Ecole polytechnique federale de Lausanne

Master Thesis

Master in Computational Science and Engineering

Describing information influence in social media with coupling inference methods

Author:
Cyril Vallez[*,†]

Principal Supervisors:
Dr. Erik Hemberg[†]
Prof. Una-May O'Reilly[†]

Co-Supervisor:
Prof. Robert West[*]

[*]Ecole Polytechnique federale de Lausanne, Switzerland
[†] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, United States

February 15, 2023

Abstract

Quantifying influence between users on social media is an inherently difficult task, but essential in order to detect and counter disinformation. Based on the news sources shared on Twitter, this work focuses on leveraging coupling inference methods between time series of user activity to understand the dynamics of influence on social media. We compare two coupling inference methods that can be used to build graphs representing different influence relationships between users. From those graphs, we define influence measures relating to new and existing concepts in social media influence. We demonstrate our means of analysis on climate change related posts on Twitter during COP26 and COP27, and on a well known and well studied case of attempted foreign influence around the Skripal poisoning in the UK in 2018. We find that our methods allow us to detect users who spread more disinformation than average. We compare several influence measures and discover that they are only weakly correlated, implying that they do not capture the same influence types, but are complimentary.

Acknowledgment

I am grateful to all of those with whom I have had the pleasure to work during my time in Boston. This master thesis has been an intense experience for me, and a great dive into the American culture.

First and foremost, I would like to thank Erik Hemberg and Una-May O'Reilly, who trusted me from the beginning and gave me the opportunity to come to Cambridge to work with them in the ALFA lab. This work would never have been possible without their support, guidance and expertise. They are truly exceptional mentors who really cared about my work and provided help on (at least) a weekly basis. Special mention to Erik for providing feedback on at least 4 long drafts of this report!

Special thanks to Thomas Galligani for the work he had already done and that he allowed me to use, as well as for valuable discussions and for taking the time to review this document. I feel blessed for these past months alongside all of the truly exceptional members of the ALFA group: Stephen Moskal, Aruna Sankaranarayanan, Shashank Srikant, Michael Wang, and, more recently, Ethan Garza and Sam Laney. Of course this also includes Erik, Una-May and Thomas. All of you made me feel welcome and were very kind to me, qualities that are too often underrated. You may not realize it, but it did play a large role in making my stay such a good experience.

I am grateful to Ozlem Garibay, Ivan Garibay, Chathura Jayalath, Jasser Jasser, Bruce Miller, Alex Baekey, and all the MIPs ICE team for providing ideas and discussions from which this work emerges. It was a pleasure working with you and meeting you all in person in Orlando!

Finally, I want to thank Cornelia Haeringen for the time she took to proofread this report.

Contents

Chapter 1

Introduction

Detecting and countering disinformation on social media requires quantifying the influence between users, which is inherently challenging. This chapter introduces our work. In Section 1.1, we introduce social media, disinformation, and climate change concepts. Sections 1.2 and 1.3 present the research questions on which this work is based, and the contributions we make to the literature. Section 1.4 defines terminology more precisely.

1.1 Social media, disinformation and climate change

Since the appearance of social networks, there has been interest in determining the most important actors, and studying how people interact and influence each other in such structures. The advent of Online Social Networks (OSNs) such as Facebook or Twitter provided a massive amount of data to analyze social networks. Knowing the influence one user has over the network is essential for many applications, one of the most obvious being marketing [18, 28, 41]. Other applications range from human behavior [15, 43], to communication [13], to political science [3].

OSNs are now so big that they have become a major part of daily communications, and represent most of information diffusion in society. An increasing number of people tend to rely almost exclusively on their platforms for getting news [20, 33]. In such a vast ecosystem where fact checking and moderation are difficult, disinformation, misinformation and fake news thrive [22, 42]. Misinformation may be defined as false or inaccurate information that is spread, regardless of intent to mislead, meaning that the spreader may be unaware of the nature of the content he is sharing. On the other hand, disinformation is targeted misinformation: the spreader intentionally shares misleading information. Note that in both cases, the information itself may not need to be false, but may be deliberately presented in such a way as to be misleading. The effects of such erroneous information on society are broad, from loss of trust in government, scientific community, or medicine, to exacerbation

of polarization and division within and across communities.

On top of that, OSNs make use of recommender systems [19]. Such tools are used to determine what content is being shown to each individual, depending on all the informations the platform has accumulated about said individual. They try to infer as much as possible about users, in order to show them relevant ads, popular personalities to follow, and posts the user is likely to find interesting and read; the goal being to maximize user interest, and thus time spent on the platform. In this way, recommender systems shape user preferences and guide choices, both individually and socially. This poses a range of ethical problems about their use [21]. There are growing concerns about how their algorithmic power could also (unintendedly) help propagate misinformation [5, 38]. Indeed, they tend to increase the social fracture between groups of different ideologies, and encourage the creation of echo chambers.

In the last years, many disinformation campaigns have been orchestrated on social media, coming from different actors with different purposes. These campaigns usually target scattered events such as the 2016 [7] or 2020 [32] US elections, the Skripal poisoning case in the UK [26] or the Black Lives Matter (BLM) movement [34]. In some cases, the intent is obvious (which does not mean that the campaign is ineffective), such as to shift vote attitudes towards a given candidate for presidential elections, or deny state responsibility in the Skripal example. Occasionally the strategy is intended to amplify discord in order to try to undermine a government, institution or community.

"To solve the climate crisis, we must also tackle the information crisis" [14]. When it comes to existential threats, such as climate change, disinformation can prove especially harmful. This is because misinformation about climate change has confused the public, led to political inaction, and stalled support for or led to rejection of mitigation policies [6]. Contrary to the Skripal poisoning, climate change is not a single event in time to which malicious actors can respond by launching a short-term disinformation campaign. On the contrary, it is a well known and well studied phenomenon that continuously takes place due to human activities. Nevertheless, some climate related events are way more publicized than others, potentially making them targets to particularly virulent influence campaigns. Among such events, the different Conferences of the Parties (COP) are particularly well advertised. It is reasonable to assume that the amount of false informations spreading is larger during such events.

1.2 Motivation and research questions

Disinformation on social media desperately needs countermeasures. This is especially true for content relating to climate change. But how can we hope to counter something if we do not even know that it is happening or has already happened? This is the underlying motivation for this work. Can we build a detector able to discover influence operations on social media? Given that some influence operations have been extensively studied, such as the Skripal case [26], can we build a framework that would retrospectively capture its influence dynamics, and use this framework for other events such as the COP26 and COP27?

More generally, how can we measure information influence on social media data collected around given events? How general can such measures be?
Some theories already exist on how information is flowing in social networks, and how users are exposed to different degrees of information, for example the concepts of homophily and echo chambers. We raise related questions, such as what are useful empirical measures of influence on social media, and how do they relate to existing sociology concepts and theories? How sensitive are those measures to different parameters and data?
Finally, what are important properties of information sources?

To answer these questions, this work focuses on leveraging climate change related social media data from Twitter, centered on COP26 and COP27 events, to find and describe patterns of influence attempts relating to disinformation campaigns. To more generally describe actions and influence relationships between those actions, we use the trustworthiness score from NewsGuard [23] of the news sources contained in the URLs shared by the users in their posts. We then create graphs of influence describing how tweets from one user have influenced tweets of another (and in what way), by using coupling inference methods that examine time series representing the action frequency of users.

1.3 Contribution

The contributions of this thesis to the scientific world are the following :

- We create, curate, and describe 2 new datasets, consisting of Twitter data relating to climate change, centered around COP26 and COP27 events (Section 3.1.2).
- We extend an existing framework for measuring information influence on social media (Section 3.4)
- We define new influence measures and relate them to existing theories (Section 3.6.2)
- We compare two different coupling detection methods (transfer entropy and joint distance distribution) on three different datasets (Section 5.2)

- We verify generalizability of our framework on different timelines and type of events (Sections 4.1, 4.2, and 4.3)
- We perform sensitivity analysis of our methods on different parameters (Section 4.4)
- We study different influence measures on social media, and how they correlate between themselves (Section 5.4)

1.4 Notations and terminology

Twitter is a microblogging platform that enables its users to share and read short public messages, known as *tweets*, and limited to 140 characters. Tweets can contain plain text, URLs (Uniform Resource Locator, the address of a web page), images, *mentions* of other Twitter users, and *hashtags*, which are words or phrases highlighted by placing the "#" symbol in front of them. Users can choose to *follow* other users, in which case they will see all the activity of the individuals they are following. Users can also interact with tweets: they can *retweet* someone's tweet (i.e. reposting the same tweet and crediting the original author), *like* it, or *reply* to it. One may also *quote* a tweet, which is a special case of a retweet, where one adds some personal text on top of the retweet.
In this work, we only consider 2 categories, tweets and retweets. For this reason, we treat replies as tweets (the user publicly gave their opinion), and quotes as retweets (which they are a subset of).

Social networks can be represented by a digraph $G = (V, E)$, whose nodes V represent the users, and the directed edges E the interpersonal ties among them. In Twitter, the edges E can be based on the different user-user or user-tweet relationships presented above. The most common graphs are:

- follower graph: the unweighted edges E represent the follower-followee relationships between users
- retweet graph: the edges E are weighted and represent how many times one user retweeted another one
- mention graph: the edges E are weighted and represent how many times one user mentioned another one in their tweets

It is however possible to derive many other graphs representing different interactions between the users (or even between the tweets, considering the nodes V as $V = V_1 + V_2$ where V_1 is the set of users and V_2 the set of tweets posted by the users).

In the following, we use the names *Twitter metrics* or *metrics* to describe simple mathematical quantities readily available from Twitter that provide information about the social

network structure, such as the follower count or number of retweets.
We define the terms *influence measure* or *measure* to mean any formula or algorithm, combining metrics or not, that provides criteria for ranking users according to some concept of influence such as popularity or activity. Note that some metrics can also serve as an influence measure, for example the number of followers.

Let Θ be the set of all tweets containing at least one URL, Φ be the set of Twitter users, and Ψ be the set of URL domains that exist on the web. We define:

- $AU(\cdot) : \Theta \mapsto \Phi$ the mapping from a tweet to its author
- $D(\cdot) : \Theta \mapsto \Psi^n$ the mapping from a tweet to the domain of each of the n URLs contained in the tweet ($n \geq 1$).

Finally, we refer to the term *coupling inference methods*. This should be understood as methods to study and infer causality between two interacting variables in the form of *directional* coupling (i.e. drive-response relationships, information flow, or causal connectivity between the variables). Some researchers call them *causal inference methods* in the literature.

Chapter 2

Related work

Quantifying the influence of a user is a difficult task. In the first place, there is a conceptual problem. The definition of influence is somewhat loose, and there is no agreement on what is meant by an influential user. Therefore, new influence measures are constantly emerging, each of which offers different measurement criteria for different domains and applications. Riquelme and González-Cantergiani [27] provide an extensive survey of existing measures. The first studies of influence came from the marketing field [18, 28, 41], where finding a few influential individuals that could potentially guide plenty of users into using a given product could prove very beneficial as an advertisement strategy. This strategy is now called influencing the influencer [9]. The idea is simply to influence or pay a handful of users to relate whatever story or product, and hope that it will reach and have an impact on as many users of the platform as possible. While this started as a marketing strategy, the effectiveness of this approach quickly made it gain popularity as a maneuver to relate false narratives and disinformation.

Section 2.1 presents an overview of traditional measures of influence on social media. Section 2.2 describes coupling inference methods for measuring influence on social media. Table 2.1 provides an overview of the influence measures that are discussed in more details in the following.

2.1 Traditional measures of influence

Social media influence is the capacity of a user to affect how others perceive a specific problem and make a decision about it. Cha et al. [4] try to use data directly accessible from Twitter to measure influence. They compare 3 measures : *indegree* (number of followers), *retweets* (number of tweets from a user which have been retweeted), and *mentions* (number of times a given user has been mentioned in a tweet). They show that these measures are not necessarily correlated between themselves and, more importantly, that the number of

	Influence measure		Twitter metrics				content analysis	TE
			F	RT	M	L		
metric based	indegree/followers	[4]	✓	✗	✗	✗	✗	✗
	retweets	[4]	✗	✓	✗	✗	✗	✗
	mentions	[4]	✗	✗	✓	✗	✗	✗
centrality based	closeness	[8]	(✓)	(✓)	(✓)	(✓)	✗	✗
	betweenness	[8]	(✓)	(✓)	(✓)	(✓)	✗	✗
	PageRank	[24]	(✓)	(✓)	(✓)	(✓)	✗	✗
metric manipulation based	InfluenceRank	[11]	✓	✓	✓	✓	✗	✗
	IP influence	[29]	(✓)	✓	✗	✗	✗	✗
	TwitterRank	[44]	✓	✗	✗	✗	✓	✗
causality based	information transfer	[39]	✗	✗	✗	✗	✗	✓
	content transfer	[40]	✗	✗	✗	✗	✓	✓
	influence cascades	[31]	✗	✗	✗	✗	✗	✓

Table 2.1: Overview of influence measures and how they are computed. F refers to follower-followee relationships, RT to retweets, M to mentions, L to likes, and TE to transfer entropy. The symbol (✓) means that the corresponding metric may be used, but not at the same time as other metrics marked with the same symbol.

followers (indegree) is not necessarily a good representative of the impact that one user may have on the network.

Some researchers have used topological centrality measures (methods to find the most central nodes in graphs) such as *closeness* [8], *betweenness* [8], or *PageRank* [24] to determine influential users. Such measures are usually applied to the retweet graph, or follower graph (see Section 1.4). Hajian and White [11] apply a modified version of the PageRank algorithm to estimate an *Influence Rank* (IR) (an index considering an influential person as an individual who is connected to other influential individuals) and compared it to what they called *Magnitude of Influence* (MOI), a measure proportional to the number of followers who commented on, liked or propagated/retweeted a posting in the network. They show that both are correlated, but that IR and number of followers are not. Romero et al. [29] describe the *Influence-Passivity* (IP) measure. Each node in the graph is attributed an influence score, as well as a passivity score by an algorithm inspired by HITS, an algorithm for finding authoritative web pages and hubs that link to them [16]. The authors show that this measure is a very good predictor of the number of clicks URLs embedded in the tweets receive. They also compare it to the number of followers and number of retweets, as well as the PageRank score, and demonstrate that all three of those are weaker predictors, the number of retweets being the worse. Weng et al. [44] propose *TwitterRank*, an extension of the PageRank algorithm, to measure the influence of users in Twitter in a topic sensitive

way. It considers criteria of similarity between users, restricted to the topics of their tweets. However it is relatively slow, because it requires a preprocessing step for topic analysis.

2.2 Coupling inference methods for measuring influence

All the measures of influence presented above directly use metrics available from public data such as number of followers or number of retweets, or are topological measures derived from the graphs induced by such metrics. One drawback of such methods is that they are based on explicit causal knowledge (i.e. A follows B), whereas for many datasets such knowledge is not available and needs to be discovered. For this reason, researchers started to use coupling inference methods to determine the asymmetric strength of the interaction between 2 users in OSNs. Detecting causality from data is difficult and there is ongoing debate amongst scientists about the proper way to determine it.

One of the first and the most well known coupling inference method is *Granger causality* [10]. Ordinarily, regressions reflect correlations, but Granger argued that causality in economics could be tested for by measuring the ability to predict the future values of a time series using prior values of another time series. However, since the question of "true causality" is quite philosophical, one usually says "X Granger causes Y" instead of "X causes Y". More recently, Schreiber [30] derived *transfer entropy*, an information theoretic measure of the directed transfer of information between systems evolving in time. Simply put, transfer entropy from a process X to another process Y quantifies how much better we are able to predict the target process Y if we use the history of the process Y and X rather than the history of Y alone. Interestingly, Barnett, Barrett, and Seth [2] proved that transfer entropy is equivalent to Granger causality for Gaussian variables. Other coupling inference methods emerged recently, based on distances (usually euclidean) in high dimensional spaces, such as *cross mapping* [35] or *joint distance distribution* [1]. All of these measures of causality have been used in neuroscience, epidemiology, finance, and much else.

Transfer entropy has become somewhat popular to infer social influence in OSNs. Ver Steeg and Galstyan [39] introduce *information transfer* and directly use it on the time series of user activity on Twitter in order to establish influence relationships between individuals in the network. They restrict their study to tweets containing at least one URL, and show that transfer entropy between users X and Y is correlated with the number of URLs that were first tweeted by X and subsequently tweeted by Y. He et al. [12] make a very similar study, using data extracted from Tencent Weibo, a Chinese social media similar to Twitter (which was shut down in 2020). Ver Steeg and Galstyan [40] also present *content transfer* where they use transfer entropy to quantify the strength of the effect of one user's tweet content on another's. To capture content, they first compute time series of tweets by users, then transform the text of the tweets into vectors before applying transfer entropy. They

show that their measure is a good predictor of user mentions even for pairs of users not linked in the follower or mention graph. More recently, Senevirathna et al. [31] introduce *influence cascades* to study different social influence relationships (e.g. how a new post influences sharing the post as opposed as to how a new post influences contributing to the post) using transfer entropy. They investigate the different influence relationships within and across two platforms, namely GitHub and Twitter.

Chapter 3

Methodology

In this chapter, we describe the datasets and methods we use. Sections 3.1 and 3.2 respectively depict the datasets and data processing we apply to these datasets. Sections 3.3 and 3.4 present how we derive time series of user activity, and use them to create influence graphs. We then explain how we extract influence cascades in Section 3.5, and how we compute influence measures derived from the influence graphs in Section 3.6.

3.1 Datasets

In order to study influence spread on Twitter, we need data from the website. At the time of writing, Twitter offers free access to their API[1] to researchers. Using this access, one is allowed to extract up to 10 million tweets a month from the platform. In order to get data, one must formulate a query consisting of a set of keywords and/or properties to match in the tweets (e.g. "dog lang:en" would match all tweets containing the word dog and written in English), along with the time period in which to search for. The API will respond with all the tweet objects matching the query in the provided time frame. These tweet objects consist of the text of the tweet, as well as a large quantity of metadata, such as information about the author, information about the parent tweet in case of a retweet or reply, or URLs and hashtags contained in the text.

Table 3.1 shows a high-level description of the datasets we use. The exact queries we formulate are available in Table A.7.

3.1.1 Skripal dataset

We calibrate and study our method on a well known disinformation campaign. Sergei Skripal is a former Russian intelligence officer who worked as a double agent for the UK. He

[1] https://developer.twitter.com/en/docs/twitter-api

Dataset	Type	Topic	Year	Start date	End date
Skripal	spontaneous	poisoning	2018	2018-03-04	2018-04-15
COP26	scheduled	climate change	2021	2021-10-18	2021-11-26
COP27	scheduled	climate change	2022	2022-10-24	2022-12-02
RandomDays	scattered	climate change	2020-2022	-	-

Table 3.1: Datasets overview. The date format is YYYY-MM-DD. The start date is always inclusive, and the end date exclusive.

was arrested by Russia's Federal Security Service (FSB) in 2004, convicted of high treason and sentenced to 13 years in prison. In 2010, he was released as part of a spy exchange and settled in the UK.

On March 4, 2018, Sergei Skripal and his daughter Yulia, who was visiting from Moscow, were poisoned with a Russian-developed nerve agent called Novichok. They were hospitalized in Salisbury District Hospital in critical condition. The British intelligence service investigated the incident as an attempted double murder by the Russians.

This attempted double murder was heavily publicized, and on March 12, 2018, Theresa May, prime minister of the UK at the time, publicly attributed responsibility to Russia. In response, the Russians started a large disinformation campaign to try to deny their culpability, notably through news outlets such as RT and Sputnik. Ramsay and Robertshaw [26] published a large report documenting the disinformation campaign. They show that one of the major strategies was to Flood the Zone (FTZ), i.e. publish a large amount of separate and sometimes contradictory narratives about the event, making it harder for genuine news to reach an audience.

In this case, the data was not obtained directly via the Twitter API, but through BrandWatch[2], a third party allowing data collection from social media. We formulated the same query directly to the Twitter API but obtained less data this way. We believe this is because a non-negligible amount of tweets and users was banned from the platform since the event. We therefore decided to use the BrandWatch data. We asked for every tweet containing the words "Skripal" or "novichok" (and corresponding hashtags), containing at least one URL, and written in English, between March 4, 2018, and April 14, 2018, included. The dates are summarized in Table 3.1.

