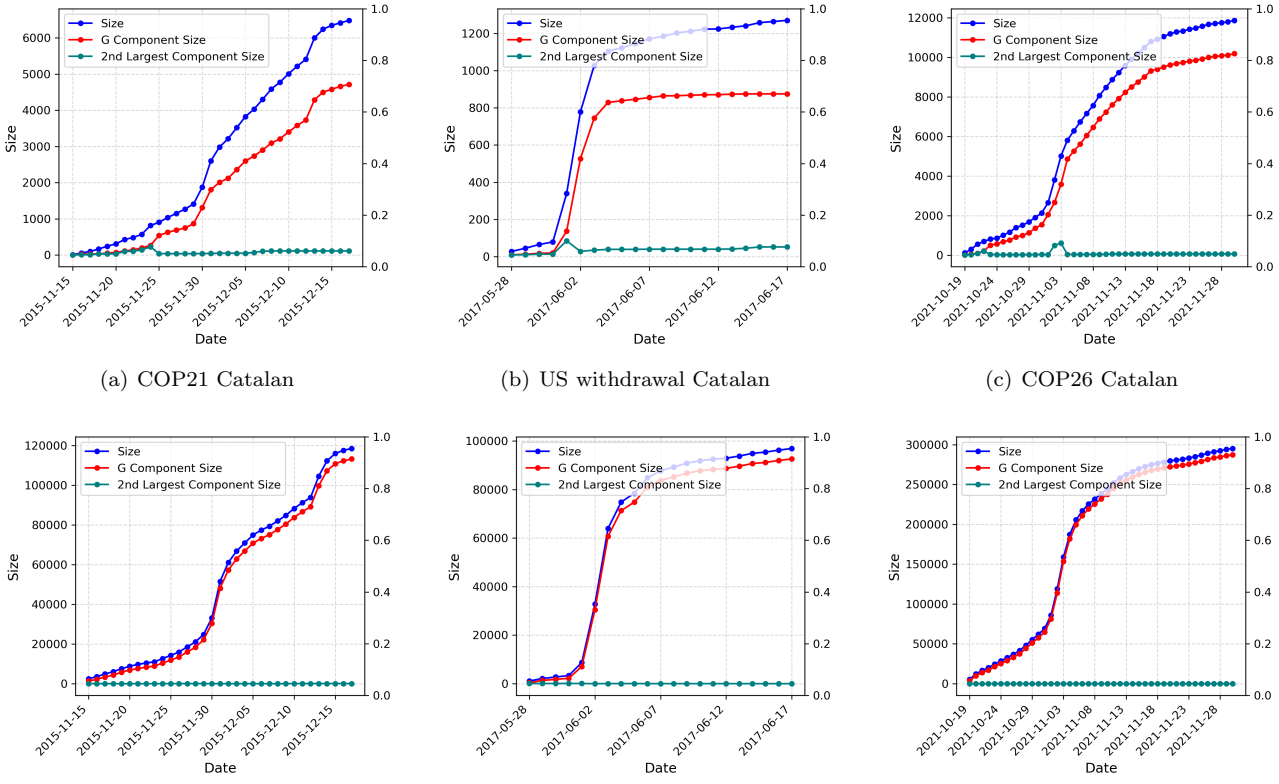
Regardless of the method used, determining whether a graph exhibits a modular partition is essential. While the method can provide an optimal graph partition, its significance lies in the existence of a community structure. In-depth analysis, comparison of various methods, visual inspection of the network, and conducting statistical evaluations, when possible, are necessary to validate the robustness of the discovered partition.

B. Size and Giant Component

For the moment, this section will evaluate the complete graph, by that we mean that the edges will represent any kind of interaction between nodes. Once we have our graphs, we can compute its size, which is the number of nodes. Another object we can obtain is the giant component. This component is a subgraph of our graph and represents the largest group of interconnected nodes. For this type of graph based on a Twitter discussion, we expect the size of the giant component and the total size to be similar. Generally, users discussing the same topic tend to interact with one another, contributing to the formation of a cohesive group. The ratio between the size of the giant component and the total size provides insight into how connected the system is, reflecting the extent of percolation. Percolation in social network-based systems offers valuable perspectives on the spread of ideas and opinions. Although we do not focus on percolation here, you can find more information in Annex VI.

The typical behavior of a "size vs. time" curve is like the one represented in all FIG 7 graphs. A topic starts to become recurrent and ends up becoming a trending topic as users interact extensively with each other. At a certain point, the curve becomes saturated as there are no new users contributing to the discussion. Even so, discussions on Twitter are varied and have different forms as in the case of COP21. In these we observed an overlap of two debates. In the zoomed time period in FIG 4, there are two distinct peaks: one at the beginning of COP21 and one at the end. This gives rise to the two jumps observed in the Spanish and Catalan graphs. The case of the US withdrawal follows a typical Twitter discussion profile, but the sharper peak makes the curve grow and saturate faster.

There are several sources of noise that could affect our analysis. For example: inaccurate queries, which could result in data not relevant to the case study. However, the most significant source of noise with the highest probability of occurrence is the linguistic dynamics between Catalan and Spanish. Given the close relationship between these languages, it is probable that users tweeting in Catalan interact with users tweeting in Spanish, introducing noise in our dataset.



(a) COP21 Catalan (b) US withdrawal Catalan (c) COP26 Catalan

(d) COP21 Spanish (e) US withdrawal Spanish (f) COP26 Spanish

FIG. 7. Sizes and giant component sizes of the graphs over time. The top graphics are for Catalan cases, and the bottom graphics are for Spanish discussions. From left to right, we have COP21, US withdrawal from COP21, and COP26.

We can find this cross-language dynamic in other cases, but it notoriously affects Catalan-specific networks because it is usual to find language switches due to the bilingualism of Catalan-speaking people. Given the volume of data from Spanish tweet networks we cannot appreciate how this language crossover affects.

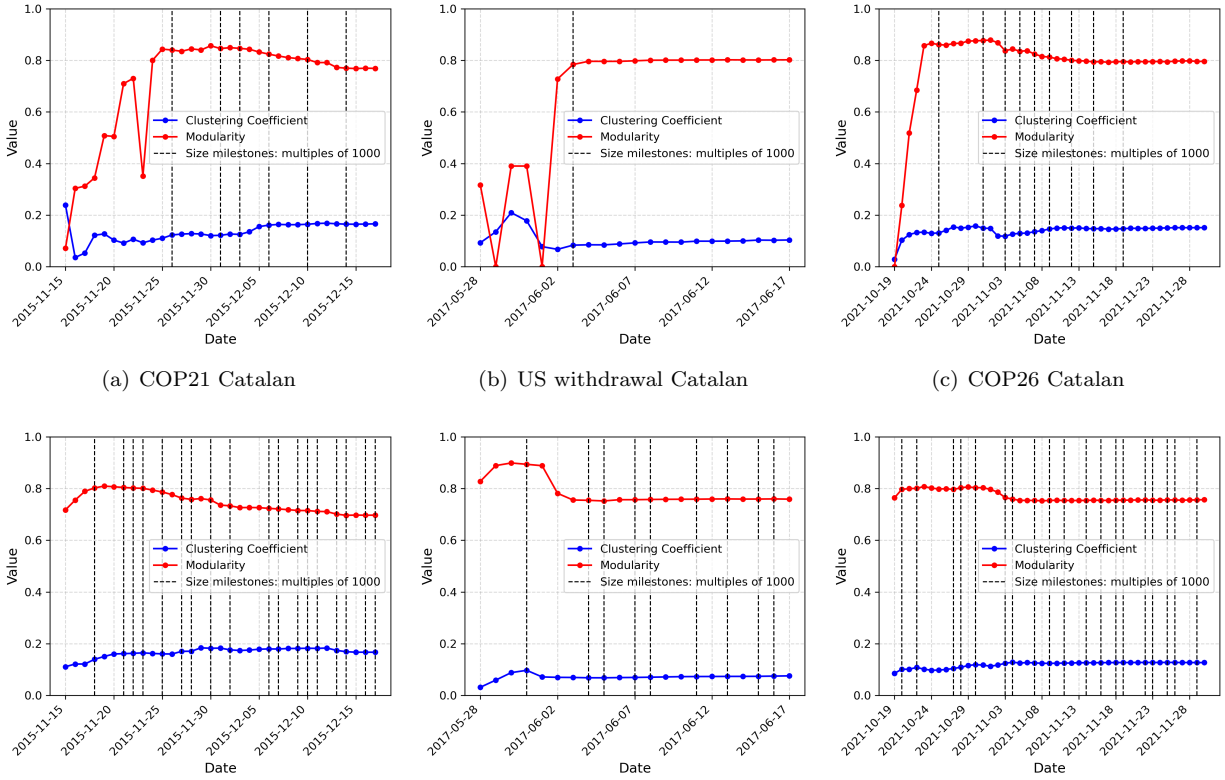
To mitigate this effect on networks of tweets in Catalan, data from both languages could be cross-referenced. It can be expected that there will be an appreciable volume of interactions with Spanish tweets. With this we would take into account these interactions and thus estimate the structure of a more complete network, since as we can see the size ratio in TABLE I is much smaller than in the Spanish tweet networks.