[2] https://www.brandwatch.com/

3.1.2 COP26 & COP27 datasets

Our primary interest is to investigate influence attempts related to climate change. For this reason, we identified climate events susceptible to be victims of disinformation campaigns. The Conference of the Parties (COP) is the supreme decision-making entity of the United Nations Framework Convention on Climate Change (UNFCCC). The UNFCCC was formed in 1994 to stabilize the greenhouse gas emissions and to protect the earth from the threat of climate change. As of 2019, the number of member countries in the UNFCCC has reached 197. The COP summits usually take place every year (an exception was made in 2020, during the covid pandemic, at which time COP26 was rescheduled to the following year) for the parties to meet and decide on how to act in favor of climate.

COPs are the largest climate summits in the world, and are very publicized, making them a target for disinformation. King, Janulewicz, and Arcostanzo [14] document and study disinformation about climate change at COP26. The report stated "to solve the climate crisis, we must also tackle the information crisis".

We focus on the two last COP events, namely COP26 in Glasgow, Scotland, between October 31, 2021, and November 12, 2021, and COP27 in Sharm El Sheikh, Egypt, between November 6, 2022, and November 18, 2022. For each of these events, we queried the Twitter API for tweets containing the explicit words "climate change", "climate crisis", "climate emergency", "climate action" or "global warming" (and corresponding hashtags), containing at least one URL, and written in English. To allow for fair comparison, we asked the API for tweets before and after the events as well, using the same number of days before and after as during the event. The exact dates can be examined in Table 3.1.

3.1.3 RandomDays dataset

In order to evaluate our results, we query the Twitter API for climate data for a random period of time. However, as we cannot be sure that no climate event happened during a period of roughly two weeks, we decided to query random, separate days. We randomly picked 13 days between January 1, 2020 and December 31, 2022, and used the same query as for the COP26 and COP27 datasets for each of these days. We then re-aggregated the data as if it was continuous (i.e. as if it had happened during 13 consecutive days) keeping only the time of the tweets and discarding the date. Since we want to use this dataset as a baseline on which no major disinformation campaign occurred, we believe this approach is less likely to capture special events than asking for a random 13 days period and hoping that no substantial climate event took place in said period (during which disinformation may flow heavily).

3.2 Data processing

Now that we have defined our datasets, we present the data processing pipeline we apply on each of these tweets prior to using them to compute influence. We rely on three key concepts : the notions of stratification, actions and actors.

3.2.1 Stratification of the dataset

We define as stratification a method designed to separate the data, in order to compare results across different groups. In the following study, we use temporal stratification, meaning that we split the data based on specific dates of the tweets around an event. Let dataset $D \subset \Theta$ be the set of tweets posted in the time period T. We split the time interval into 3 sub-intervals:

$$T = \{T_b, T_d, T_a\} \quad \text{s.t.} \quad \begin{cases} T_b \cup T_d \cup T_a = T \\ T_i \cap T_j = \emptyset \ \ \forall \ \ i, j \in \{b, d, a\} \end{cases} \tag{3.1}$$

where the indices b, d, a are chosen to mean Before, During and After respectively. This is because we stratify around a given event in order to compare how information is flowing before, during, and finally after the event.

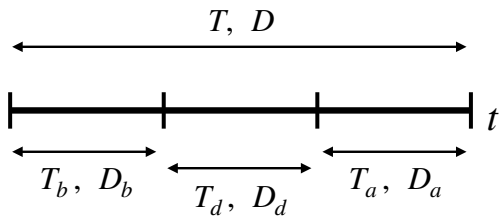


Figure 3.1: Representation of the temporal stratification defined in Equation 3.2.

Let x be a tweet posted at time t. Then we define the groups, or *strata*, as:

$$D_s = \{x \in D \mid t \in T_s\} \quad \forall \ s \in \{b, d, a\} \tag{3.2}$$

Note that the properties in Equation 3.1 also hold for the strata D_s. Explicitly we have:

$$\begin{cases} D_b \cup D_d \cup D_a = D \\ D_i \cap D_j = \emptyset \ \ \forall \ \ i, j \in \{b, d, a\} \end{cases} \tag{3.3}$$

Figure 3.1 provides a visual representation of the stratification process. This procedure allows us to compare how volume and characteristics of influence change over time at proximity of important events. In all cases, we make sure that the length of the time period

spanned by each group is the same for one dataset.

	Before		During		After	
	Start	End	Start	End	Start	End
Skripal	2018-03-04	2018-03-18	2018-03-18	2018-04-01	2018-04-01	2018-04-15
COP26	2021-10-18	2021-10-31	2021-10-31	2021-11-13	2021-11-13	2021-11-26
COP27	2022-10-24	2022-11-06	2022-11-06	2022-11-19	2022-11-19	2022-12-02

Table 3.2: Stratification dates for each of the datasets presented in Sections 3.1.1 and 3.1.2. The date format is YYYY-MM-DD. The start date is always inclusive, and the end date always exclusive.

Table 3.2 presents the dates on which we apply the stratification process for the Skripal, COP26, and COP27 datasets. Note that we do not apply stratification to the RandomDays dataset.

3.2.2 Actors - users

We define actors as the entities spreading influence. Thus, one choice of actors is to use all unique Twitter users present in the dataset. However, one could also decide to aggregate a community of users with similar beliefs as a single actor in order to understand how influence is flowing between different communities and groups of users. In the following, we study the impact of such choices by first using all users, and then creating aggregates of users based on different properties. But we always constrain an actor to have a minimum *activity rate* (or tweet count) that we set to a minimum of 3 tweets in the time period corresponding to the stratum considered. Let H_s , $s \in \{b, d, a\}$ be the set of actors inside stratum s. We enforce:

$$AR_{T_s}(A) \geq 3 \quad \forall\, A \in H_s \tag{3.4}$$

where $AR_{T_s}(\cdot) : H_s \mapsto \mathbb{N}$ is the activity rate (or tweet count) mapping an actor to the number of their tweets in the time period T_s (corresponding to the stratum D_s). Note that $AR_{T_s}(\cdot)$ only counts the number of tweets, *excluding* retweets.

Most of the time, we are going to use as actors all individual users, in which case the actors H_s are given by:

$$H_s = \{A \mid x \in D_s\,,\ AU(x) = A\,,\ AR_{T_s}(A) \geq 3\} \tag{3.5}$$

Note that actors are usually different inside each stratum, i.e. in general $H_b \neq H_d \neq H_a$.

We define the mapping $AC(\cdot) : D_s \mapsto H_s$ from a tweet to the actor who posted it. It may happen that a tweet does not map to any actor. For example, when actors are defined to be all users as in Equation 3.5, all tweets from users with an activity rate less than 3 do not map to any actor. We discard all these tweets, i.e. we perform the remapping:

$$D_s \mapsto \{x \in D_s \mid AC(x) \in H_s\} \tag{3.6}$$

After this operation, each tweet $x \in D_s$ has a corresponding actor.

3.2.3 Actions - sharing URLs

We define actions as: what is the influence attempt from one actor to the other. Remember from Section 3.1 that all tweets acquired from Twitter contain at least one URL. To define actions, we use NewsGuard ratings. NewsGuard [23] is an organization actively committed to combat disinformation. Their team of journalists reviewed and rated all the news sources that account for up to 95% of online engagement. They use a strict and transparent policy based on 9 criteria[3] (that can be freely checked on their website) to give a trust score between 0 and 100 to each of the news sources they review. It is important to understand that each of these ratings were produced by humans following rigorous rules. According to NewsGuard, a news website which obtained a score larger than 60 can be considered as trustworthy (T), and untrustworthy (U) otherwise. We obtained the list of ratings from NewsGuard as of September 2022. It contains 8145 unique news domains having been fully rated. Using this list of rated domains, we are able to extract the domains contained in the URLs of the tweets in our datasets, and match it against NewsGuard data to give it a rating. Each tweet is then attributed an action of T if it shares news from trustworthy outlets, and U if it shares news from websites that are considered as not trustworthy by NewsGuard:

$$AT(x) = \begin{cases} \text{U if } NG(D(x)) \leq 60 \\ \text{T if } NG(D(x)) > 60 \end{cases} \qquad \forall x \in D_s \tag{3.7}$$

where $NG(\cdot) : \Psi^n \mapsto \mathbb{N}$ maps n web domains to the NewsGuard trustworthiness score of the first that matches the NewsGuard list. This means that tweets containing more than one URL that matches the NewsGuard list are attributed the action category of the first URL (in the order they appear in the text of the tweet) that matches the list. Tweets for which the domain of the URL is not present in the NewsGuard list are discarded from D_s.

Note that actions are the same across strata and actors.
In the following, we define the set of all possible actions as $S = \{a_j\}_{j=1}^{L}$. For the T/U actions defined in Equation 3.7, we have $L = 2$ and $S = \{U, T\}$.

[3] https://www.newsguardtech.com/ratings/rating-process-criteria/

3.3 Time series creation

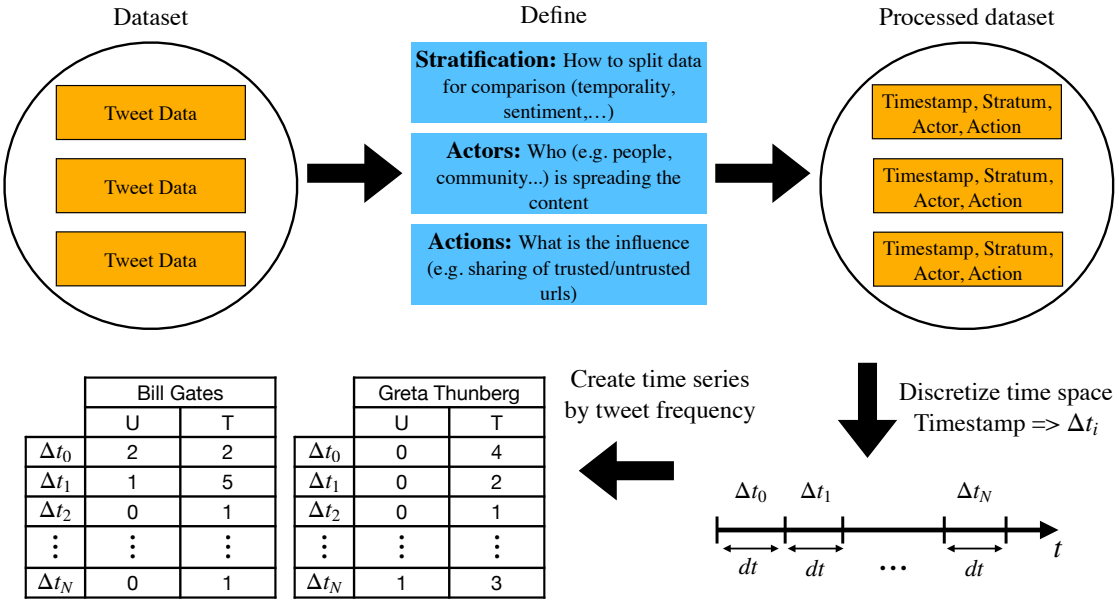


Figure 3.2: Data processing pipeline used to create time series of actions frequency per actor. This is a simplified example where only 2 actors are shown.

After having defined a group (stratum), actor and action for each tweet in the dataset, we compute time series of action frequency per actor, per stratum. To this end, for each of the strata s, we discretize the time period T_s. We define T_0 and T_f as the start and end of the interval T_s respectively, i.e. $T_s = [T_0, T_f]$. For a given *time resolution* dt, we create bins $\{\Delta t_i\}_{i=0}^{N}$ of size dt such that

$$\begin{aligned} \Delta t_i &= [t_i, t_{i+1}[\ , \quad i = 0, 1, ..., N \\ t_i &= T_0 + i \cdot dt \ , \quad i = 0, 1, ..., N+1 \end{aligned} \tag{3.8}$$

We enforce $t_N < T_f \leq t_{N+1}$, i.e. the last bin Δt_N must contain T_f. The number of bins is determined implicitly by setting dt, and depends on the T_0 and T_f observed.

Then, for each actor inside the set H_s corresponding to the stratum s, we create time series for each action by counting the number of tweets for which the timestamp is contained in each of the bins $\{\Delta t_i\}_{i=0}^{N}$. The time series creation process is similar to computing a

histogram of the tweets with bins $\{\Delta t_i\}_{i=0}^{N}$ for a given stratum, actor and action. Let $X_{a_j}^{A} \in \mathbb{R}^{N+1}$ be the time series for actor $A \in H_s$ and action $a_j \in S$:

$$(X_{a_j}^{A})_i = |\{x \in D_s \mid AC(x) = A\,,\ AT(x) = a_j\}|\,, \quad i = 0, ..., N \tag{3.9}$$

where $|\cdot|$ denotes the cardinal number of a set (i.e. the number of elements in the set). Note that the number of bins (thus the time series length) is similar for all strata D_s corresponding to the same dataset D (see Section 3.2.1).

A visual representation of the time series creation process is given in Figure 3.2. For one dataset D, one ends up with a total N of different time series:

$$N = L \cdot \sum_{s \in \{b,d,a\}} |H_s| \tag{3.10}$$

where each part of the sum consists of the time series inside a different stratum.

3.4 Coupling inference methods and influence graphs

From the time series derived in Section 3.3, our goal is now to use coupling inference methods in order to detect how actions from one actor influenced actions from another actor. Suppose we have two actors, A and B, and actions $\{a_j\}_{j=1}^{L}$. Then the influence A exerted on B is defined as the square matrix :

$$(I_{A,B})_{i,j} = d(X_{a_i}^{A}, X_{a_j}^{B})\,, \quad i, j = 1, ..., L \tag{3.11}$$

where $d(\cdot,\cdot)$ is a coupling inference method between time series. Note that in general such methods are not symmetric, hence $I_{A,B} \neq I_{B,A}$. In the case of T/U actions, the explicit form is

$$I_{A,B} = \begin{bmatrix} d(X_U^A, X_U^B) & d(X_U^A, X_T^B) \\ d(X_T^A, X_U^B) & d(X_T^A, X_T^B) \end{bmatrix} \tag{3.12}$$

We then construct an *influence graph* for each stratum as follows :

- Compute the pairwise influence matrix between all actors, i.e. use Equation 3.11 to compute all influence relationship between all actors.
- Define a directed graph $G = (V, E)$ where nodes V are the actors, and there is a directed edge $e = (A, B)$ if the influence matrix between actors A and B has at least one non-zero component. In this case, attribute matrix $I_{A,B}$ as the edge weight.

Figure 3.3 describes the graph creation process from the time series. In the end, one obtains as much influence graphs as there are strata. In this work we consider two different coupling inference methods $d(\cdot,\cdot)$ to compute the graphs: *transfer entropy* (TE) and *joint distance distribution* (JDD).

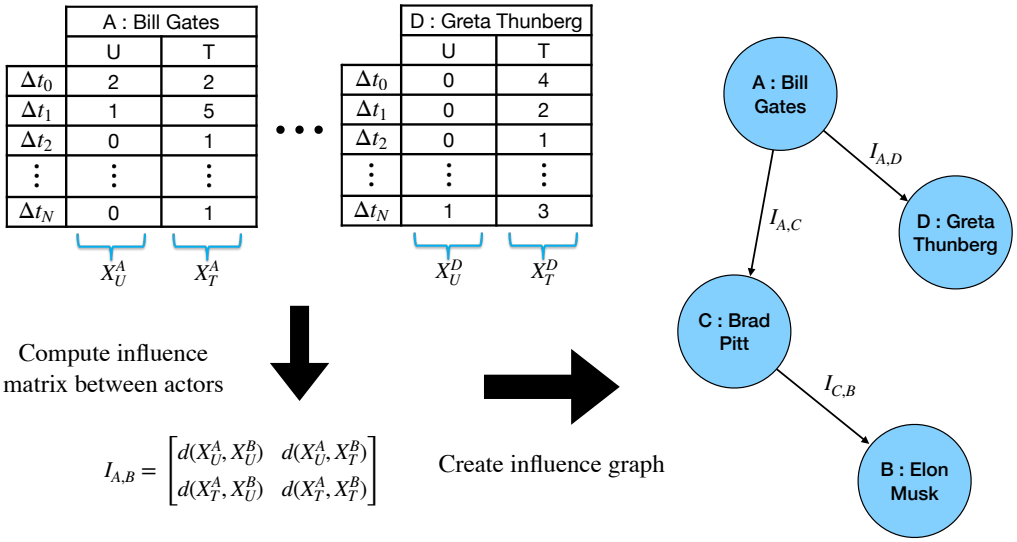


Figure 3.3: Influence graph creation from time series of actions per actor. This is a simplified example with only 4 actors.

3.4.1 Transfer Entropy

The first method we use is transfer entropy. Introduced by Schreiber [30] in 2008, it is formally defined between two random discrete variables X and Y with values in sample space Ω_X and Ω_Y respectively as :

$$T_{X\to Y} = \sum_{(y_{n+1}, y_n^{(k)}, x_n^{(l)})} p(y_{n+1}, y_n^{(k)}, x_n^{(l)}) \log \frac{p(y_{n+1}|y_n^{(k)}, x_n^{(l)})}{p(y_{n+1}|y_n^{(k)})} \tag{3.13}$$

where y_n and x_n respectively represent the state of variable Y and X at time n, and $y_n^{(k)} = (y_n, ..., y_{n-k+1})$ and $x_n^{(l)} = (x_n, ..., x_{n-l+1})$ are short-hand for k-dimensional and l-dimensional delay embedding vectors of state. The sum runs on all possible combinations of states that variables X and Y can assume at adjacent times. This means that the sum runs on all possible vectors in the space $\Omega_Y \times (\Omega_Y)^k \times (\Omega_X)^l = (\Omega_Y)^{k+1} \times (\Omega_X)^l$ such that

$(y_{n+1}, y_n^{(k)}, x_n^{(l)}) \in (\Omega_Y)^{k+1} \times (\Omega_X)^l$.

In practice, the usual choices for k and l are almost always $k = l$, and most of the time this reduces even further to $k = l = 1$ for computational reasons, which means 1-dimensional delay. In the following, we also decide to use $k = l = 1$. This means that transfer entropy reduces to

$$T_{X\to Y} = \sum_{(y_{n+1},y_n,x_n)\in(\Omega_Y)^2\times\Omega_X} p(y_{n+1}, y_n, x_n) \log \frac{p(y_{n+1}|y_n, x_n)}{p(y_{n+1}|y_n)} \tag{3.14}$$

However, in our case the variables X and Y represent the tweet frequency and take values in $\Omega_X = \Omega_Y = \mathbb{N}$. Thus, in order to simplify the sample space and limit the computations, we follow an approach used in computational neuroscience [25], and we binarize our variables. This means that instead of looking at the number of tweets posted in each time bin Δt_i, we check if *at least one tweet was posted* in the bin Δt_i (similarly to the approach of at least one spike in the time interval in neuroscience). Mathematically speaking, this means that we transform the N-dimensional time series $X \in \mathbb{R}^N$ using

$$X_i = \begin{cases} 0 & \text{if } X_i = 0 \\ 1 & \text{if } X_i > 0 \end{cases} \qquad i = 1, ..., N \tag{3.15}$$

Using such a binarization of the tweet frequencies, Equation 3.14 reduces to

$$T_{X\to Y} = \sum_{(y_{n+1},y_n,x_n)\in\{0,1\}^3} p(y_{n+1}, y_n, x_n) \log \frac{p(y_{n+1}|y_n, x_n)}{p(y_{n+1}|y_n)} \tag{3.16}$$

in which the triplet (y_{n+1}, y_n, x_n) can take values in at most $2^3 = 8$ possible states. This reduces the computational cost. In order to evaluate the probabilities in Equation 3.16, we compute the histogram of realization of the states for each of the 8 possible states, and deduce the probabilities from it. This is because we can only observe the time series for one actor once, in contrast to repeating a stochastic experiment multiple times, and using Monte-Carlo methods to estimate probabilities of the underlying time series for example.