TABLE I shows that the Spanish graphs have a higher $\frac{Giant\ component\ size}{Total\ size}$ ratio than the Catalan ones. In Spanish cases, users are more likely to interact with others who tweet about the studied topic. Although Catalan and Spanish cases exhibit similar responses to events, the Catalan systems are proportionally smaller, making them more vulnerable to noise.

Let's now examine the size of the second-largest component. The pattern differs between the two languages but remains consistent within each language across different case studies. In the case of Catalan, we observe a distinct pattern: the curve starts at zero, sharply rises to a peak, and then rapidly declines to zero again. This pattern suggests the emergence of an additional component beyond the main one. By definition, this component is not connected to the primary component. These two components appear to represent opinions that remain unconnected for a certain period before merging. This peak is closely related to the concept of percolation and emerges immediately after the percolation threshold, marking the transition from disconnected components to a partially or fully connected network.

Graph	Total size	G Component size	Fraction
COP 21 Ca	6.482	4.714	0.7272
COP 21 Es	118.464	113.166	0.9553
US Withdrawal Ca	1.270	875	0.6890
US Withdrawal Es	96.908	92.617	0.9557
COP 26 Ca	11.864	10.185	0.8584
COP 26 Es	295.153	287.272	0.9733

TABLE I. Summary of the final parameters from FIG 7. Total size is the total number of nodes (users) involved in each network (Twitter discussion), G Component size is the number of nodes belonging to the giant component, and Fraction is the ratio of the G Component size to the total size.



(a) COP21 Catalan (b) US withdrawal Catalan (c) COP26 Catalan

(d) COP21 Spanish (e) US withdrawal Spanish (f) COP26 Spanish

FIG. 8. Clustering and modularity of the graphs along time. It also appears the milestones, this milestones are the number of nodes, here are represented by a dotted line for every 1000 or 10000 nodes added to the graph depending on the number of total nodes.

The difference between the overall system size and the giant component's size indicates the presence of non-interacting nodes that don't establish connections within the giant component.It's important to note that the lack of interaction doesn't necessarily mean an absence of all forms of engagement; it could indicate that our query failed to capture interactions with these "independent" nodes.

For Spanish, we observe that this parameter consistently remains at zero throughout the entire time period. This observation indicates that interactions largely revolve around the initial tweets once the discussion starts. The ratio of the giant component's size to the overall system's size is nearly one at each time step within the studied period, implying an effective percolation threshold of zero. This means that the system remains entirely connected throughout the study period. Additionally, the emergence of independent nodes is considerably less due to the issues mentioned above.

C. Clustering and Modularity

In this section, we observe the values of these two quantities over time for the network, considering all nodes and all types of interactions between users. In FIG 8, a similar behavior is observed for the stationary state of each case in terms of clustering and modularity. However, there is a difference between the Spanish and Catalan behaviors due to the graph size. Initially, the Catalan case exhibits a lot of noise, with low values of clustering and modularity indicating a few nodes that do not interact with others. As the discussion develops, the graphs eventually converge to the same pattern of values as the Spanish case: a final modularity around 0.760 and clustering around 0.132. Exact values can be found in TABLE II.

Graph	Modularity	Clustering
COP 21 Ca	0.769	0.166
COP 21 Es	0.696	0.167
US Withdrawal Ca	0.801	0.104
US Withdrawal Es	0.760	0.076
COP 26 Ca	0.780	0.151
COP 26 Es	0.756	0.127

TABLE II. Summary of the final values for Modularity and Clustering in FIG 8.

A modularity value of 0.760 is relatively high, suggesting a highly structured network with distinct communities. Nodes within the network are grouped into cohesive

subsets with relative stronger interactions within each community compared to between communities. This structure indicates the presence of different groups with specific interaction patterns.

The clustering coefficient, measures how many neighbors of a node are connected to each other, has a value of approximately 0.131. This relatively low value implies that while the network exhibits a modular structure, the number of local connections (clusters) compared to all possible connections is moderate.

These values together suggest that the network has a structure organized into communities, with some connections between these communities. This configuration may represent a network in which well-defined groups of nodes interact intensively within each group, but there is also some connectivity between these groups.

With these values of modularity and clustering, in the context of a social platform-based network, this pattern may indicate the presence of nodes (*hubs*) with numerous interactions that are somehow connected. This structure resembles a star-like pattern, with hubs at the ends that interact with a large number of nodes. Given the social nature of the platform, it's plausible that these hubs represent social leaders with a large following, connected through intermediary nodes that interact with multiple leaders.

D. Assessing polarization

Polarization relates to the distribution of opinions in a society and, therefore, to quantify it, the statistical characteristics of such a distribution will have to be considered [13].

To measure polarization as a network property of interactions between users, it should be reflected in the structure of the network by creating two separate communities (groups of nodes internally very well connected and sparsely connected to each other). This makes it possible to measure the degree of polarization in a system only by observing the structure of interactions within it [14].

Since we are looking for two groups of people with similar opinions, based on the above networks we will filter them to keep only the subgraph of retweets. In Twitter when you agree with a tweet and want to share his opinion the retweet is the interaction you exert on that tweet. We will use the Kernighan–Lin algorithm [15] to separate the network imposing two similar sized subgraphs. This way we want to find users who agree with the existence of climate change and the climate change denialists.

The magnitude we will compare is the $z - score = \frac{p_0 - \overline{p}}{\sigma_p}$ being p_0 the modularity of the real graph, $\overline{p}$ is the mean value of modularity of the random version of each graph and σ_p is the standard deviation of the randomized graph.

We compared three types of partitions to obtain more information about the retweet network. The first one is the one we will call "Polarization". This is the partition that we obtain with the Kernighan-Lin algorithm. The second is "Louvain". This partition into Louvain communities with resolution $r = 1$ [16]. This will be the one we will use as a control. It is based on maximizing modularity so we should always get the highest z-score. Finally we have "Louvain 95%" we call it so because we will use a Louvain algorithm with a resolution such that at least 95% of the total nodes of the network are within the two largest communities. What we are comparing is whether our measure of polarization is accurate assuming that we have two groups of similar size with different opinions.

Time Period	Polarization		Louvain $r = 1$		Louvain 95%	
	Value	z-score	Value	z-score	Value	z-score
COP21 Ca	0.38	1.90	0.79	14.53	0.71	12.04
COP21 Es	0.32	19.41	0.71	56.59	0.63	48.37
US With. Ca	0.38	0.54	0.79	7.77	0.43	1.51
US With. Es	0.37	4.99	0.78	19.26	0.72	17.39
COP26 Ca	0.41	3.36	0.82	17.73	0.75	15.10
COP26 Es	0.43	54.93	0.79	111.66	0.55	73.45

TABLE III. The table summarizes: in each row the period we are talking about, and in each column we have the three partitions described with their absolute value of the retweet network and the z-score as described above.

The values vary depending on the size of the network; the bigger it is, the more stable its random version is, which results in a smaller σ_p and a higher z-score. The high values of standard deviation for small networks in their random versions could also be studied further, as there is no need to exhibit this behavior. As a general rule, we can observe that the modularity according to the Louvain partition is always the largest. This is expected because by definition, this partition maximizes modularity. Then, we can notice a significant difference between the z-scores of the modularity calculated with the polarization partition and those calculated with Louvain 95%.

Moreover, this value provides us with information about the network's shape. In the case that the network tends to form a single large community, all z-scores should be low since the modularity definition implies the presence of modules. In the case that the network is divided into many communities, we would expect a low z-score for the polarization partition. However, the z-scores of the Louvain partition and Louvain 95% should be comparable and high. Another scenario is that we have polarity in two similar groups. In this case, all three values will be comparable and quite high.

Looking at TABLE III, we can see that for the COP21 discussion in Spanish, the two Louvain $r = 1$ values are comparable and high compared to the polarization partition's z-score. We can also observe how the polarization partition's z-score for the COP26 in Spanish is only half of that of the Louvain partition. This polarization partition z-score is the highest we obtain.