3.4.2 Joint Distance Distribution

In order to efficiently use transfer entropy, we had to make a few sacrifices, binarizing the tweet count and naively estimating probabilities due to lack of data. For this reason, we investigate other coupling inference methods that may be simpler and do not have the previously mentioned drawbacks. We now present joint distance distribution (JDD), a method proposed by Amigó and Hirata [1] in 2017. It has the advantage of relying on distances in high-dimensional spaces instead of probabilities, which make it a very attractive choice

Algorithm 1: Joint Distance Distribution (JDD) Algorithm

Input: Time series $\{X_i\}_{i=1}^N \in \mathbb{R}^N$ and $\{Y_i\}_{i=1}^N \in \mathbb{R}^N$, distance $\delta(\cdot,\cdot)$, embedding size d, number of bins B.

for $i = 1, 2, ..., N - d + 1$ **do**
  Compute vectors $x_i = (X_i, X_{i+1}, ..., Xi + d - 1) \in \mathbb{R}^d$
  Compute vectors $y_i = (Y_i, Y_{i+1}, ..., Yi + d - 1) \in \mathbb{R}^d$
end
for $i = 1, 2, ..., N - d + 1$ **do**
  for $j = i, i + 1, ..., N - d + 1$ **do**
    Compute distance $\delta(x_i, x_j)$
    Compute distance $\delta(y_i, y_j)$
  end
end
Normalize each distance collection $\delta(x_i, x_j)$ and $\delta(y_i, y_j)$ into $[0, 1]$
Divide evenly the interval $[0, 1]$ of distances $\delta(x_i, x_j)$ into $2B$ bins I :
for $b = 1, 2, ..., 2B$ **do**
  Define interval $I_b = \left(\frac{b-1}{2B}, \frac{b}{2B}\right]$
  $\delta_{\min}(b) = \min_{1 \le i < j \le N-d+1} \{\delta(y_i, y_j) \mid \delta(x_i, x_j) \in I_b\}$
end
$\Delta = \{\delta_{\min}(B + b) - \delta_{\min}(b) \mid b = 1, 2, ..., B\}$
Test the null hypothesis H_0 : the mean of Δ is 0 using a right-sided t-test at confidence level α

Output: p-value for the hypothesis and binary variable : presence of a coupling $X \to Y$ if H_0 is rejected, no such coupling otherwise.

in our case where probabilities cannot be correctly estimated. The exact procedure is described in Algorithm 1. The idea behind this algorithm relies on Takens's delay embedding theorem [36]. For more details, the reader is referred to the cited material.

We use $d = 5$, $B = 10$, and we take the squared Euclidean distance in $\mathbb{R}^d$ as the distance $\delta(\cdot,\cdot)$. Moreover, note that the output of this procedure is discrete (coupling or no coupling) instead of continuous, in contrast to transfer entropy for which we obtain a real value quantifying the coupling strength.

3.4.3 Surrogate testing

The success of coupling inference methods is however unreliable. The detection of the coupling between two variables depends on the strength of the coupling itself, and the amount

of noise present in each observed time series. This issue also extends to studies other than causality. For this reason, *surrogate data testing* [17, 37] plays a crucial role in various methods (not only coupling inference methods), as it allows for thorough statistical evaluations to determine that the results are not just random, but are actually a distinctive characteristic of the system in question. The ideal solution would be to calculate the value of the same measure for many instances of the system's dynamics and derive a precise estimate from the resulting distribution of values.

In practice, we do not have access to multiple instances of the same dynamics. The surrogate data method involves comparing a specific aspect of the data (a distinguishing statistic) with the distribution of that same aspect found in a set of fabricated signals (surrogates). These surrogates have the same attributes as the original data set, but lack the specific property that is being examined.

In our case, since the coupling inference methods we use are searching for relations in the temporal coherence of the time series, we simply use surrogates corresponding to random permutations of the time series indices, i.e. we destruct the temporal relation between them. Explicitly, we use the procedure in Algorithm 2

Algorithm 2: Surrogate testing procedure

Input: time series $X \in \mathbb{R}^N$ and $Y \in \mathbb{R}^N$, coupling inference method $d(\cdot,\cdot)$, threshold T, statistic $S(\cdot)$, number of surrogates M

Compute $d(X,Y)$
if $d(X,Y) > T$ **then**
  Initialize list L of size M
  for $i = 1, 2, ..., M$ **do**
    Initialize X' randomly shuffled permutation of X
    $L[i] = d(X', Y)$
  end
  Compute statistic $B = S(L)$ `// e.g. median or maximum`
  if $d(X,Y) > B$ **then**
    return 1 `// The measure` $d(X,Y)$ `is significant`
  else
    return 0 `// The measure` $d(X,Y)$ `is not significant`
  end
end
return 0 `// If we end-up here` $d(X,Y)$ `was lower than the threshold`

Output: Binary variable : 1 for a significant coupling $X \to Y$, 0 otherwise

Note that for joint distance distribution, we use the p-value as result of $d(X,Y)$, and

thus use comparisons $d(X,Y) < T$ and $d(X,Y) < B$ in the conditional "if" branches respectively (instead of $d(X,Y) > T$ and $d(X,Y) > B$).

We use a threshold on the value of the measure before testing for surrogates. This approach has two advantages. First, it allows to reduce the time complexity of the whole method. Indeed, using M surrogates will multiply the computational time by M without threshold. Secondly, it allows consistency between TE and JDD. JDD is formulated as an hypothesis testing, so we still need to use a threshold to reject the hypothesis or not, even in the presence of surrogate testing. For this reason, we can use a similar approach for TE.

In order to choose values for the threshold, and a good statistic $S(\cdot)$ for the collection of surrogate values, we test different combinations on random time series. Using both false positives and true positives results, we could draw the Receiver Operating Characteristic (ROC) curves for each statistic $S(\cdot)$ and find the optimal parameters that would maximize true positives while minimizing false positives. Figure 3.4 shows the number of false positive obtained for different combinations. We first try to also find the number of true positives for a predefined synthetic temporally linked system, such as

$$Y_i = \begin{cases} 2X_{i-1} + \sigma + 2X_{i+1} & \text{if } i = 2, ..., N-1 \\ \sigma + 2X_{i+1} & \text{if } i = 1 \\ \sigma + 2X_{i-1} & \text{if } i = N \end{cases} \tag{3.17}$$

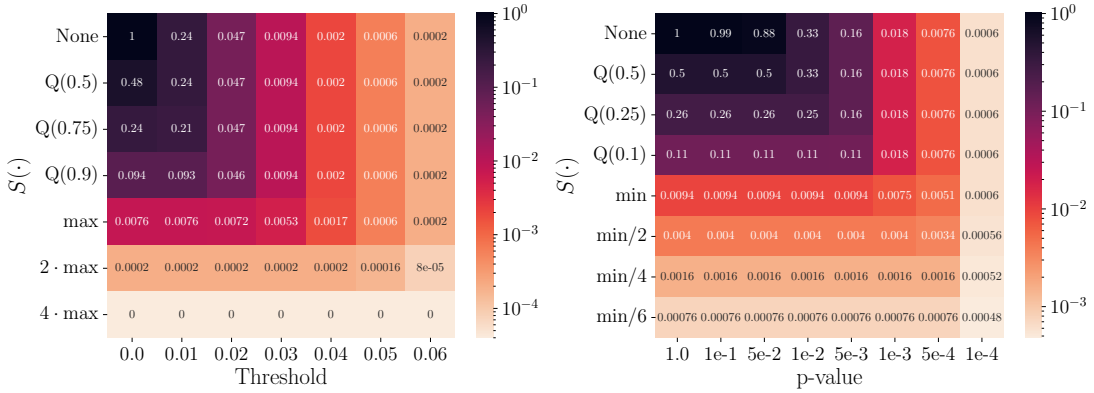
where σ represents random noise.

However, we noticed that the ROC curves highly depend on the random system (Equation 3.17), i.e. changing the constants and/or using only X_{i-1} or X_{i+1} would completely change the shape of the curves. Thus, choosing values this way would never generalize to our datasets. For this reason, we only chose thresholds and statistic using false positive rate in Figure 3.4. Choosing a combination giving 0 false positives is however very likely to fail to detect any true positive as well. For this reason, we choose reasonable combinations giving approximately the same false positive rate for both methods.

In the following, we chose a threshold $T = 0.04$ and statistic $S(\cdot) = \max(\cdot)$ for TE, and a p-value of 0.001 and statistic $S(\cdot) = \min(\cdot)/4$ for JDD.

3.5 Influence cascades

We call *influence cascade* the path of the actors and actions starting from an actor A who acted on their own (i.e. motivated by an external stimulus) and who influenced other actors to take actions, and the actions that those actors then influenced other actors to take, and so forth until the last actor's actions do not influence anyone else. From an influence



(a) TE (b) JDD

Figure 3.4: False positive rate for different thresholds and statistic $S(\cdot)$ for TE (a) and JDD (b). Here $Q(p)$ represent the quantile of the distribution at probability p, i.e. $Q(0.5)$ is the median, $Q(0.75)$ is the 3rd quartile etc...

graph, influence cascades start from nodes having 0 indegree but non-zero outdegree, i.e. actors who were not influenced by anyone but influenced other. From such actors, one may then simply follow the graph links until there are no more edges to connect new actors. For example, in the graph depicted in Figure 3.3, one may observe an influence cascade starting at actor A, Bill Gates, at level 0 of the cascade, then reaching actors C and D (Brad Pitt and Greta Thunberg) in level 1, and finally stopping with actor B, Elon Musk, in level 2. We call the last level (in this case level 2) the maximum depth of the cascade.

However, simply extracting the actors (and edges between them) present in the cascades does not allow us to describe what type of influence was exerted between them (i.e. what components on the influence matrix were non-zero). For this reason, similarly to Senevirathna et al. [31], we decide to aggregate actors at each depth (level) of the cascades, and compute the amount of influence of each type.

Let $\Lambda_i \quad i = 0, 1, ..., n$ be the set of all actors present at level i of the cascade (n being the maximum depth). We define

$$\Gamma_{i,i+1} = \{I_{A,B} | A \in \Lambda_i, B \in \Lambda_{i+1}\} \qquad i = 0, ..., n-1 \tag{3.18}$$

as being the set of influence matrices from actors at level i of the cascade to actors at level $i+1$. Then, the normalized social influence an action $a_k \in S$ has on an action $a_l \in S$ between levels i and $i+1$ of the cascade is defined as

$$\gamma_{i,i+1}(a_k, a_l) = \frac{\sum\limits_{I \in \Gamma_{i,i+1}} I_{k,l}}{\sum\limits_{N=0}^{n-1} \sum\limits_{I \in \Gamma_{i,i+1}} I_{k,l}} \tag{3.19}$$

Having defined these quantities, we now visualize the influence cascades through a Sankey diagram. In this diagram, each node at each level represents an action $a_j \in S$, and the flows between nodes of action $a_k \in S$ in level i and node of action $a_l \in S$ in level $i+1$ represent the quantity $\gamma_{i,i+1}(a_k, a_l)$ defined in Equation 3.19. Figure 3.5 shows an example of such a Sankey diagram.

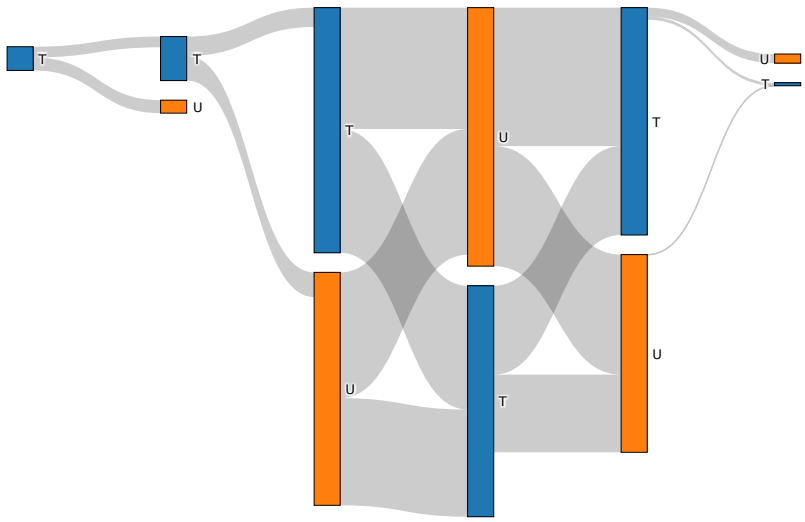


Figure 3.5: Example of an influence cascade, in the case of actions being defined as $\{T, U\}$. The cascade has 6 levels ($n = 5$), and starts from a user sharing Trustworthy (T) URLs. The height of the bars (nodes) is the maximum between the amount of flow in and out.

3.6 Influence measures

In this part, we describe the influence measures we use to classify actors. We first present some usual measures, and then describe the ones we derive from the influence graphs defined above.

3.6.1 Traditional measures

In order to validate and compare the results from our approach, we also compute some traditional measures of influence. In Chapter 2, we already described some of the possible techniques. We decided to focus on the 4 following measures :

- *tweet count*: the total number of tweets, usually reflecting bots or extremely virulent users. In disinformation campaigns, this is a useful measure as it allows to quickly find the most aggressive spammers of one (or multiple) content.
- *follower count*: the total number of followers, obtained at the time at which we obtained the data from Twitter. This measure presents users who have a very large audience.
- *retweet count*: the total number of retweets, reflecting user ability to create content that others find interesting.
- *I score*: influence score, defined by Romero et al. [29], which is a good predictor of the number of times a URL shared by someone in a tweet will be clicked on.

Out of those 4 influence measures, the first 3 are straightforward to obtain, as they are directly available from the Twitter data. The I score however needs to be computed, and we now present the procedure to obtain it.

The Influence-Passivity (IP) algorithm The I score we use is in fact the Influence (I) component of the Influence-Passivity (IP) algorithm from Romero et al. [29].

First, generate the following graph $G = (V, E)$: the nodes V are users who tweeted at least 3 URLs. The edge $e = (i, j)$ exists if user j retweeted a URL posted by i at least once. In this case, the edge has weight $w_{i,j} = \frac{S_{i,j}}{Q_i}$ where Q_i is the number of URLs that i mentioned and $S_{i,j}$ is the number of URLs mentioned by i and retweeted by j. Note that we do not take into account retweets from one user to themselves, meaning that we do not allow self-loops in the graph.

For every edge $e = (i, j) \in E$, define the acceptance rate $u_{i,j}$ and rejection rate $v_{i,j}$ in the following way:

$$u_{i,j} = \frac{w_{i,j}}{\sum\limits_{k:(k,j)\in E} w_{k,j}}$$
$$v_{i,j} = \frac{1 - w_{i,j}}{\sum\limits_{k:(i,k)\in E} (1 - w_{i,k})} \tag{3.20}$$

where $k{:}(k,j) \in E$ means all indices k such that the edge $(k,j) \in E$, i.e. the edge (k,j) exists.

Using these quantities, one may then follow the iterative procedure described in Algorithm 3 to obtain the Influence (I) and Passivity (P) scores of each nodes of the graph G, that is each user. Both these scores are bounded in the interval $[0,1]$. The influence score is a measure of the importance of the user in the graph, while the passivity of a user describes how difficult it is for other users to influence them. We only keep the I score in our case, and discard the P score.

Algorithm 3: IP algorithm

Input: Graph $G = (V, E)$, acceptance and rejection matrices u and v, tolerance α.

Initialize $I^0 = (1, 1, ..., 1) \in \mathbb{R}^{|V|}$
Initialize $P^0 = (1, 1, ..., 1) \in \mathbb{R}^{|V|}$
Initialize residual $R = \infty$
Initialize iteration count $m = 0$
while $R > \alpha$ **do**
    $m = m + 1$
    for $i = 1, ..., |V|$ **do**
        $P_i^m = \sum\limits_{j:(j,i)\in E} v_{j,i} I_j^{m-1}$ `// Use` I^{m-1}`, i.e. previous iteration`
        $I_i^m = \sum\limits_{j:(i,j)\in E} u_{i,j} P_j^m$ `// Use` P^m`, i.e. current iteration`
    end
    Normalize each component of I^m by $\sum_{k=1}^{|V|} I_k^m$
    Normalize each component of P^m by $\sum_{k=1}^{|V|} P_k^m$
    Update residual $R = ||I^m - I^{m-1}||_1 + ||P^m - P^{m-1}||_1$
end

Output: I^m and P^m

In Algorithm 3, $||\cdot||_1$ represent the L-1 norm, i.e. the sum of the absolute values of a

vector. Algorithm 3 converges rather quickly (tens of iterations) to a reasonable tolerance that we set to $\alpha = 10^{-3}$, as can be seen in Figure 3.6.

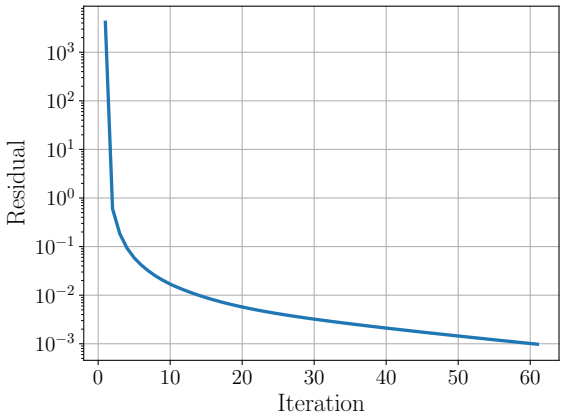


Figure 3.6: Convergence plot of the IP algorithm applied to the Skripal dataset. The procedure converges in about 60 iterations to the prescribed residual of 10^{-3}.

3.6.2 Centrality measures based on influence graphs

We now describe the influence measures that we derive from the influence graphs defined in Section 3.4. Since the graphs directly capture coupling relationships between actors' actions, it is natural to derive influence from the graph structure and topology. In particular, we focus on the following measures:

- *outdegree*: for each node (actor) in the influence graph, outdegree directly indicates how many other actors this actor influenced.
- *betweenness*: the betweenness centrality measures how important a node is in the graph, in the sense of shortest path to every other node. In our case, it is an indication of how well one actor serves as a bridge to influence many other users.

However, since the influence graphs we defined in Section 3.4 have matrices as edge weights (see Equation 3.11), we can make some distinctions between different types of influence. Indeed, we can measure outdegree and betweenness based on the influence graph (with matrices on edges), or we can split the graph into L^2 sub-graphs induced by each component of the influence matrices (i.e. components $d(X^A_{a_i}, X^B_{a_j})$ for two actors $A, B \in D_s$ and actions $a_i, a_j \in S\,,\ i, j = 1, ..., L$). For the U/T actions defined in Equation 3.7, we have 4 influence types:

- T-T: represents an *echo chamber* of trustworthy news sharing.

- T-U: represents a *credibility cross-over* of people reacting to trustworthy news by sharing untrustworthy ones.
- U-T: represents a *credibility cross-over* of people reacting to untrustworthy news by sharing trustworthy ones.
- U-U: represents an *echo chamber* of untrustworthy news sharing.

In turn, we define outdegree and betweenness for each of the 4 influence types presented, by computing these measures on the sub-graphs corresponding to each component of the influence matrices.

Chapter 4

Results

We now present all the results we obtained. We first focus on the results we derived with the Skripal dataset in Section 4.1. Then, Sections 4.2 and 4.3 show the results we got using the COP26 and COP27 datasets, respectively. Section 4.4 depicts the impacts of user aggregation on the influence graphs. In Section 4.5, we dive into the web domains and articles shared by users who were found to be the most influential by our methods.

4.1 The Skripal case

We begin by describing results obtained for the Skripal dataset (Section 3.1.1).

4.1.1 Dataset description

We obtained data related to the Skripal incident for the months of March and April 2018. However, the poisoning only happened on March 4, 2018. In their work, Ramsay and Robertshaw [26] only study articles published for the month of March, from March 4 to March 31. We extend this period, studying the coverage from March 4 to April 14 included. We divide this interval into 3 distinct 14 days periods, that we call before, during and after the disinformation campaign (see Table 3.2). In practice, the timeline of the campaign is obviously not as clear, however it seems reasonable to assume that the campaign started rather slowly at the beginning, then increased in intensity during mid/end of March (around the time at which Theresa May attributed the responsibility to Russia), and then started slowing down again. We will use this distinction for stratification. We use the T/U action classification described in Section 3.2.3, and actors are defined inside each strata as individual users having posted at least 3 tweets during the corresponding period (Equation 3.5).