E. Visual Results on Comparing Natural Modularity versus Polarization

In this section, we utilize Gephi to create visualizations of the graphs, enabling a visual comparison. Additionally, we identify the main hubs within the largest communities.

For the purposes of this section, we will focus on two graphs: the retweet graphs for the COP21 and COP26 periods in Spanish.

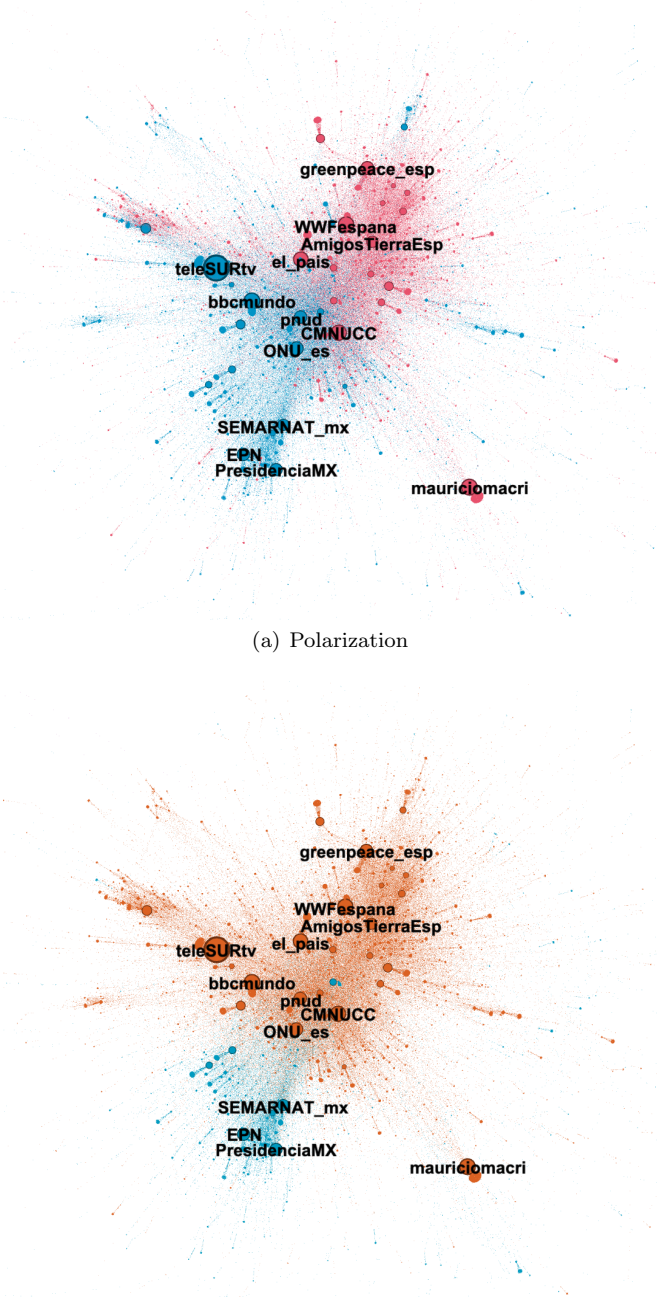
1. *Visualization of the COP21 Spanish Graph*

Looking at FIG 9, it's apparent that distinct community separation is not immediately evident. When using the Louvain 95 partition as a mathematical reference, we observe that enforcing size equality among communities leads to a noteworthy shift of many nodes, including hubs, from one community to another. This transition seems unnatural due to the significant disparity in community sizes. Specifically, the Louvain partition has an 83-17 ratio, whereas polarization enforces a 50-50 node distribution between communities. This divergence in sizes affects the polarization hypothesis.

Focusing on FIG 9 b), we examine the nodes that have received the most retweets (the hubs, represented larger in the images) to understand why they are in different communities.

The hubs in the largest community are: **teleSURtv**, **bbcmundo**, and **mauriciomacri**. **teleSURtv** is considered an alternative media outlet with headquarters in South America, focused on providing visibility to marginalized voices. It is associated with an innovative ideology, implying a bias towards progress. **bbcmundo** is a news portal that impartially describes global events, and its international recognition among Spanish speakers justifies its presence in the largest community. **Mauricio Macri** is a prominent Argentine politician who served as the president of Argentina from 2015 to 2019. His ideology falls within the center-right of the political spectrum. Despite this, during his presidency, he did not prioritize climate change policies. This anomaly can be attributed to his presidency coinciding with the COP21 period.

EPN, **PresidenciaMX** and **SEMARNAT_mx**. **EPN** (Enrique Peña Nieto) strongly supported the measures approved in the Paris Agreement and was associated with the *Partido Verde Ecologista de México*. The audience following Enrique Peña and Mauricio Macri likely differs significantly, leading to their separation into distinct communities. Enrique Peña Nieto was also the president during the COP21 period. **PresidenciaMX** was the official account of Enrique Peña's party and shared a similar ideology with him. **SEMARNAT_mx** is the official account of Mexico's Secretary of Environment and Natural Resources, actively supporting climate measures and demonstrating a strong awareness of global climate issues. We've identified a community that shares similar environmental values, although it is not as large as initially expected.



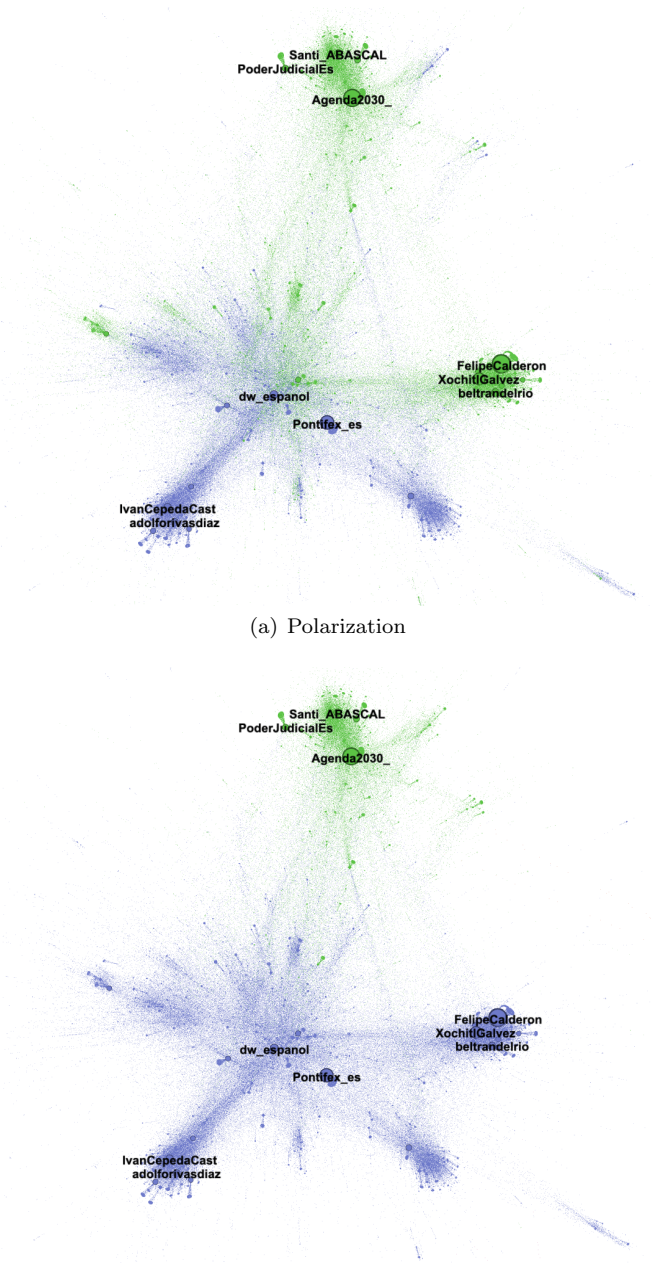
(a) Polarization

(b) Louvain 95%

FIG. 9. The top image depicts the COP21 Spanish graph colored according to the polarization partition. The bottom image shows the same graph using the Louvain 95 partition.

Given the ideological misalignment of Mauricio Macri within his community, we hypothesize that the separation into communities may be influenced by geographic factors. The larger community covers all of South America, while the smaller community primarily encompasses Mexico, located in North America.

2. Visualization of the COP26 Spanish Graph



(a) Polarization

(b) Louvain 95%

FIG. 10. The top image illustrates the COP26 Spanish graph colored according to the polarization partition. The bottom image shows the same graph using the Louvain 95 partition.