Figure 4.1 describes actions and actor distributions across the strata. Figure A.1 shows actors with the most followers in the dataset. Table 4.1 shows the ranking of the top 10

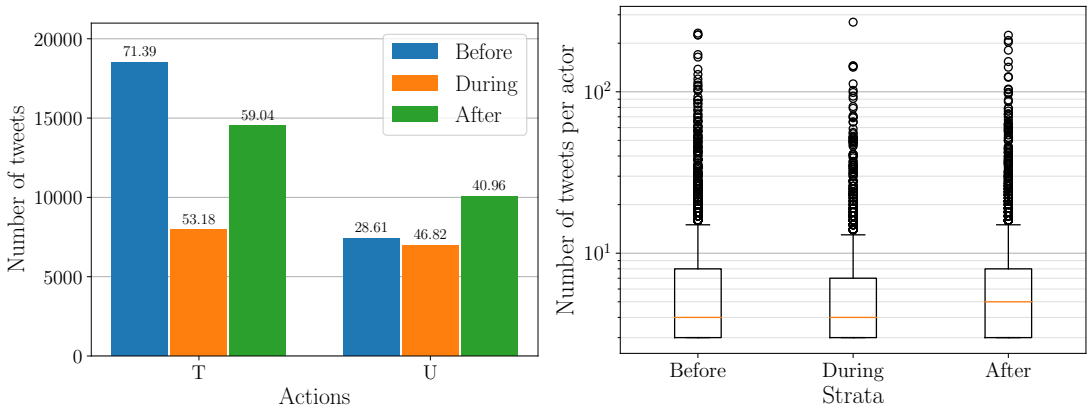


Figure 4.1: (a) Actions distribution across strata (the proportion inside each stratum is indicated as a number above each bar) and (b) distribution of the number of tweets per actor across strata. This data represents the Skripal dataset.

most important users, for each of the traditional measures described in Section 3.6.1.

4.1.2 Influence graphs using joint distance distribution

Using the T/U action classification, the influence between 2 actors is defined as in Equation 3.12. There are 4 types of influence, namely T-T, T-U, U-T, and U-U. Each of these types represents a different kind of edges in the influence graphs. Figure 4.2 shows the number of edges of each type obtained in the graphs, as well as the mean number of actors that were reached at each depth of the influence cascades derived from the influence graphs. Since the action distribution is unbalanced (see Figure 4.1), not all kinds of edges can exist between each actor (some actors may not have tweeted untrustworthy or trustworthy URLs). For this reason, Figure 4.2 b) shows the count of edges of each type, normalized by the number of reachable edges of this type in the graph.

Table 4.2 shows the top 10 most influential actors according to centrality measures derived from the influence graphs (see Section 3.6.2). In some cases, less than 10 actors have non-zero value for one or more centrality measure. In this case, we indicate such missing data with a dash (-). Table A.1 shows the same data.

Before			
tweet count	follower count	retweet count	I score
ferozwala	nytimes	Billbrowder	CraigMurrayOrg
RLSRUSSIANNEWS	CNN	RT_com	Billbrowder
_dpaj	BBCBreaking	CraigMurrayOrg	PiersRobinson1
_NoMoreExcuses	TheEconomist	Independent	VanessaBeeley
Independent	Reuters	NBCNews	ShoebridgeC
RT_com	FoxNews	DrDenaGrayson	MaxBlumenthal
mlnangalama	WSJ	BBCBreaking	MoonofA
wherepond	TIME	CBSNews	NeilClark66
5150power	AP	KremlinTrolls	BBCBreaking
newsbloktwit	HuffPost	guardian	tnewtondunn

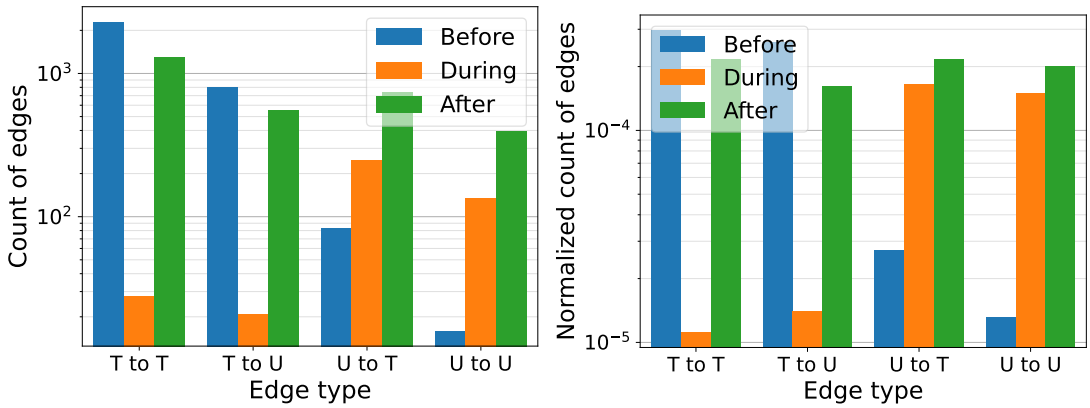
During			
tweet count	follower count	retweet count	I score
ferozwala	nytimes	RT_com	CraigMurrayOrg
mlnangalama	Reuters	CraigMurrayOrg	NeilClark66
RLSRUSSIANNEWS	WSJ	NeilClark66	PaulCraigRobert
_dpaj	cnni	ShoebridgeC	ShoebridgeC
RotenbergBros	guardian	Ian56789	RusEmbUSA
srbnews0	CBSNews	RTUKnews	21WIRE
5150power	NBCNews	Billbrowder	ProfessorsBlogg
new16media	AJEnglish	haynesdeborah	_irishrepublic
RT_com	Newsweek	Independent	OffGuardian0
newsbloktwit	FRANCE24	_irishrepublic	haynesdeborah

After			
tweet count	follower count	retweet count	I score
_dpaj	cnnbrk	RT_com	CraigMurrayOrg
EdWardMDBlog	nytimes	RTUKnews	skwawkbox
srbnews0	CNN	CraigMurrayOrg	NeilClark66
ferozwala	BBCBreaking	PrisonPlanet	RussiaInsider
RT_com	BBCWorld	NeilClark66	PrisonPlanet
mlnangalama	Reuters	ShoebridgeC	PaulCraigRobert
ali9l9	FoxNews	BBCBreaking	JohnWight1
RLSRUSSIANNEWS	WSJ	Independent	CRG_CRM
jondknight	TIME	Ian56789	ShoebridgeC
notiven	washingtonpost	zerohedge	21WIRE

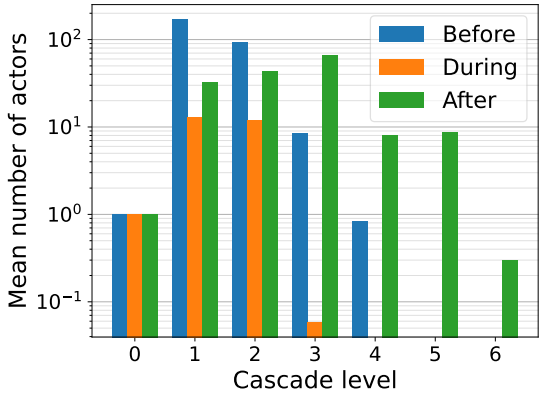
Table 4.1: Top 10 most important users using traditional influence measures, for each of the strata. This table was obtained using the Skripal dataset.

4.1.3 Influence graphs using transfer entropy

Figure 4.3 represents the count of edges in the case of the graphs obtained using TE as coupling inference measure. As above, Tables 4.3 and A.2 display the top 10 most influential actors according to centrality measures, using the whole graph and each edge type respectively.



(a) (b)



(c)

Figure 4.2: a) count of each type of edges, b) count of each edge type normalized by the number of reachable edges, and c) mean number of actors at each level of the influence cascades derived from the graph. Data obtained using JDD on the Skripal dataset.

4.1.4 Correlations between measures

Figure 4.4 presents the correlation between each of the traditional and centrality based (using the TE graph) influence measures, for each of the strata.
We define as correlation the following *intersection* or *overlap* measure: let Ω_i bet the set consisting of the top 50 most influential users according to measure i. Then we compute the overlap $\mathrm{Ov}(\Omega_i, \Omega_j)$ between the sets of actors Ω_i obtained by measure i and Ω_j obtained by measure j as:

Before		During		After	
outdegree	betweenness	outdegree	betweenness	outdegree	betweenness
conju_re	Daily_Star	RT_com	RT_com	Angelus1701	Angelus1701
Daily_Star	nytimesworld	newsroll	_dpaj	JudeJack	JudeJack
Telegraph	Telegraph	ferozwala	-	HillestadNils	OnlBreakingNews
nytimesworld	Redpolitics	JJorbyn	-	americandailys	gabriellesct
Juancarlos_Mike	Londied39	lisa_alba	-	starandsixpence	HillestadNils
mayasdolly	Harley_Woody	chootchyface	-	puffin1952	NewsBlogged
ReciteSocial	grauniad_news	paris_2015	-	_ThePage	puffin1952
Redpolitics	Daily_Express	RLSRUSSIANNEWS	-	NewsBlogged	wittich
Harley_Woody	warringworld	ConversationUK	-	NewsDingo	marvellous997
qkode	farhaadaarif	ALLREDToDoRoJo	-	amandasome	bassmadman

Table 4.2: Top 10 most important users using centrality measures based on the influence graphs, for each of the strata. Graph derived using JDD on the Skripal dataset.

Before		During		After	
outdegree	betweenness	outdegree	betweenness	outdegree	betweenness
peterpobjecky	NBCNightlyNews	sengeezer	sengeezer	Hadrienr974	thebrkg
NewStatesman	TheUrbanNewz	NewsAboutLife	NewsAboutLife	rlangjournalist	JiriParkes
farhaadaarif	farhaadaarif	Russianation	KremlinTrolls	JiriParkes	Ian56789
JJorbyn	NewStatesman	bdnews24	andro1711	grauniad_news	Nildam85
everydayisabirt	zhouhuasheng06	Orgetorix	Tufairi	guardian	MSNBC
notiven	newsbloktwit	notiven	RTUKnews	Joanvanderlinge	grauniad_news
NBCNightlyNews	ValuBit	RLSRUSSIANNEWS	bettabettanesi	myamigocouk	farhaadaarif
Bill_Owen	mlnangalama	psic88	Russianation	andy_s_64	WhirlwindWisdom
livier_i	notiven	RTUKnews	AlanMcpartlands	hangen_claude	TheAllRadar
CordeliaAppleb1	Kostian_V	Tufairi	RLSRUSSIANNEWS	vrai777	cgnetwork

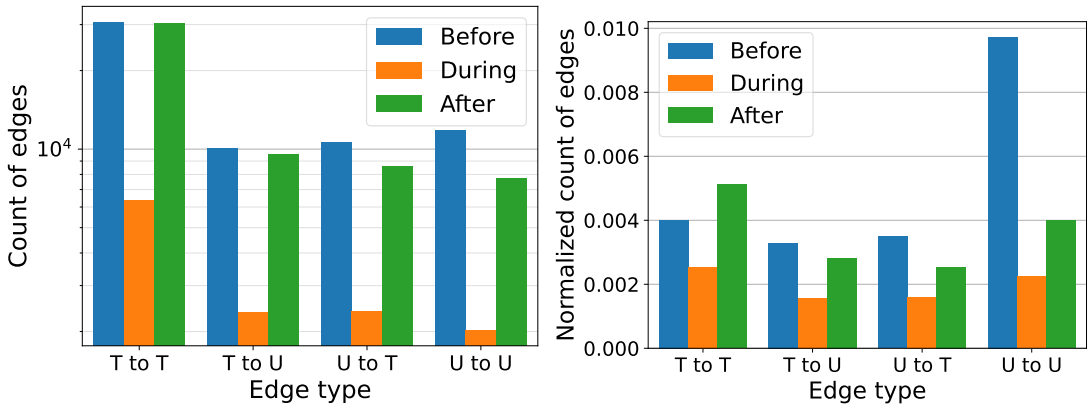
Table 4.3: Top 10 most important users using centrality measures based on the influence graphs, for each of the strata. Graph derived using TE on the Skripal dataset.

$$\mathrm{Ov}(\Omega_i, \Omega_j) = \frac{|\Omega_i \cap \Omega_j|}{|\Omega_i|} \tag{4.1}$$

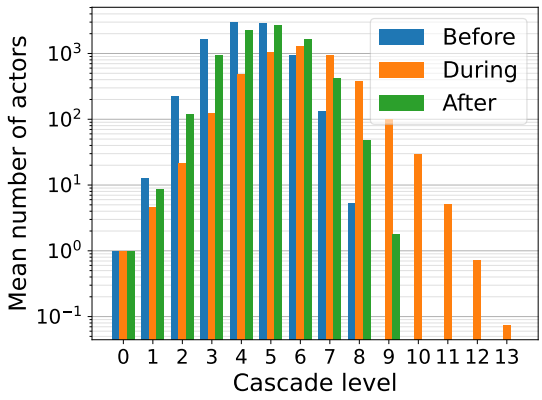
Note that Equation 4.1 does not take into account the actual ranking of each actor in the set, only their presence. We only take the 50 highest ranked actors in each set. Including too many actors could cause random ties in rank due to zero score (many users have 0 retweets or followers for example) and incur bias in the correlation between influence measures. This is also why we do not compute these correlations with centrality measures coming from the JDD graphs. Indeed, Table 4.2 shows some missing values for only the top 10 users, meaning that taking the top 50 would result in more than 40 actors being present at random because they have score 0.

4.2 COP26 dataset

In this part, we present results based on the COP26 dataset presented in Section 3.1.2.



(a) (b)



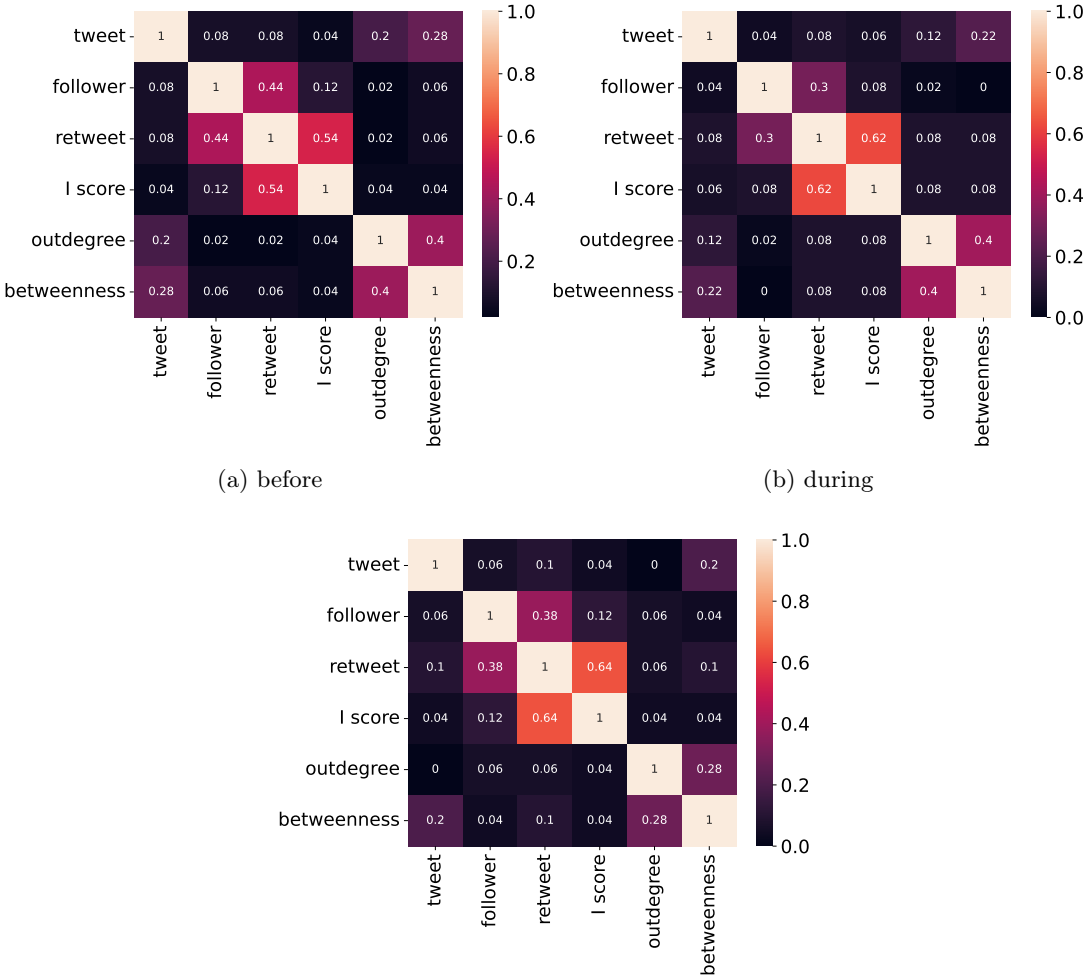
(c)

Figure 4.3: a) count of each type of edges, b) count of each edge type normalized by the number of reachable edges, and c) mean number of actors at each level of the influence cascades derived from the graph. Data obtained using TE on the Skripal dataset.

4.2.1 Dataset description

In the same manner as before, we stratify the dataset into 3 groups based on the time of the tweets. Each group spans an equally spaced period of 13 days before, during and after the COP26 summit (see Table 3.2). We use the T/U classification as actions, and actors for each stratum are defined as users having posted at least 3 tweets in the corresponding time period as in Equation 3.5.

Figure 4.5 represents the distribution of actions and actors for each of the strata. Figure A.2 represents the actors in the dataset with the most followers. Finally, Table 4.4 shows



(a) before

(b) during

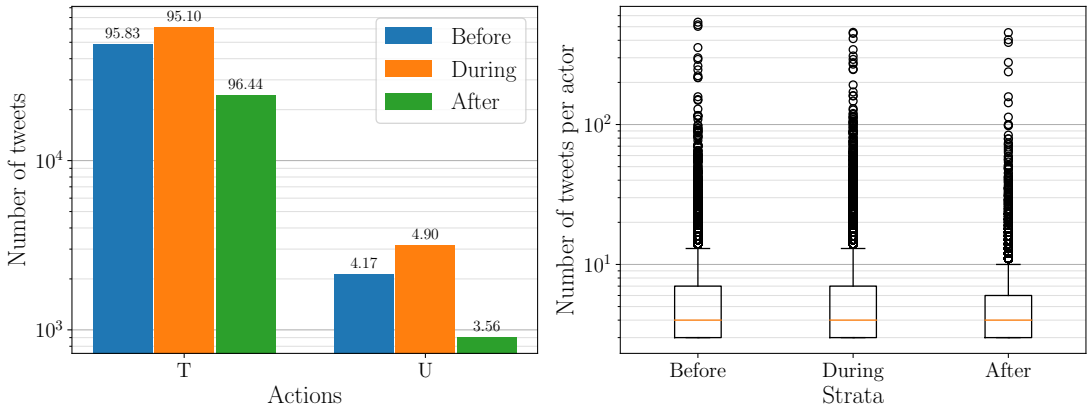
(c) after

Figure 4.4: Correlation matrix between each measure for the Skripal dataset. The centrality measures outdegree and betweenness refer to the TE graph.

the highest-ranked actors for each traditional measure and stratum.

4.2.2 Influence graphs using joint distance distribution

This part focuses on results obtained when using JDD as coupling inference measure. We now use the RandomDays dataset presented in Section 3.1.3 in order to compare results to a control describing what the influence graph looks like when there are no influence



(a) (b)

Figure 4.5: (a) Actions distribution across strata (the proportion inside each stratum is indicated as a number above each bar) and (b) distribution of the number of tweets per actor across strata. This data represents the COP26 dataset.

campaigns going on (this assumption is made based on how the RandomDays dataset was collected). Figure 4.6 shows the count and normalized count of each type of edges obtained, as well as the mean number of actors at each depth in the influence cascades derived from the graphs.

Table 4.5 shows the top 10 most influential users according to the centrality measures based on the influence graphs. Table A.3 presents the same quantities but split into each edge type category.

4.2.3 Influence graphs using transfer entropy

In this section, we use TE as the coupling inference method in equation 3.12 to derive the influence graphs. Figure 4.7 represents the distribution of each type of edges in the graphs, as well as the number of actors at each level of the influence cascades.

Tables 4.6 and A.4 present the top 10 most influential users according to centrality measures derived from the influence graphs.

4.2.4 Correlations between measures

Similarly to Section 4.1.4, we are interested in how all of these influence measures are linked between themselves. Figure 4.8 shows the correlation matrix between influence measures

Before			
tweet count	follower count	retweet count	I score
TinTincognito	CNN	mikegalsworthy	mcannonbrookes
bobhillbrain	nytimes	PaulEDawson	samanthamaiden
Eric2017w	BBCWorld	Reuters	MrKRudd
Independent	NatGeo	ProfStrachan	mikegalsworthy
rpujolvives	TheEconomist	SenWhitehouse	RDNS_TAI
great_thunberg	Reuters	washingtonpost	GeorgeMonbiot
RFrumpf	FoxNews	TIME	billmckibben
TimMelino	WSJ	CNN	BenFranta
gridpointwx	washingtonpost	nytimes	PeterKGeoghegan
ReutersScience	TIME	guardianeco	dwallacewells

During			
tweet count	follower count	retweet count	I score
bobhillbrain	NASA	rapplerdotcom	MrKRudd
TinTincognito	CNN	nytimes	GeorgeMonbiot
Independent	nytimes	Reuters	CarolineLucas
great_thunberg	BBCBreaking	CNN	PeterKGeoghegan
reno_ralph	BBCWorld	EdwardJDavey	SenWhitehouse
TimMelino	TheEconomist	BBCNews	MichaelEMann
rpujolvives	Reuters	CarolineLucas	KateAronoff
gridpointwx	FoxNews	newsmax	hausfath
indy_climate	WSJ	AP	EdwardJDavey
ReutersScience	washingtonpost	Independent	InsiderIntl

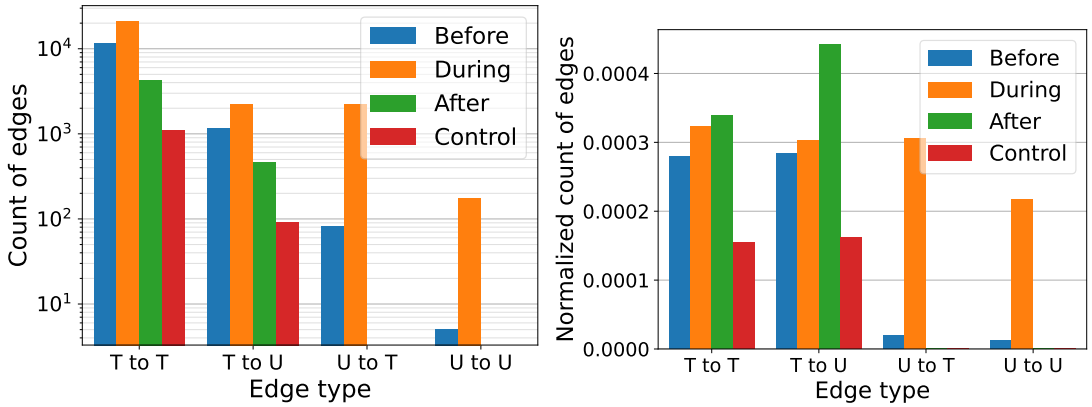
After			
tweet count	follower count	retweet count	I score
rpujolvives	CNN	PrisonPlanet	lloydalter
bobhillbrain	nytimes	PaulEDawson	guardianeco
TinTincognito	BBCWorld	washingtonpost	billmckibben
gptnshl	NatGeo	sunlorrie	GeoffreySupran
great_thunberg	TheEconomist	wef	dwallacewells
raphclimbot	Reuters	MichaelEMann	GreenRupertRead
Independent	WSJ	ClimateReality	peter_j_wood
TimMelino	washingtonpost	BBCWorld	AssaadRazzouk
gridpointwx	TIME	nytimes	MichaelEMann
_DrFrusci	Forbes	Independent	WeatherProf

Table 4.4: Top 10 most important users using traditional influence measures, for each of the strata. This table was obtained using the COP26 dataset.

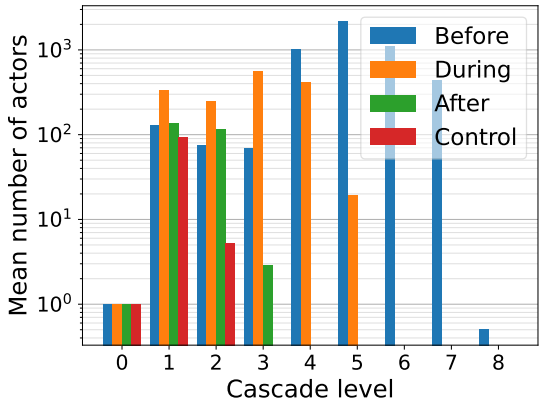
for each stratum. We use Equation 4.1 to compute correlation.

4.3 COP 27 dataset

We now present results when using our framework on the COP27 dataset presented in Section 3.1.2. We proceed in the same way as for the COP26 dataset. The strata are temporal: before, during, and after the actual COP27 summit (see Table 3.2). We use the RandomDays dataset as a control.



(a) (b)



(c)

Figure 4.6: a) count of each type of edges, b) count of each edge type normalized by the number of reachable edges, and c) mean number of actors at each level of the influence cascades derived from the graph. Data obtained using JDD on the COP26 dataset.

Actions are the T/U categories derived from NewsGuard ratings (Equation 3.7), and actors are consistently defined in the same manner as before: for each stratum, they are users having tweeted at least 3 times in the corresponding time window (see Equation 3.5).

4.3.1 Dataset description

Figures 4.9 and A.3 depict the actions and actor distributions, and represent the actors with the most followers as a wordcloud respectively. On the wordcloud, the size of the word

Before	
outdegree	betweenness
MichiganRadio	jlitwinetz
jlitwinetz	MichiganRadio
katydaigle	aawsat_eng
jilevin	jilevin
YV5SEL	katydaigle
BrendanCarton	davidtomkins
DebsF319	TurboKitty
TurboKitty	NelsonGich
LatinoLdnOnt	smorffer
Daily_Express	YV5SEL

During	
outdegree	betweenness
TheDisproof	klausammann
AugustEve2012	bdollabills
drsohailmahmood	TheDisproof
klausammann	AugustEve2012
paulinepark	paulinepark
globaltimesnews	drsohailmahmood
AandNoa	cnni
Daily_Record	bpolitics
unherd	MSNBC
EnviroEdgeNews	EnvDefenseFund

After	
outdegree	betweenness
DataAugmented	jftaveira1993
BrianMcHugh2011	BrianMcHugh2011
jftaveira1993	Surly01
EsgWire	EsgWire
business	latimes
Orgetorix	Orgetorix
latimes	AndyVermaut
AndyVermaut	business
commondreams	Eire353
insideclimate	IndianExpress

Table 4.5: Top 10 most important users using centrality measures based on the influence graphs, for each of the strata. Graph derived using JDD on the COP26 dataset.

Before	
outdegree	betweenness
GlobalUnion3	RobotChange
JunkScience	TinTincognito
HarwoodEdu	Surly01
eduCCateGlobal	HarwoodEdu
Surly01	GlobalUnion3
RobotChange	eduCCateGlobal
weatherindia	AlexWitzleben
ManishKhurana	Independent
margreis9	weatherindia
DclareDiane	RFrumpf

During	
outdegree	betweenness
eduCCateGlobal	RobotChange
HarwoodEdu	HarwoodEdu
greenprofgreen	TinTincognito
JunkScience	eduCCateGlobal
physorg_space	TimMelino
zyiteblog	gridpointwx
Surly01	highcountrynews
pablorodas	joincurby
openDemocracy	weatherindia
LatinoLdnOnt	LehtmanMaria

After	
outdegree	betweenness
HarwoodEdu	RobotChange
DanAlbas	TinTincognito
highcountrynews	GailWaldby
eduCCateGlobal	HarwoodEdu
EnvHamilton	Surly01
JM_Coppede	eduCCateGlobal
nprworld	AlexWitzleben
TinTincognito	CCLSVN
DenisPetit2233	highcountrynews
checkupcbc	CelloMomOnCars

Table 4.6: Top 10 most important users using centrality measures based on the influence graphs, for each of the strata. Graph derived using TE on the COP26 dataset.

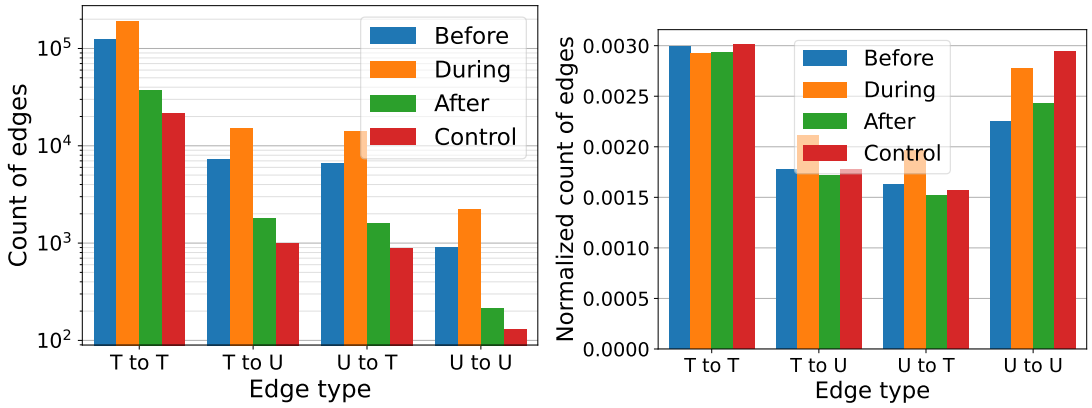
is proportional to the number of followers the actor has.

Table 4.7 shows the top 10 most important users according to the 4 traditional measures of influence described in Section 3.6.1.

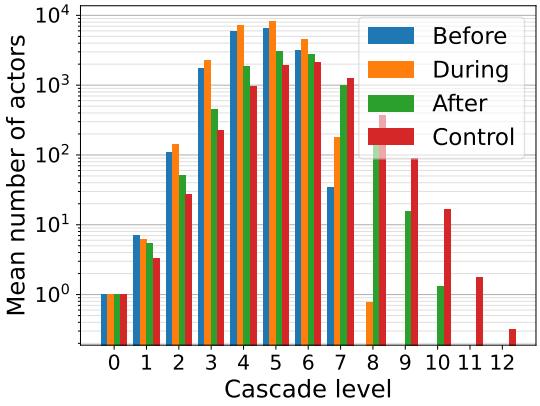
4.3.2 Influence graphs using joint distance distribution

In this section, we present the results we obtain when using our influence graphs framework on the dataset presented above. We use joint distance distribution to compute the influence matrix between each actor. Figure 4.10 shows statistics (count and normalized count) for each type of edges that we obtained after computing the influence graphs, as well as the mean number of actors at each depth of the influence cascades.

Tables 4.8 and A.5 respectively present the top 10 most influential actors using the centrality measures defined in Section 3.6.2, for the whole graph and divided into each edge



(a) (b)



(c)

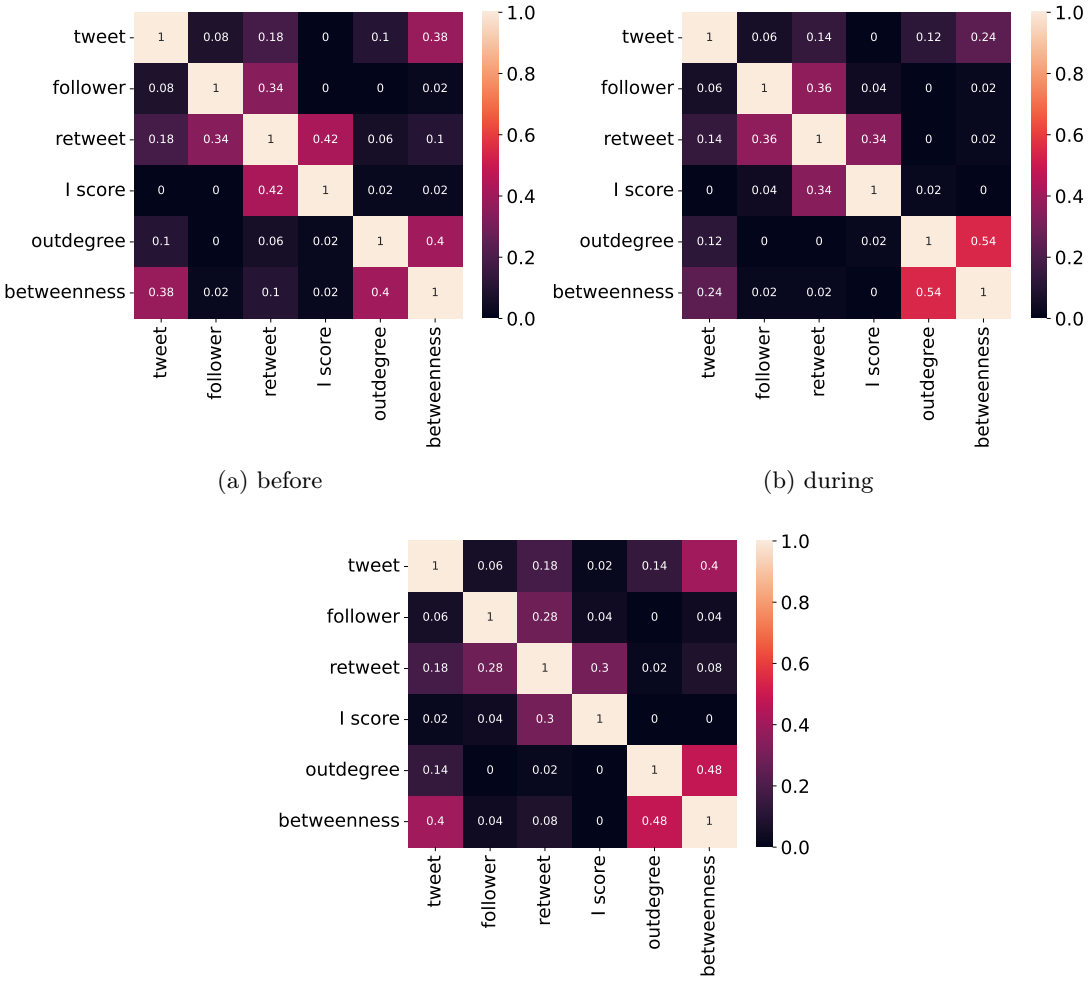
Figure 4.7: a) count of each type of edges, b) count of each edge type normalized by the number of reachable edges, and c) mean number of actors at each levels of the influence cascades derived from the graph. Data obtained using TE on the COP26 dataset.

type.

4.3.3 Influence graphs using transfer entropy

We now use transfer entropy to derive the influence graphs on the COP27 dataset. Figure 4.11 shows the number of edges in the influence graphs, as well as the mean number of users reached at each level of the influence cascades.

Tables 4.9 and A.6 depict the rank of the top 10 most influential actors as defined by



(a) before

(b) during

(c) after

Figure 4.8: Correlation matrix between each measure for the COP26 dataset. The centrality measures outdegree and betweenness refer to the TE graph.

the centrality measures derived from the influence graphs.

4.3.4 Correlations between measures

Figure 4.12 presents the correlation between each influence measure, that we compute using Equation 4.1. The centrality measures (indegree and betweenness) are the ones obtained using the transfer entropy graphs, as there are too many actors tied with a score of 0 for

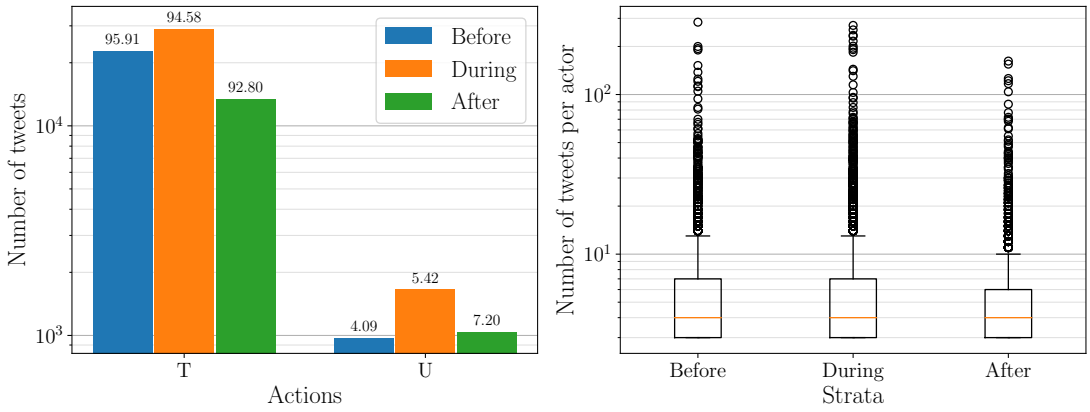


Figure 4.9: (a) Actions distribution across strata (the proportion inside each stratum is indicated as a number above each bar) and (b) distribution of the number of tweets per actor across strata. This data represents the COP27 dataset.

joint distance distribution based graphs.

4.4 Sensitivity to aggregation of actors

In this part, we focus on studying how the influence graph method behaves if we start to aggregate multiple individual users into a single actor. This is motivated by the fact that we already know that individual users are not equal in term of potential influence. An individual who posted only 3 tweets (i.e the minimum to be included as an actor), but has 0 retweets and very few followers is less likely to be influential in the network compared to someone who posted a lot of tweets, generated a lot of retweets, or has a large number of followers (i.e. a large audience) for example. Aggregating users will greatly reduce the number of actors, which in turn reduces the computational complexity of the method since it scales quadratically with the number of actors.

This allows to study the robustness of the method if actors were to be defined as a mix of individual users and communities or groups of users. For example, one may first run community detection algorithms on the social media graphs in order to find people with similar political ideas or ideologies, and then define those communities as actors in order to understand how influence is flowing between communities as well as between users.

In order to study the influence of actors aggregations, we use the following setup :

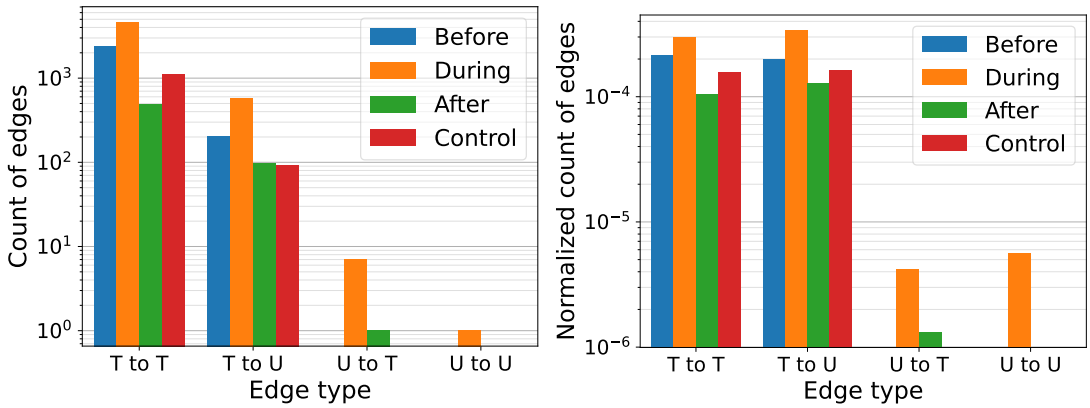
Before			
tweet count	follower count	retweet count	I score
Aaron89410013	CNN	ShellenbergerMD	RogerHallamCS21
great_thunberg	nytimes	BjornLomborg	ProfBillMcGuire
MrMatthewTodd	NatGeo	dwallacewells	CarolineLucas
indy_climate	TheEconomist	JunkScience	JamesGDyke
Independent	Reuters	toadmeister	StopCambo
bullshitjobs	FoxNews	nytimes	WSOnlineNews
BlasphemousBan1	WSJ	ProfStrachan	cflav
THE_Mr_Z	washingtonpost	RogerHallamCS21	ProfStrachan
WeLnever	TIME	MrMatthewTodd	LouiseB_NY
bobhillbrain	Forbes	CarolineLucas	ShellenbergerMD

During			
tweet count	follower count	retweet count	I score
Independent	CNN	BjornLomborg	AssaadRazzouk
great_thunberg	nytimes	AssaadRazzouk	MikeHudema
MrMatthewTodd	BBCWorld	toadmeister	ProfBillMcGuire
PhydellaLL	NatGeo	JunkScience	CharlieJGardner
indy_climate	TheEconomist	nytimes	BjornLomborg
ReutersScience	Reuters	MikeHudema	hausfath
Outside1791	FoxNews	nationalpost	JimBair62221006
climate	WSJ	Reuters	EliotJacobson
UncleChopperRIP	washingtonpost	MrMatthewTodd	fossiltreaty
TopClimateNews	TIME	washingtonpost	ClimateComms

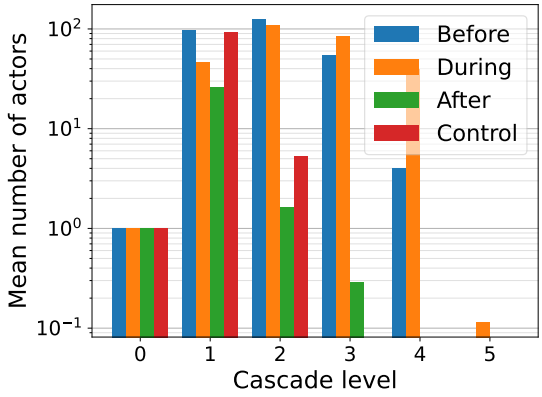
After			
tweet count	follower count	retweet count	I score
great_thunberg	CNN	JunkScience	CarolineLucas
PhydellaLL	nytimes	nytimes	SierraClub
UncleChopperRIP	NatGeo	BjornLomborg	dwallacewells
TopClimateNews	TheEconomist	CarolineLucas	BjornLomborg
indy_climate	Reuters	toadmeister	ProfBillMcGuire
Outside1791	WSJ	ECOWARRIORSS	DocsEnvAus
BlasphemousBan1	washingtonpost	business	LobbyForClimate
Independent	TIME	NewYorker	ProfTerryHughes
matt_syk34	Forbes	UNClimateSummit	BrentToderian
br00t4c	AP	BrentToderian	surfinhambone

Table 4.7: Top 10 most important users using traditional influence measures, for each of the strata. This table was obtained using the COP27 dataset.