It is noticeable that this network exhibits distinguishable communities. Despite this, the same issue persists. Forcing a 50-50 split in a network where Louvain found a 79-21 ratio results in a significant number of nodes moving to the main community when transitioning partitions.

In the largest community, we find the following hubs: **FelipeCalderon**, **Pontifex_es**, and **XochitlGalvez**.

Felipe Calderón, the president of Mexico at that time, was a strong advocate of ideas shared during climate summits. He is known as the Honorary President of the "New Climate Economy" Commission. **Pontifex_es** is Pope Francis's official account, known for advocating extensively for COP26 and holding worldwide influence. Despite language barriers, Pope Francis is a central node in the network and one of the most retweeted accounts. **XochitlGalvez**, a Mexican presidential candidate, aligns with COP26 measures and has an environmentally conscious mentality.

On the other hand, we find: **Agenda2030_**, **Santi_ABASCAL** and **PoderJudicialEs**.

Agenda2030_ is a parody account that comments on various topics, including the 2030 agenda. Its proximity to Santiago Abascal is due to his denialist stance. Notably, the presence of Agenda2030_ in the query has led to unrelated tweets being included, aiding in identifying the denialist position. Santiago Abascal, leader of the Spanish ultra-right party, openly opposes environmentalist ideas, a stance consistent with denialism. The relationship with the **Agenda2030_** account is thus clarified. **PoderJudicialEs** is the official account of the Spanish General Council of the Judiciary's Communications Office, representing an official institution. While impartial, it garners significant support from the conservative sector, particularly after publishing a ruling against a member of the Podemos party.

Although geographic distribution of hubs may contribute to the community split, we have finally achieved the desired partition-supporters of climate change measures versus denialists. The large community might further divide into smaller country-based communities within Spanish-speaking America.

V. CONCLUSIONS

Throughout this paper, we have endeavored to provide a framework for conducting sociological studies using social platforms as data sources. Building upon this foundation, we conducted a data collection on Twitter discussions, focusing on a specific topic: climate change.

In our temporal analysis, we did not identify the typical usage pattern of Twitter, characterized by a decline in 2017 for the Spanish dataset and a surge for the Catalan dataset. However, we did observe peaks of activity related to a specific subtopic: climate change intertwined with political discourse.

Employing the formalism of complex networks, we illuminated the underlying structure of this data. Furthermore, our objective was to detect polarization within the discussions, identifying distinct groups with contrasting views on the concept of climate change.

A significant point of consideration is that attempting to segregate this discussion into two equally-sized graphs using the Kernighan-Lin algorithm did not yield successful results. This became apparent when comparing the graphs resulting from various partitioning methods, re-

vealing substantial discrepancies in community sizes. In hindsight, our pursuit would have been more accurate had we incorporated explicit negationist terms in our search query.

As demonstrated, leveraging social platforms for sociological studies offers a robust option. The capacity for in-depth and versatile data analysis is substantial. Additionally, the immediacy of responses to social events aligns with our expectations. Given that social platforms encompass a significant portion of society, investigations grounded in this data source provide comprehensive insights.

In the context of my specific case study, I acknowledge that certain elements, such as potential cross-linguistic connections, were not factored into the analysis. Integrating this dimension would have enhanced the understanding of the Catalan graphs. Furthermore, I've recognized the importance of maintaining a specific range of network sizes. Utilizing an excessively small graph introduces noise that distorts accurate representation, while employing overly large graphs strains computational resources beyond their capacity. Striking a balance between these constraints is crucial to ensuring accuracy and feasibility in the study's findings.

This underscores the necessity of clearly defining research objectives when designing a sociological study.

Regarding the visual analysis of graphs based on community separation through modularity, we've identified that many partitions exhibit irregular sizes. This aspect renders the polarization algorithm used inadequate, as it inherently promotes size equality. Furthermore, these partitions tend to align with geographic distinctions. In the case of the COP26 graph, we uncovered substantial environmentalist support from Mexico and a notable denialist faction from Spain.

VI. ANNEX: PERCOLATION THEORY

In essence, percolation deals with the connectivity of the components of a system. Imagine a network in which individual nodes represent individuals or entities, and connections between nodes symbolize interactions or relationships. Percolation theory studies how the connectivity of these nodes changes as connections are randomly added or removed.

In the field of complex social networks, percolation can be applied to understand the diffusion of opinions or information. Each node in the network corresponds to an individual, and the edges represent their interactions. As the percolation parameters change, certain critical points emerge that define the behavior of the system.