- We find the 500 most influential users according to 3 traditional measures of influence: follower count, retweet count and I score. These 500 users will be used as individual actors.
- We then aggregate the remaining users in bins of size N according to their influence value for each of the 3 influence measures. For example, if we have $N = 10$, it means that users having ranks 501 to 510 according to the influence measure will be aggregated as one actor, then users having ranks 511 to 520 into a second actor, etc... Note that at some point the influence measure may give a large portion of users a score of 0 (i.e. nobody retweeted them or they do not have any followers). At this



(a) (b)



(c)

Figure 4.10: a) count of each type of edges, b) count of each edge type normalized by the number of reachable edges, and c) mean number of actors at each levels of the influence cascades derived from the graph. Data obtained using JDD on the COP27 dataset.

point, the rank between users is ill defined and the aggregation is made uniformly at random between users.

- We compute the influence graphs with the actors being a mixture of individual users and aggregates of users, and study the number of edges in the graph involving aggregates (i.e. edges coming in or out of aggregate actors).

Figure 4.13 presents the result of this study when varying the size of the aggregates (N), for each of the 3 traditional influence measures mentioned, as well as for both joint distance distribution (left column) and transfer entropy (right column). Figure 4.14 shows

Before	
outdegree	betweenness
BurningClock	BurningClock
Reuters	HBreen2
HBreen2	guardian
axios	tdzarnick
WashTimes	EarthAccounting
ConversationEDU	EveningStandard
tdzarnick	IEyeOfTheStorm
guardian	PoliticsKulture
EarthAccounting	indy_climate
EveningStandard	LordGittins

During	
outdegree	betweenness
auto_news_feed	Vastuullisuus
Vastuullisuus	auto_news_feed
ECIU_UK	jftaveira1993
Telegraph	PoliticsKulture
Daily_Express	dicklibertyshow
PoliticsKulture	TheEconomist
TheEconomist	Telegraph
TheCanaryUK	BurningClock
BurningClock	Daily_Express
dicklibertyshow	rpujolvives

After	
outdegree	betweenness
SafetyPinDaily	PhydellaLL
jftaveira1993	ECOWARRIORSS
DrBobBullard	BullardCenter
AP_Climate	FriendsOScience
TIME	-
indy_climate	-
PhydellaLL	-
k__ei	-
BullardCenter	-
CCLSVN	-

Table 4.8: Top 10 most important users using centrality measures based on the influence graphs, for each of the strata. Graph derived using JDD on the COP27 dataset.

Before	
outdegree	betweenness
RobotChange	RobotChange
Surly01	Reuters
ECIU_UK	TopClimateNews
7adair	Surly01
newscientist	GISP_Tweets
margreis9	GlobalUnion3
riv39525750	brenda_spiller
mommom_dayton	ecobearwitness
GISP_Tweets	AlexWitzleben
Reuters	HBreen2

During	
outdegree	betweenness
RobotChange	RobotChange
empyreanprotoc1	TopClimateNews
bernieT36	empyreanprotoc1
GlobalUnion3	GlobalUnion3
TopClimateNews	joincurby
SocMedBoost	FreshElecSolar
JunkScience	Surly01
AltayErgun	Robert76907841
anamafalda1992	danspena
FreshElecSolar	JunkScience

After	
outdegree	betweenness
CAROL11959252	RobotChange
RobotChange	TopClimateNews
empyreanprotoc1	empyreanprotoc1
EIA_News	BizSustainably
TopClimateNews	ClubAdaptation
BizSustainably	foodandwater
maxboykoff	CAROL11959252
brave0nft	danspena
BMDD10	GlobalUnion3
BernThemAll	ecobearwitness

Table 4.9: Top 10 most important users using centrality measures based on the influence graphs, for each of the strata. Graph derived using TE on the COP27 dataset.

the percentage that the aggregate actors represent out of all actors in the dataset, as a function of the size N.

4.5 News sources shared by the most influential

4.5.1 RT and Sputnik as vectors of disinformation

We examine the news sources shared by the users who's discussions of untrustworthy URLs our methods find to have the most influence on the actions of other users to discuss untrustworthy URLs (i.e. we investigate U-U edges in the influence graphs). The first idea is the following: Ramsay and Robertshaw [26] showed that the news sources "RT" (rt.com) and "Sputnik" (sputniknews.com) are mostly responsible for the influence campaign during the Skripal poisoning event. Both of these websites are considered as untrustworthy by NewsGuard. Using the 10 most influential actors on the U-U edges for both JDD and TE (Tables A.1 and A.2 respectively), we investigate how those users compare to the rest of

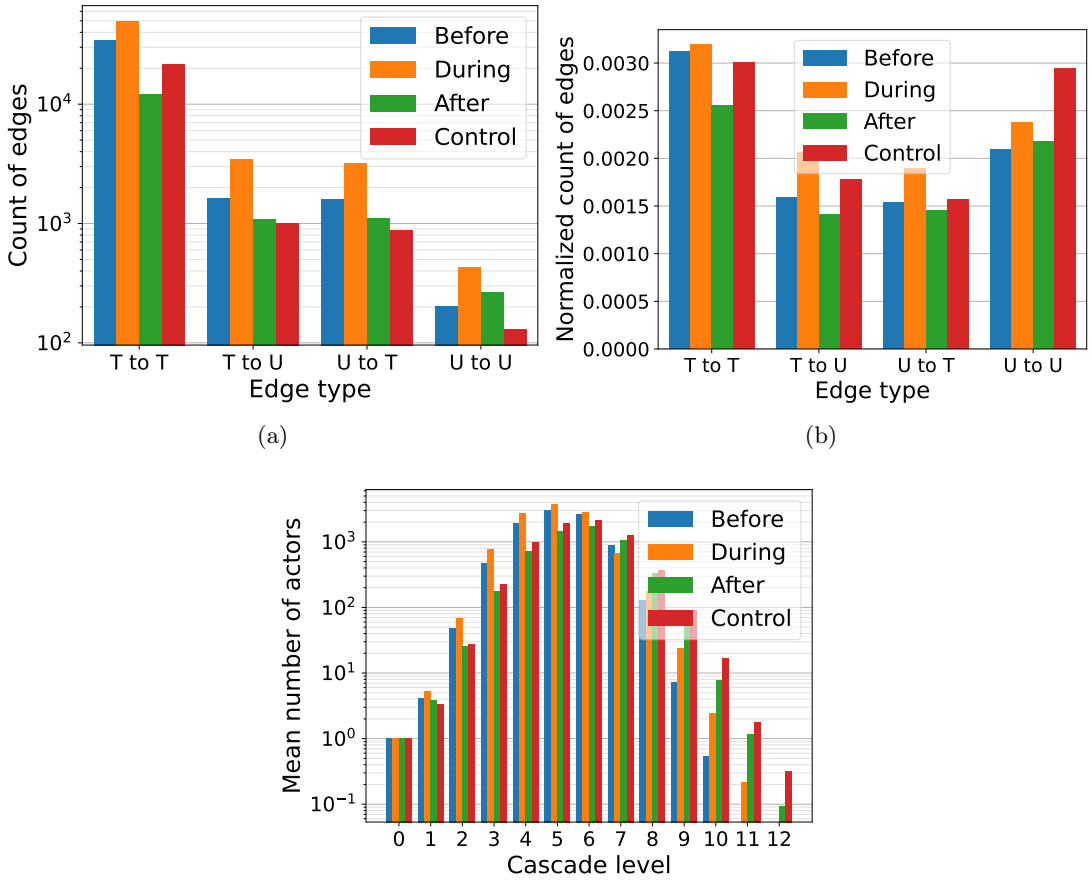


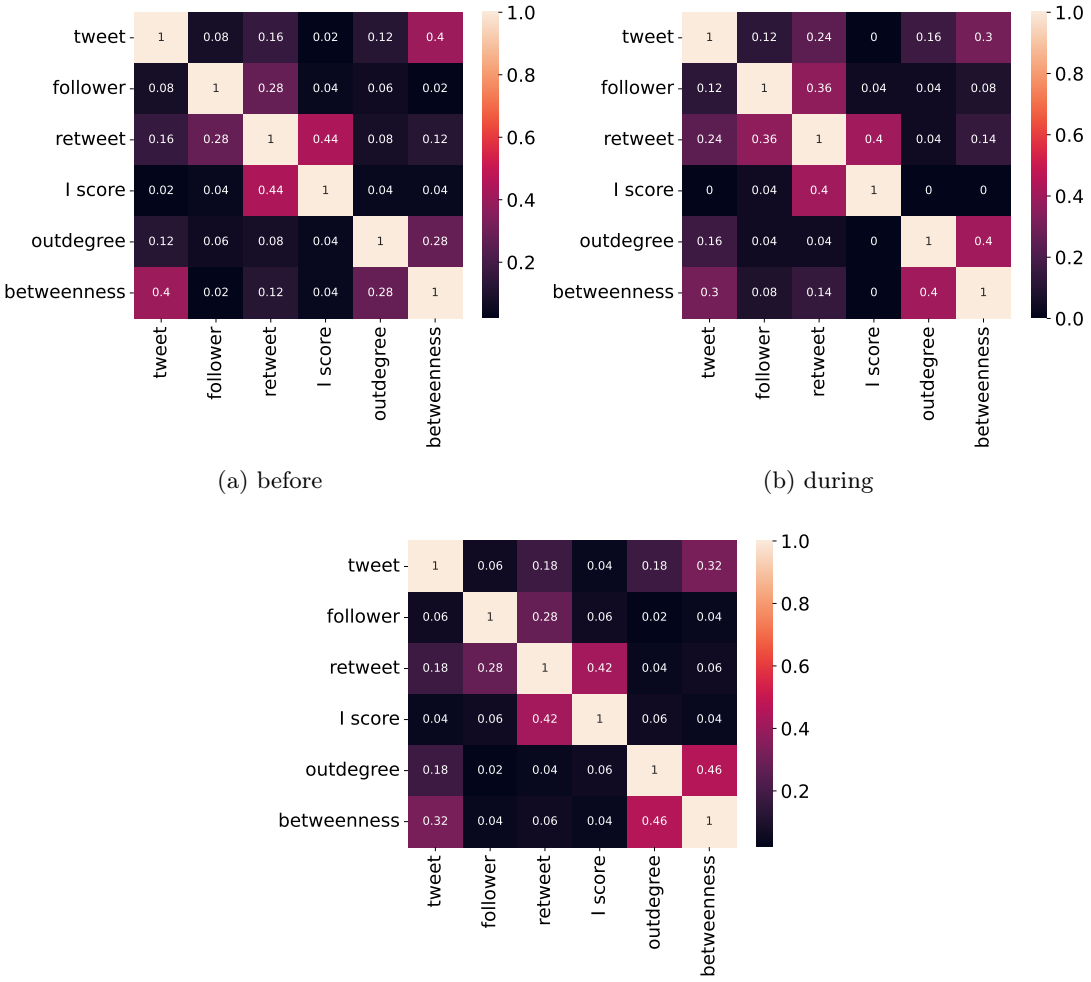
Figure 4.11: a) count of each type of edges, b) count of each edge type normalized by the number of reachable edges, and c) mean number of actors at each levels of the influence cascades derived from the graph. Data obtained using TE on the COP27 dataset.

the users when sharing RT or Sputnik news.

Tables 4.10 and 4.11 respectively compare the mean number of RT and Sputnik URLs shared by users, for all users having shared untrustworthy (U) URLs, and the top 10 most influential users on the U-U edges for both JDD and TE.

4.5.2 COP26 disinformation sources

We study what type of news sources individuals labeled as the most influential (according to outdegree on the U-U edges) shared during COP26. Table A.3 shows that JDD only



(a) before

(b) during

(c) after

Figure 4.12: Correlation matrix between each measure for the COP27 dataset. The centrality measures outdegree and betweenness refer to the TE graph.

finds 5 users from whom all U-U edges in the graph are coming. For this reason, we restrict our analysis to the top 5 most influential users on the U-U edges for both JDD and TE.

For both JDD and TE, we extract the domain of each untrustworthy (U) URL shared by each of the 5 most influential users, and plot them on Figure 4.15. The affiliation of each news source to a state or political party was made according to the wikipedia[1] page

[1] https://www.wikipedia.org/

	Skripal dataset	Top 10 JDD	Top 10 TE
Before	4.8	-	32.9
During	3.9	33.6	15.1
After	6.3	20.7	10.9

Table 4.10: Mean number of tweets containing a "rt.com" URL per user who shared untrustworthy news. The top 10 JDD and TE users are the 10 most influential actors according to outdegree on the U-U edges (see Tables A.1 and A.2).

	Skripal dataset	Top 10 JDD	Top 10 TE
Before	1.1	-	4.0
During	1.9	28.7	19.5
After	1.4	11.5	0

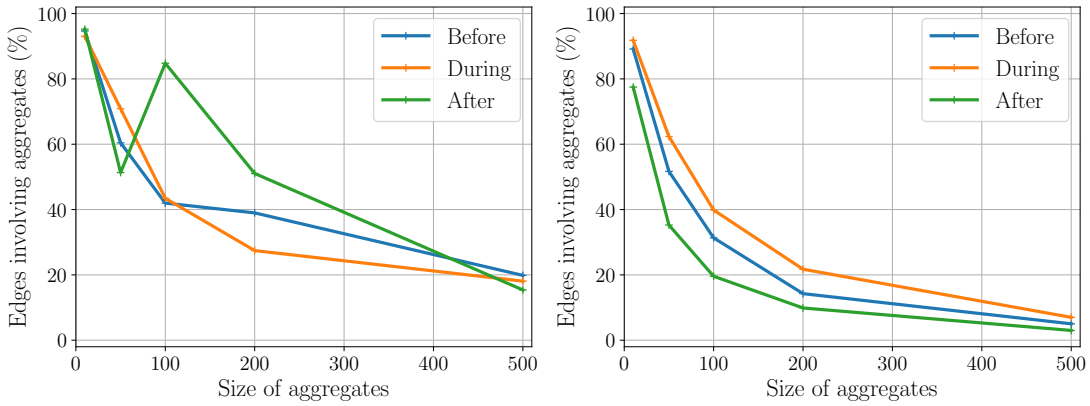
Table 4.11: Mean number of tweets containing a "sputniknews.com" URL per user who shared untrustworthy news. The top 10 JDD and TE users are the 10 most influential actors according to outdegree on the U-U edges (see Tables A.1 and A.2).

dedicated to the news source. If the news outlet does not have a wikipedia page or if the political affiliation is not clearly mentioned, we label it as "other". Note that we make the distinction between American far-right affiliated news sources, and far-right affiliated news sources from other countries.

Table 4.12 shows some of the articles to which the URLs extracted in Figure 4.15 a) point to. The actual link to the web article is included. For the full URLs (not just symbolic links), see the replication of the table in the Appendix (Table A.8).

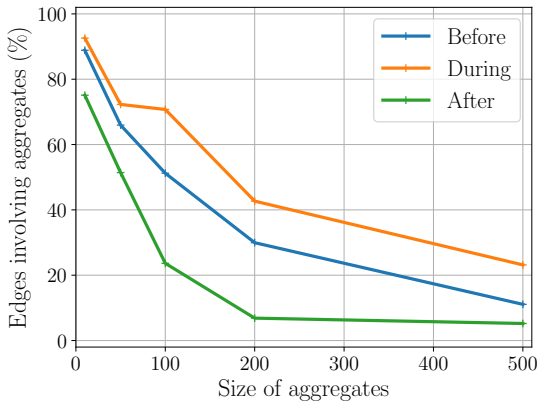
news outlet	article title
globaltimes.cn	Western backbiting over China's coal production completely unjustified (url)
	China keeps its promises on climate, so should the US: Global Times editorial (url)
cgtn.com	China tells U.S. to take actions, not empty words on climate change (url)
	China highlights 'arduous efforts' it has made to fight climate change (url)
breitbart.com	Xi Jinping Scolds World on Climate Change While China Keeps Polluting (url)
	Watch: Sleepy Joe Biden Struggles to Stay Awake During Climate Change Summit (url)
thegatewaypundit.com	Biden's Marxist Treasury Nominee Says the Quiet Part Out Loud on Fossil Fuel Industry: "We Want Them to go Bankrupt if We Want to Tackle Climate Change" (VIDEO) (url)
zerohedge.com	Watch: Al Gore's Latest 'Solution' To Climate Change Is Mass Surveillance (url)

Table 4.12: Example of articles shared by the top 5 most influential users according to JDD (Figure 4.15 a)).

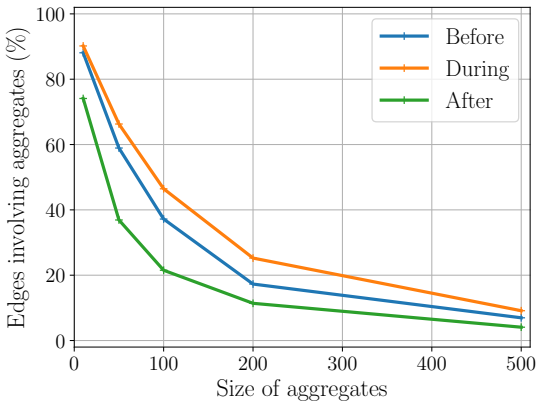


(a) follower count, JDD

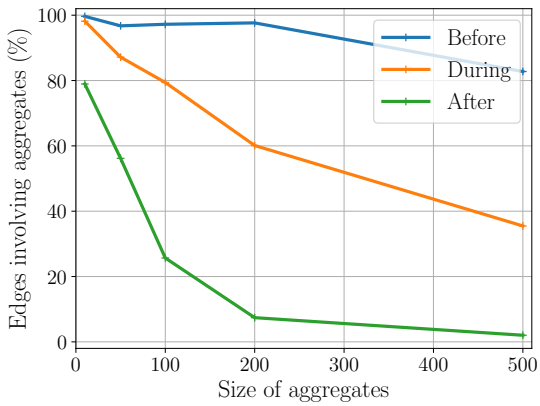
(b) follower count, TE



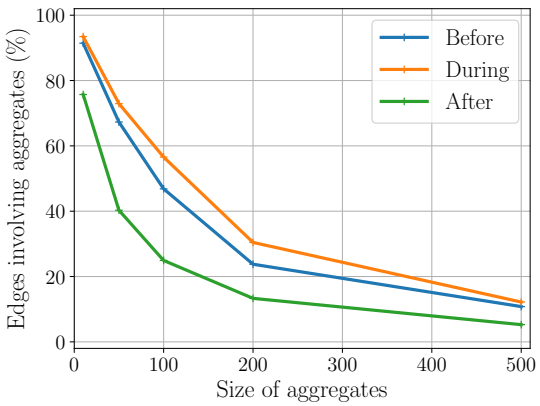
(c) retweet count, JDD



(d) retweet count, TE



(e) I score, JDD



(f) I score, TE

Figure 4.13: Percentage of graph edges involving an actor defined as an aggregate of multiple users (i.e. edge to or from an aggregate). a) and b) show the results when aggregating based on follower count for JDD and TE respectively, c) and d) when aggregating based on retweet count, and e) and f) when aggregating based on I score.

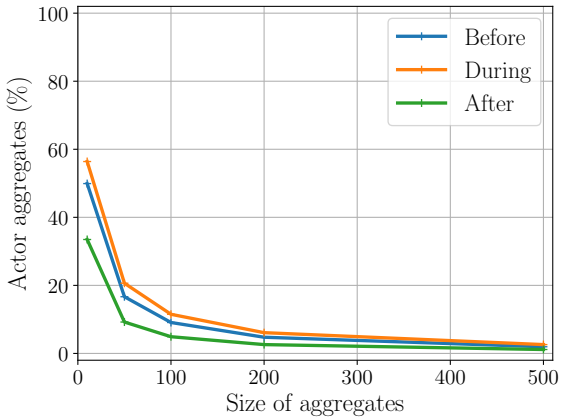
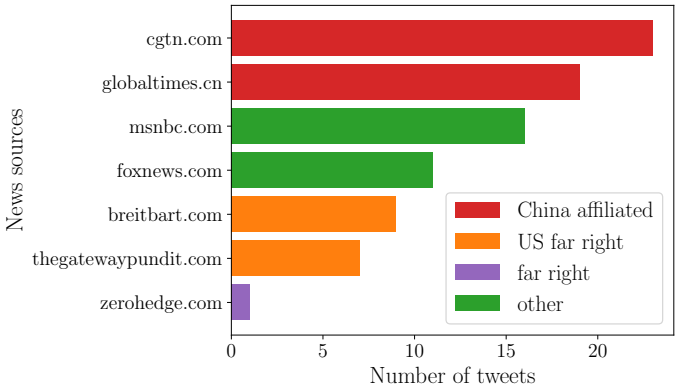
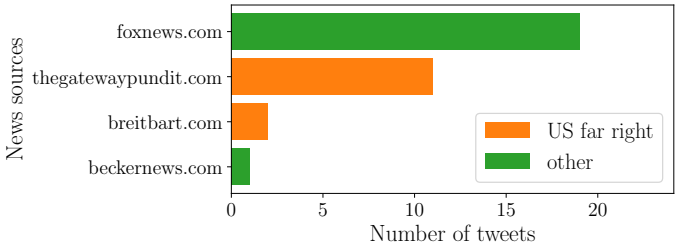


Figure 4.14: Percentage of aggregate actors out of all actors. With lower aggregate size, we need more aggregates to span all individual users, and thus the percentage of aggregates in the total of actors is larger.



(a) JDD



(b) TE

Figure 4.15: News outlets mentioned by the 5 most influential users according to outdegree on edge U-U during COP26, for both JDD and TE graphs.

Chapter 5

Discussion

In this chapter, we discuss the results obtained in Chapter 4. Section 5.1 treats the volume and distribution of data for the different datasets. In Section 5.2, we compare the influence graphs obtained when using JDD or TE. The users detected using different influence measures are analyzed in Section 5.3. Section 5.4 examines the correlations between different influence measures. In Section 5.5, we discuss actor aggregation and its impacts on the influence graphs. Section 5.6 investigates the news sources shared by some users that our methods find as the most influential. Finally, some possibilities of future work and extension of our framework are presented in Section 5.7.

5.1 Volume of data

Figures 4.1, 4.5 and 4.9 show the distributions of trustworthy and untrustworthy URLs shared on Twitter for each events (Skripal poisoning, COP26 and COP27 respectively). They also show the distribution of the number of tweets per user. Remember that they only present the data for users having an activity rate of at least 3 tweets. The distribution of the number of tweets per user is almost a constant for each time window inside a dataset, as well as across datasets, meaning that the distribution of people's activity on Twitter is approximately constant (for people with a minimum activity of 3 tweets), independently of the type of events or the evolution and popularity gain of the platform (the Skripal data goes back to 2018, while COP26 was in 2021 and COP27 in 2022).

However, the distribution and volume of trustworthy and untrustworthy URLs shared by users heavily depends on the kind of event, as well as the time window around the event itself. The Skripal poisoning, a spontaneous incident that resonated worldwide, displays a very large proportion of untrustworthy URLs: between 28.61% in the first period to 46.82% when the event was attributed to Russia. In comparison, for COP26 and COP27, the proportion of untrustworthy URLs is never higher than 7.20%. Interestingly, when the

proportion of untrustworthy URLs is the highest for Skripal (during), it is not the amount of untrustworthy URLs that increased, but the volume of trustworthy URLs that decreased (about 2 times lower than for the other 2 time periods). This means that the flow of trustworthy URLs temporarily stopped after the attribution of the poisoning to Russia while the flow of untrustworthy URLs kept roughly constant.

5.2 Comparison between JDD and TE based influence graphs

One of the main goals of this report is to characterize and compare JDD and TE as methods to capture influence between actors' actions. Figures 4.2, 4.6 and 4.10 show the number of edges and the reach of the influence cascades when using JDD for the Skripal, COP26 and COP27 datasets respectively. Figures 4.3, 4.7 and 4.11 present the same quantities when we derive the influence graphs with transfer entropy (TE). From these figures, it is clear that the influence graphs obtained with TE are more dense than the graphs obtained with JDD. Moreover, the normalized proportion of edges of each type is more evenly distributed across the time periods for TE, whereas it is harder to predict for JDD. The influence cascades are consistently longer (they span more levels) and reach more actors at a given level for TE compared to JDD. We hypothesize that this is due to the underlying structure of the graphs which is more dense with TE, as previously mentioned.

Since COP26 and COP27 are scheduled events (as opposed to spontaneous events such as Skripal), we have access to a control set (see Section 3.1.3). Figures 4.7 and 4.11 show that after proper normalization, the count of edges for COP26 and COP27 cannot be correctly differentiated from the count of edges obtained for the control set when the graphs are derived with TE. The same is true for the influence cascades: there are no clear differences between results obtained for the control and the actual COP26 and COP27 data. It is therefore not possible to exclude the possibility that the links between actors obtained with TE are the product of pure chance, i.e. noise incurred by the transfer entropy computation. For JDD, we can observe a clear difference with the control set. Indeed, for COP26 the method captures large spikes of U-U and U-T influence during the event, which are not present at all for the control. The amount of edges for T-T and T-U influence is also much larger than for the control. For COP27 however, the number of edges obtained is quite similar to the control set.

5.3 Active spreaders of Russian narratives

According to Ramsay and Robertshaw [26], we know that RT and Sputnik are the main news outlets responsible for the Russian disinformation campaign. Tables 4.1, 4.2, 4.3, A.1 and A.2 all show the most influential users for the Skripal dataset, according to different

influence measures. We are able to extract some interesting insights from these tables. First, the account "RT_com" is omnipresent for traditional measures: it appears in the top 10 according the the tweet count for all strata, and is either first or second for the retweet count for all periods as well. It is however only detected by the JDD-based influence graphs, and only for the stratum "during", but for which case it is classified as the first most influential user according to both betweenness and outdegree on the edge type U-U. As its name indicates, the account is the official Twitter account of the news source RT. We note that the influence measures derived from the TE-based graphs detect the UK-based antenna of RT, the account "RTUKnews".

However, both JDD-derived and TE-derived measures fail to detect the account "CraigMurrayOrg", number 1 most influential user for all strata according to I score, and also highly ranked by retweet count. The account belongs to Scottish author and ex-diplomat Craig Murray. While we did not find evidence of any affiliation between Murray and Russia, the author kept posting articles on his blog denying (or at least strongly questioning) Russia culpability of the attempted murder, and some of his articles were directly quoted and published by RT, with headlines such as "Craig Murray: Opposition figure Navalny may possibly have been targeted by Russian state, but Western narrative doesn't add up"[1]. The I score measure consistently finds the accounts belonging to Neil Clark ("NeilClark66") and Charles Shoebridge ("ShoebrideC") for all strata. Both are UK citizens and regular pundits on RT. They are however not detected by our methods, except Shoebridge who is found by the JDD-based betweenness measure on the edge type T-U, but only for the stratum "after".

JDD-based influence measures detected the account "newsroll", a partially automated account (now banned from Twitter) who was actively spreading disinformation (according to an article posted by the Digital Forensic Research Lab on Medium in late March 2018[2]). None of the other measures detected this account. Finally, the account "Ian56789" was found by both the retweet count measures, as well as TE-based measures. This account, originally labeled as a Russian bot designed to spread disinformation[3], was later claimed by a UK citizen who publicly went on Sky News to prove his identity[4].

[1] https://www.rt.com/russia/500013-navalny-targeted-state-western-narrative/

[2] https://medium.com/dfrlab/trolltracker-stale-narratives-in-response-to-expelled-diplomats-3afeed88ee1

[3] https://www.theguardian.com/world/2018/apr/19/russia-fake-news-salisbury-poisoning-twitter-bots-uk

[4] https://www.polygraph.info/a/twitter-troll-is-actually-uk-citizen/6741811.html

5.4 Different measures for different influence definitions

In Section 3.6, we defined different ways to rank the users according to different influence measures. In Chapter 4, we presented different tables showing the users and their rank according to these measures. Figures 4.4, 4.8 and 4.12 show the correlation defined in Equation 4.1 between these influence measures for the Skripal, COP26 and COP27 datasets respectively. From these correlation matrices, one may observe the following:

- Overall the correlation between different influence measures is low.
- The largest consistent correlation between datasets and strata within the datasets is between retweet and I score. This is expected due to how I score is computed (based on the retweet graph).
- We note some other consistent correlations: between outdegree and betweenness, retweet and follower, and betweenness and tweet.

While the link between retweet count and I score is expected, the other correlations reveal interesting properties of the influence graphs. First, their structure is such that users directly influencing a lot of other users are also acting as bridges between other actors in the graph (correlation between outdegree and betweenness). The link between betweenness and tweet count suggests that the more content a user posts, the more likely this user is going to be a relay of influence between people. Finally, the correlation between retweet count and follower count shows that with a larger audience, it is easier to spread a message that at least some in the audience will find interesting.
However, the relatively low values for these correlations stress that each measure is in fact capturing different types of information, and influence definitions.

5.5 Impact of user aggregation

Figure 4.13 demonstrates that TE is far more robust to actor aggregation than JDD. Indeed, the number of edges involving aggregates in the graph is almost a perfect constant as a function of the size of the aggregate, independently of how the aggregates are chosen. On the other hand, the proportion of edges in or out of aggregates strongly depends on how the aggregates are created for JDD. The behavior is not even the same across the strata, when aggregating in the same manner for JDD. For example, when the I score is used to create the bins of users, the vast majority of the influence graph is composed of edges between aggregates in the stratum "before COP26", independently of the size of the bins. But in the stratum "after COP26", as the bin size increases, the graph is quickly made of very few edges between aggregates, the majority of the edges being between individual users.

Figure 4.14 maps the size of the aggregates (or bins of users) to the proportion of actors defined as aggregates out of all actors. For the smallest size of bins tested (i.e. 10), the aggregates represent between about 35% and 60% of all actors. But the edges involving them represent between 80% and 100% of the influence graph in all cases (for both JDD and TE and all aggregation strategies). And this trend is observed for all bin sizes: the proportion of edges from the aggregates is always larger than the proportion the aggregates themselves represent in the actors. This is even more marked for JDD than TE. This suggests that both JDD and TE are sensitive to the amount of data available for the actors: defining one actor as an aggregation of multiple users results in time series of action frequencies less sparse for this actor, and in turn a bigger proportion of edges in the graph.

5.6 Echo chambers of untrustworthy news sources

As mentioned in Section 3.6.2, edges of the type U-U in the influence graphs are related to echo chambers of untrustworthy news sharing. Tables 4.10 and 4.11 show that for the Skripal dataset, the outdegree measure based on the influence graphs obtained with JDD finds users tweeting on average a lot more of articles from "rt.com" and "sputniknews.com" than the same measure based on influence graphs with TE. Since we know from the study of Ramsay and Robertshaw [26] that RT and Sputnik are the principal vectors of the Russian disinformation campaign, we can say that the JDD-based influence graphs found users more implicated in propagating the Russian narratives than the TE-based graphs.

In Figure 4.15, it is clear that the users found by the JDD-based graph on COP26 are sharing more news from sources that may be qualified of questionable. In order to understand what kind of articles were shared from these news sites, we manually examined the articles to which the URLs point. Table 4.12 presents some of the articles we found that immediately reflected the presence of misinformation/disinformation. China-affiliated news outlets seem to share anti-US and anti-Western climate narratives while greatly exaggerating their own climate actions. From the far right American side, we can observe the counter anti-Chinese narrative, as well as articles demeaning Biden's climate awareness and administration.
Such findings provide evidence that JDD-based influence graphs can identify users who are more involved in influence campaigns on Twitter than TE-based graphs.

5.7 Future work

The choices of stratification, actions and actors we made are handy to compare how users are influencing each other by sharing URLs. However, the influence graph method described in this work is agnostic to these choices. For example, one could study influence

based on different actions, or actions based on a different categorization of the URLs (Mainstream/Fringe instead of Trustworthy/Untrustworthy, or even combine both into 4 possible actions). It is also interesting to perform community detection on the Twitter graphs to find groups of users with the same beliefs/ideas and study influence between such communities.

It is also possible to integrate and develop the influence graph method presented in this work into tools to further investigate and analyze disinformation campaigns. Moreover, one could validate and expand results from the analyses we made on other cases of known disinformation campaigns.

Finally, actions could benefit from a topic or context analysis. Indeed, if someone posts "Look at the fake news I found ! https://rt.com", the action associated to the tweet will be U, i.e. sharing of untrustworthy news. But in this case, the author actually exposed the disinformation attempt to other users, thus helping mitigate its effect on the network.

Chapter 6

Conclusion

In this work, we explored the use of two coupling inference methods to derive influence graphs, and detect disinformation campaigns on social media. NewsGuard labeling of news sources was used to categorize the URLs that users share on Twitter into trustworthy or untrustworthy content. We focused on climate change related data around two major climate events, the COP26 and COP27. In order to compare and benchmark our results, we also studied the Skripal poisoning, for which we know that Russia launched a large scale disinformation campaign on Twitter. We compared different influence measures, some readily available from the Twitter data such as follower count or retweet count, and other derived from our influence graphs formulation. We find that these measures have relatively low correlation between themselves, showcasing that they do not capture the same types of influence. Finally, we observe that using joint distance distribution as coupling inference method results in sparser influence graphs, and is more sensitive to the choice and aggregation of actors, but finds individuals who seem more active in spreading disinformation than when using transfer entropy.

Bibliography

[1] José M Amigó and Yoshito Hirata. "Detecting directional couplings from multivariate flows by the joint distance distribution". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.7 (2018), p. 075302.

[2] Lionel Barnett, Adam B Barrett, and Anil K Seth. "Granger causality and transfer entropy are equivalent for Gaussian variables". In: *Physical review letters* 103.23 (2009), p. 238701.

[3] Robert M Bond et al. "A 61-million-person experiment in social influence and political mobilization". In: *Nature* 489.7415 (2012), pp. 295–298.

[4] Meeyoung Cha et al. "Measuring User Influence in Twitter: The Million Follower Fallacy". In: *Proceedings of the International AAAI Conference on Web and Social Media* 4.1 (May 2010), pp. 10–17. DOI: 10.1609/icwsm.v4i1.14033. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14033.

[5] Federico Cinus et al. "The effect of people recommenders on echo chambers and polarization". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 90–101.

[6] John Cook, Peter Ellerton, and David Kinkead. "Deconstructing climate misinformation to identify reasoning errors". In: *Environmental Research Letters* 13.2 (2018), p. 024018.

[7] Robert Faris et al. "Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election". In: *Berkman Klein Center Research Publication* 6 (2017).

[8] Linton C Freeman. "A set of measures of centrality based on betweenness". In: *Sociometry* (1977), pp. 35–41.

[9] Andrea Galeotti and Sanjeev Goyal. "Influencing the influencers: a theory of strategic diffusion". In: *The RAND Journal of Economics* 40.3 (2009), pp. 509–532.

[10] Clive WJ Granger. "Investigating causal relations by econometric models and cross-spectral methods". In: *Econometrica: journal of the Econometric Society* (1969), pp. 424–438.

[11] Behnam Hajian and Tony White. "Modelling influence in a social network: Metrics and evaluation". In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE. 2011, pp. 497–500.

[12] Saike He et al. "Identifying peer influence in online social networks using transfer entropy". In: *Pacific-Asia workshop on intelligence and security informatics*. Springer. 2013, pp. 47–61.

[13] Elihu Katz and Paul F Lazarsfeld. *Personal influence: The part played by people in the flow of mass communications*. Routledge, 2017.

[14] Jennie King, Lukasz Janulewicz, and Francesca Arcostanzo. "Deny, deceive, delay: Documenting and responding to climate disinformation at COP26 and beyond". In: (2022).

[15] Maksim Kitsak et al. "Identification of influential spreaders in complex networks". In: *Nature physics* 6.11 (2010), pp. 888–893.

[16] Jon M Kleinberg. "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[17] Gemma Lancaster et al. "Surrogate data for hypothesis testing of physical systems". In: *Physics Reports* 748 (2018), pp. 1–60.

[18] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. "The dynamics of viral marketing". In: *ACM Transactions on the Web (TWEB)* 1.1 (2007), 5–es.

[19] Linyuan Lü et al. "Recommender systems". In: *Physics reports* 519.1 (2012), pp. 1–49.

[20] Katerina Eva Matsa et al. "Western Europeans under 30 view news media less positively, rely more on digital platforms than older adults". In: (2018).

[21] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges". In: *Ai & Society* 35.4 (2020), pp. 957–967.

[22] Patricia Moravec, Randall Minas, and Alan R Dennis. "Fake news on social media: People believe what they want to believe when it makes no sense at all". In: *Kelley School of Business research paper* 18-87 (2018).

[23] *NewsGuard Ratings*. https://www.newsguardtech.com/solutions/newsguard/. Accessed: 2022-09-01.

[24] Lawrence Page et al. *The PageRank citation ranking: Bringing order to the web*. Tech. rep. Stanford InfoLab, 1999.

[25] Stefano Panzeri et al. "Correcting for the sampling bias problem in spike train information measures". In: *Journal of neurophysiology* 98.3 (2007), pp. 1064–1072.

[26] Gordon Ramsay and Sam Robertshaw. "Weaponising news: RT, Sputnik and targeted disinformation". In: (2019).

[27] Fabián Riquelme and Pablo González-Cantergiani. "Measuring user influence on Twitter: A survey". In: *Information processing & management* 52.5 (2016), pp. 949–975.

[28] Everett M Rogers, Arvind Singhal, and Margaret M Quinlan. "Diffusion of innovations". In: *An integrated approach to communication theory and research*. Routledge, 2014, pp. 432–448.

[29] Daniel M Romero et al. "Influence and passivity in social media". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2011, pp. 18–33.

[30] Thomas Schreiber. "Measuring information transfer". In: *Physical review letters* 85.2 (2000), p. 461.

[31] Chathurani Senevirathna et al. "Influence Cascades: Entropy-Based Characterization of Behavioral Influence Patterns in Social Media". In: *Entropy* 23.2 (2021), p. 160.

[32] Karishma Sharma, Emilio Ferrara, and Yan Liu. "Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 US Presidential Election". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 908–919.

[33] Elisa Shearer and Jeffrey Gottfried. "News use across social media platforms 2017". In: (2017).

[34] Catharine Starbird, Ahmer Arif, and Tom Wilson. *Understanding the structure and dynamics of disinformation in the online information ecosystem*. Tech. rep. University of Washington Seattle United States, 2018.

[35] George Sugihara et al. "Detecting causality in complex ecosystems". In: *science* 338.6106 (2012), pp. 496–500.

[36] Floris Takens. "Detecting strange attractors in turbulence". In: *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*. Springer. 2006, pp. 366–381.