A few interactions between users are not enough to transport opinions. As connections increase (analogous to the addition of edges in percolation), groups of nodes (users) are formed and within each of them the opinions specific to that group (cluster) are disseminated.

Criticality

Related to this property is a phase transition. Below a critical point, known as the percolation threshold, we have the isolated group phase, which confines opinions into small clusters. Beyond the threshold a giant connected group emerges, which allows opinions to percolate throughout the network.

ACKNOWLEDGMENTS

I sincerely thank all the people who have contributed significantly to the achievement of this Master's Thesis.

First and foremost, I would like to express my deep appreciation to my TFM director, Dra. Luce Prignano, for her invaluable guidance, support and dedication throughout this process. Her knowledge and resilience in the face of adversity have been very helpful to me in moving forward with this project. I would also like to thank the ideas provided by my TFM co-director, Dr. Emanuele Cozzo. They have helped me to complete a project of real value.

I would also like to thank my professors in the master's program, especially Dra. Maria Angeles Serrano and Dr. Marian Boguñá, for their commitment and teachings, which have broadened my understanding of complex networks. Their dedication and passion have awakened in me a great interest in this subject.

I would also like to thank my classmates, specifically Mireia Olives, Emma Oriol and Elena Fernandez for their support and the exchange of ideas that has helped me to clarify concepts.

I would like to express my gratitude to the University of Barcelona for providing me with the necessary resources and making education so accessible on a personal and scientific level.

I thank my family and friends in specific my partner Alexei Granados and my cousin Andrea Castillo for their constant support, understanding and encouragement throughout this academic journey.

Finally, I wish to acknowledge all the people who have indirectly contributed to this work, who through published articles or books I have been able to acquire the knowledge to capture it in this project.

[1] Chen, T. H. Y., Salloum, A., Gronow, A., Ylä-Anttila, T., & Kivelä, M. (2021). Polarization of climate politics results from partisan sorting: Evidence from Finnish Twittersphere. Global Environmental Change,

71, 102348.

[2] Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., ... & Baronchelli, A. (2022). Growing polarization around climate change on social media. Nature Climate Change, 12(12), 1114-1121.

[3] Twitter, (Retrieved 2023). API Documentation. Twitter https://developer.twitter.com/en/docs/twitter-api

[4] Summers, E. and Ruest, N. (2023) Twarc (v2.14.0). Retrieved from https://github.com/docnow/twarc

[5] MINISTERIO PARA LATRANSICIÓN ECOLÓGICA Y EL RETO DEMOGRÁFICO, (Retrieved 2023) *Resultados de la COP21*. Gobierno de España https://www.miteco.gob.es/es/cambio-climatico/temas/cumbre-cambio-climatico-cop21/resultados-cop-21-paris/default.aspx

[6] CEPAL, N. (2018). The 2030 agenda and the sustainable development goals: An opportunity for Latin America and the Caribbean.

[7] Naciones Unidas, (Retrieved 2023) *COP26: Juntos por el planeta*. Naciones Unidas https://www.un.org/es/climatechange/cop26

[8] Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.

[9] Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. nature, 393(6684), 440-442.

[10] Barabási, A. L., and Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

[11] Guimera, R., and Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. nature, 433(7028), 895-900.

[12] Álvarez, R. (2017-09-29). *Twitter está duplicando el número de caracteres hasta los 280 por tweet, el cambio más radical desde su nacimiento*. Xataca https://www.xataka.com/servicios/

[13] Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., ... & Holman, B. (2017). Understanding polarization: Meanings, measures, and model evaluation. Philosophy of science, 84(1), 115-159.

[14] Salloum, A., Chen, T. H. Y., & Kivelä, M. (2022). Separating polarization from noise: comparison and normalization of structural polarization measures. Proceedings of the ACM on human-computer interaction, 6(CSCW1), 1-33.

[15] Niño, J. O. P. (2020). Detección de comunidades en redes: Algoritmos y aplicaciones. arXiv preprint arXiv:2009.08390.

[16] (Retieved 2022-11-28). Louvain method. Wikipedia. https://en.wikipedia.org/wiki/Louvain_method.

[17] Li, M., Liu, R. R., Lü, L., Hu, M. B., Xu, S. and Zhang, Y. C. (2021). Percolation on complex networks: Theory and application. Physics Reports, 907, 1-68.