[37] Marco Thiel et al. "Twin surrogates to test for complex synchronisation". In: *EPL (Europhysics Letters)* 75.4 (2006), p. 535.

[38] Antonela Tommasel and Filippo Menczer. "Do Recommender Systems Make Social Media More Susceptible to Misinformation Spreaders?" In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 550–555.

[39] Greg Ver Steeg and Aram Galstyan. "Information transfer in social media". In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 509–518.

[40] Greg Ver Steeg and Aram Galstyan. "Information-theoretic measures of influence based on content dynamics". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, pp. 3–12.

[41] Demetris Vrontis et al. “Social media influencer marketing: A systematic review, integrative framework and future research agenda”. In: *International Journal of Consumer Studies* 45.4 (2021), pp. 617–644.

[42] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. “The spread of medical fake news in social media–the pilot quantitative study”. In: *Health policy and technology* 7.2 (2018), pp. 115–118.

[43] Duncan J Watts and Peter Sheridan Dodds. “Influentials, networks, and public opinion formation”. In: *Journal of consumer research* 34.4 (2007), pp. 441–458.

[44] Jianshu Weng et al. “Twitterrank: finding topic-sensitive influential twitterers”. In: *Proceedings of the third ACM international conference on Web search and data mining*. 2010, pp. 261–270.

Appendix A

Appendix

A.1 Visual representation of the actors with the most followers

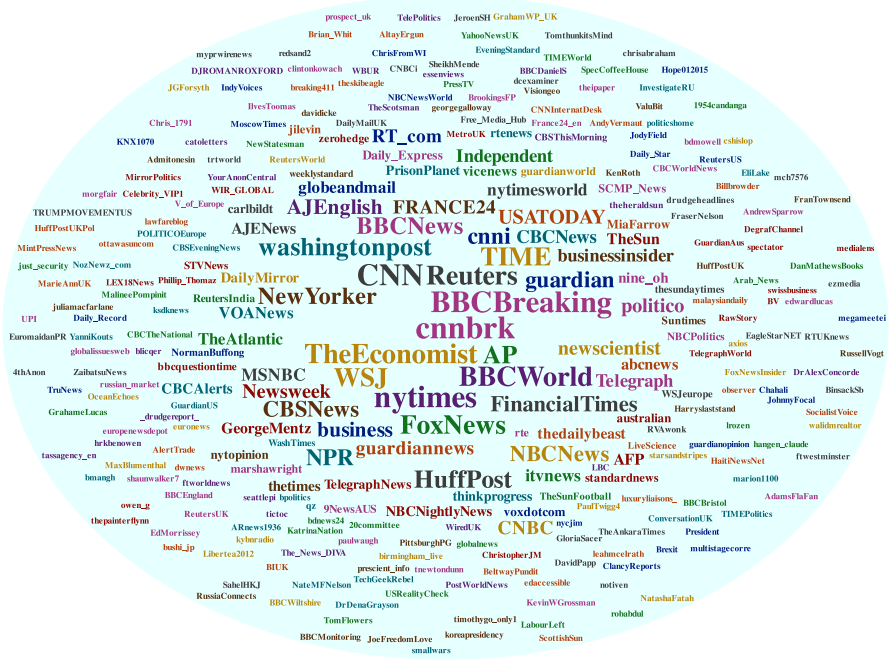


Figure A.1: Wordcloud representing the 300 actors with most followers. The size of the username is proportional to the number of followers. This is for the Skripal dataset, and spans the entire data (no stratification applied).

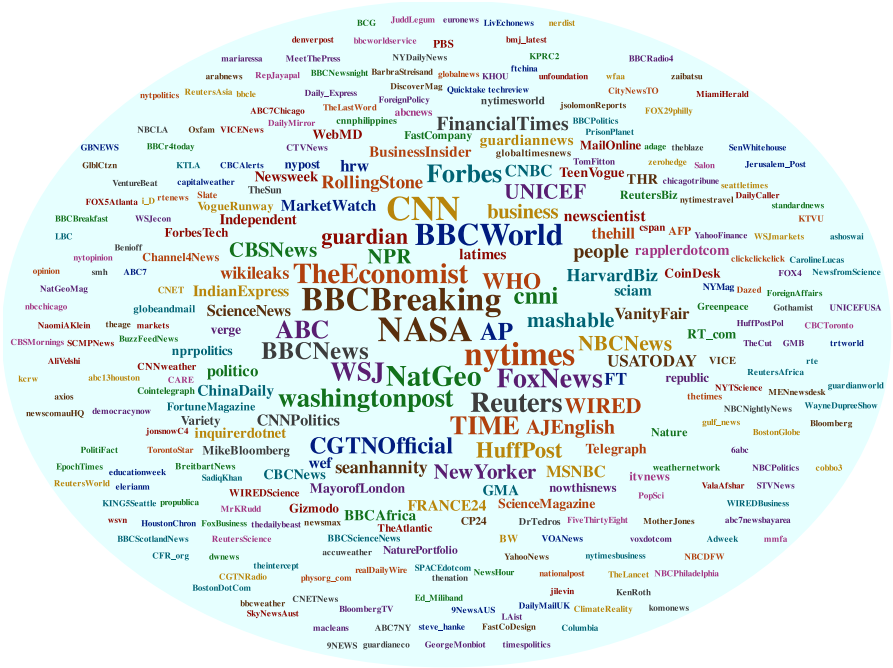


Figure A.2: Wordcloud representing the 300 actors with most followers. The size of the username is proportional to the number of followers. This is for the COP26 dataset, and spans the entire data (no stratification applied).

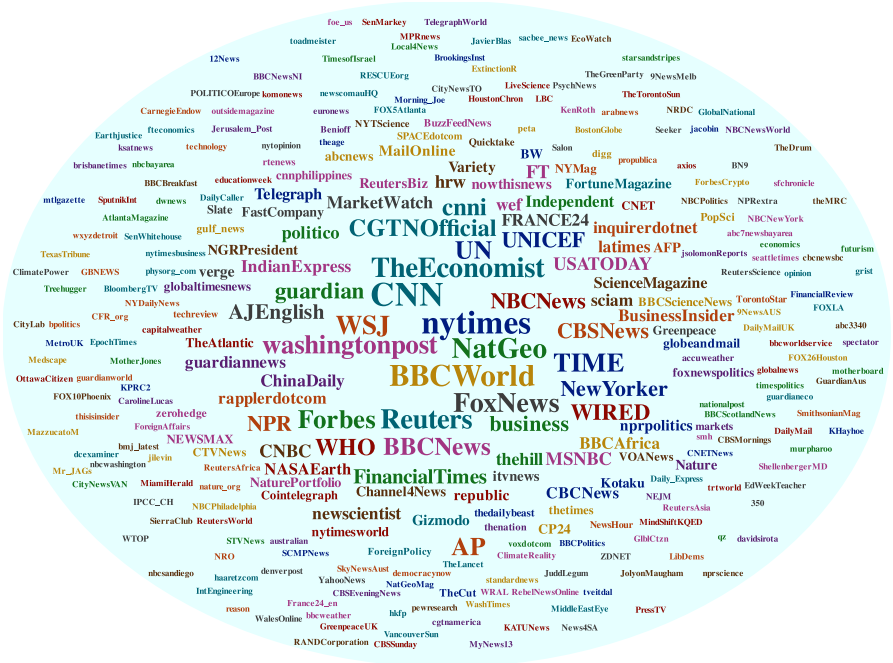


Figure A.3: Wordcloud representing the 300 actors with most followers. The size of the username is proportional to the number of followers. This is for the COP27 dataset, and spans the entire data (no stratification applied).

A.2 Most influential actors for each edge type

Before							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
nytimesworld	thebrkg	farhaadaarif	-	conju_re	conju_re	mayasdolly	mayasdolly
Daily_Star	Harley_Woody	-	-	Telegraph	Daily_Star	farhaadaarif	farhaadaarif
Telegraph	-	-	-	nytimesworld	Juancarlos_Mike	lisa_alba	lisa_alba
Redpolitics	-	-	-	Daily_Star	Telegraph	024AB	-
Londied39	-	-	-	Juancarlos_Mike	nytimesworld	mlnangalama	-
Harley_Woody	-	-	-	ReciteSocial	IndyVoices	-	-
warringworld	-	-	-	Redpolitics	thebrkg	-	-
Brought_to_You	-	-	-	Harley_Woody	grauniad_news	-	-
HoWanKwok	-	-	-	kabarberita	qkode	-	-
_NoMoreExcuses	-	-	-	qkode	ReciteSocial	-	-

During							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
-	-	-	RT_com	lisa_alba	chootchyface	RT_com	RT_com
-	-	-	-	chootchyface	lisa_alba	newsroll	newsroll
-	-	-	-	ConversationUK	ALLREDToDoRoJo	ferozwala	JJorbyn
-	-	-	-	ALLREDToDoRoJo	ConversationUK	RLSRUSSIANNEWS	ferozwala
-	-	-	-	IndpressUK	lesbonner	JJorbyn	paris_2015
-	-	-	-	_dpaj	-	paris_2015	tonybrooklyn5
-	-	-	-	lesbonner	-	QueensIceZ	QueensIceZ
-	-	-	-	-	-	joaonunes67	RLSRUSSIANNEWS
-	-	-	-	-	-	restless94110	dwilliam9940
-	-	-	-	-	-	tonybrooklyn5	lisa_alba

After							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
Angelus1701	ShoebridgeC	-	HillestadNils	Angelus1701	Angelus1701	JudeJack	JudeJack
OnlBreakingNews	-	-	flyer4life	americandailys	americandailys	starandsixpence	HillestadNils
gabriellesct	-	-	jarfizo1	puffin1952	NewsDingo	HillestadNils	SQUADDICTS
NewsBlogged	-	-	lacey9020	NewsBlogged	_ThePage	SQUADDICTS	starandsixpence
puffin1952	-	-	-	_ThePage	gabriellesct	newsbloktwit	newsbloktwit
wittich	-	-	-	amandasome	puffin1952	lordjohnalldis1	flyer4life
marvellous997	-	-	-	ExcitingAds	NewsBlogged	flyer4life	jarfizo1
bassmadman	-	-	-	wittich	ExcitingAds	mariowil7	DJSiri
newsbott3r	-	-	-	NewsDingo	worldnewsbox	myusufali1	Mr_Nick_Nasty
standardnews	-	-	-	worldnewsbox	amandasome	theforeverman	TacticalFM

Table A.1: Top 10 most important actors for each action relationship, for each stratum. This data was obtained through the JDD influence graph for the Skripal dataset.

Before							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
NewStatesman	aspals	tenaciousV56	DJSiri	NewStatesman	boone_jo	akahnnyc	peterpobjecky
NBCNightlyNews	MATErlandsson	PatrickWKavanag	shabbirh	NBCNightlyNews	valko665	RTUKnews	DontDenyThe
zhouhuasheng06	owhy3	MATErlandsson	peterpobjecky	imemilydsouza	Redsfan1977	AndrewRousak	JJorbyn
MaynTex	boone_jo	BetigulCeylan	notiven	zhouhuasheng06	Fabledsoul	BetigulCeylan	wherepond
imemilydsouza	FreeStateYank	therightarticle	DontDenyThe	MaynTex	Paparaw	peterpobjecky	shabbirh
wtfiscrackin	NhanC18	Sander_1954	ValuBit	India24News24	joejtd	Bill_Owen	BeeAHoney_
cgnetwork	wherepond	MargaretDunne13	mayasdolly	luxuryliaisons_	ScholarshipSLE	normalnorms	JuliaPolan
Kostian_V	vanjimbo	ron_ronka	uliTwitts	carslatestnews	MATErlandsson	JJorbyn	PakamamaniRenew
Telegraph	joejtd	BlameItonBHO	pablothehat	bbltweet	itvnews	GauntJohnny	TheUrbanNewz
twaddleninja	gumby4christ	KELLYCLELLAND1	PeterRosianul	hap_santos	Star24News	theaceofspuds	londonfredd

During							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
_ThePage	lei_joh	oldbid45	krstae	NewsAboutLife	pathrs	sengeezer	TheRealYoG
RotenbergBros	ShareCanadaNews	SteffiThompson	RTUKnews	bdnews24	SpyTalker	psic88	ferozwala
NewsAboutLife	mlnangalama	chinedu70204650	Cancersucks6486	Russianation	thebestbond	RTUKnews	BuggerLePanda
andro1711	JohnDelacour	Wilkmaster	OldRightie	notiven	AnthonyLehal	vivaden3	ProfessorsBlogg
Tufairi	RexZark1	davidh7426	SPVereycken	Orgetorix	IndpressUK	skinnergj	Arfatweet
DirectTrip	GeraldEvans95	AlanMcpartlands	TheUrbanNewz	Tufairi	BrendanCarton	jonasalexis2	NecktopP
sabahmajujaya	ShakPro	Russianation	CarlAntoine	BrendanCarton	amandasome	OldRightie	OldRightie
movarsi	1954candanga	bruno_paul	fluoresenz	RLSRUSSIANNEWS	franlawtheruk	Rocketnews1	RTUKnews
KremlinTrolls	AnthonyatNo1	BardenGridge	ZunguZunguZeng	NewsYunkY	GuardianUS	Mmargarites1	ali9l9
notiven	Orgetorix	terencehooson	newsbloktwit	WashTimes	Ray_Phenicie	lei_joh	zerohedge

After							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
thebrkg	openomroep	JiriParkes	STATILIVS	Hadrienr974	JuliaCharnley	rlangjournalist	Pline999
grauniad_news	BlueSeas111	NecktopP	mogabee3	grauniad_news	vrai777	JiriParkes	Revoche
MSNBC	_disbasin	Gyre07	infidelchloe	Joanvanderlinge	grauniad	Nildam85	SnakeTera
myamigocouk	jlz0z	jlz0z	Mr_Nick_Nasty	myamigocouk	conju_re	WhirlwindWisdom	infidelchloe
POLITICOEurope	julieaskew50	AdamCli	brucerisk	andy_s_64	Amorovz	lucma66	charlievictor16
newsbott3r	Lewisno1fan	StallaSimonin	evertonfc2	hangen_claude	Howdyrich	Ian56789	Char_lotte777
TheAllRadar	DowninJamaica	STATILIVS	mojos55	guardian	MicroSuperFan	new16media	LordGamblore
jondknight	WilliamDuguid1	fredwalton216	newsbloktwit	eddwilson	PamelaFalk	HowardSanjuani2	iccjock06
grauniad	Ian56789	openomroep	Ndakoma	LeeFergusson	Stephen_Gash	BryantDiananp	StephaniePetri1
SourceMerlin	Harley_Woody	ClubBayern	RTUKnews	GuardianUS	guardian	CatherinJasmin	TheUrbanNewz

Table A.2: Top 10 most important actors for each action relationship, for each stratum. This data was obtained through the TE influence graph for the Skripal dataset.

Before							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
jilevin	jlitwinetz	-	-	MichiganRadio	MichiganRadio	aawsat_eng	aawsat_eng
MichiganRadio	TurboKitty	-	-	jlitwinetz	jlitwinetz	FriendsOScience	-
jlitwinetz	Bentler	-	-	katydaigle	katydaigle	ArabNewsBiz	-
NelsonGich	great_thunberg	-	-	jilevin	LatinoLdnOnt	-	-
TurboKitty	3beesbuzz	-	-	YV5SEL	YV5SEL	-	-
katydaigle	samsondenver	-	-	BrendanCarton	Daily_Express	-	-
DebsF319	-	-	-	DebsF319	hscampoy	-	-
rapplerdotcom	-	-	-	TurboKitty	jilevin	-	-
YV5SEL	-	-	-	LatinoLdnOnt	thepsychicseer	-	-
BrendanCarton	-	-	-	Daily_Express	BrendanCarton	-	-

During							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
klausammann	klausammann	-	PepperInVegas	TheDisproof	TheDisproof	globaltimesnews	globaltimesnews
bdollabills	tryingBot05	-	Chris_1791	AugustEve2012	AugustEve2012	MSNBC	MSNBC
TheDisproof	shehzadyounis	-	-	drsohailmahmood	drsohailmahmood	PepperInVegas	PepperInVegas
paulinepark	realTuckFrumper	-	-	klausammann	klausammann	CGTNOfficial	CGTNOfficial
drsohailmahmood	RogueBalam	-	-	paulinepark	paulinepark	gridpointwx	Chris_1791
cnni	ARTHURGCARTER1	-	-	AandNoa	Daily_Record	CGTNEurope	-
AugustEve2012	AndyVermaut	-	-	Daily_Record	AandNoa	cgtnamerica	-
bpolitics	-	-	-	unherd	EnviroEdgeNews	-	-
EnvDefenseFund	-	-	-	EnviroEdgeNews	RojoRurba002	-	-
AandNoa	-	-	-	bdollabills	bdollabills	-	-

After							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
jftaveira1993	-	-	-	DataAugmented	DataAugmented	-	-
BrianMcHugh2011	-	-	-	jftaveira1993	BrianMcHugh2011	-	-
EsgWire	-	-	-	BrianMcHugh2011	business	-	-
latimes	-	-	-	EsgWire	jftaveira1993	-	-
Orgetorix	-	-	-	Orgetorix	commondreams	-	-
AndyVermaut	-	-	-	latimes	AndyVermaut	-	-
Eire353	-	-	-	business	EsgWire	-	-
IndianExpress	-	-	-	AndyVermaut	therightblue	-	-
commondreams	-	-	-	commondreams	IndianExpress	-	-
business	-	-	-	insideclimate	Orgetorix	-	-

Table A.3: Top 10 most important actors for each action relationship, for each stratum. This data was obtained through the JDD influence graph for the COP26 dataset.

Before							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
RobotChange	Robinsm86398738	Ceo_topcvstudio	drnoahross	GlobalUnion3	StefanPasti	filterednews	WeLnever
TinTincognito	WeLnever	moderateRepand1	AuEpochTimes	JunkScience	TB_Times	TheWatchmanNews	amadorn
Surly01	Rajit_Pathak	Lawlor224	Kathryn24498120	HarwoodEdu	BriannaATucker	YXiusheng	spennington33
HarwoodEdu	SidorDid	mary_shubert	CGTNOfficial	eduCCateGlobal	Empathy4Animal	cgtnafrica	wolters_ann
GlobalUnion3	NikolasKozloff	GuyThompson_Esq	MelanieAlex62	Surly01	MissPolly62	globaltimesnews	AlexWitzleben
eduCCateGlobal	Roark_Architect	dadsolarjohn	realTuckFrumper	RobotChange	RealKrisKo	MsLisaWilliams	BoSnerdley
Independent	SustainableWang	bmcarthur17	RebelNewsOnline	ManishKhurana	drpatfarrell	Marbahr16	DaysonRick
RFrumpf	freedomforusnow	buddy_dek	thaiparampil	weatherindia	jcrgordon5	WandaRufin	MontyNishimura
weatherindia	Ceo_topcvstudio	dook42_domini	GOVpsyopsCO2	margreis9	oxfamgbpress	FriendsOScience	WaterburyKevin
openDemocracy	lau56	Dragofix	StuWillner	nytclimate	CyberDigitalTec	southsher	WeirdWizardDave

During							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
RobotChange	PrisonPlanet	shihtzzu	david41032	eduCCateGlobal	UNinWashington	euronewsgreen	delmartian4
HarwoodEdu	LarsonKellie	KeillerDon	ChinaDaily	HarwoodEdu	laralogan	noticiasdl1	AdoreUSAalways
eduCCateGlobal	OohFa	Chris_1791	MaiaEnergyLtd	greenprofgreen	EvanUnoArt	DclareDiane	JJDJ1187
TinTincognito	MJW_DC	TerranEmpire	PepperInVegas	JunkScience	TKSitis	TerranEmpire	TheRebeluniter
highcountrynews	bruce_schlink	Yujinesque	ZechiniVicki	pablorodas	chicagomediaX	ArabNewsBiz	TimMelino
LehtmanMaria	WakeUpAmericaDR	NoNukeBailouts	MaryLis98915532	physorg_space	edwinhayward	delmartian4	Holly2360
joincurby	GoodKindHappy	memeorandum	ResetTheMatrix	LatinoLdnOnt	KaurananSpring	lilBee67569241	LarsLarsonShow
LatinoLdnOnt	SustainEurope	BoSnerdley	PaulaAlquist	openDemocracy	cspan	TheRebeluniter	shehzadyounis
weatherindia	RiverDartGaller	Metz1245John	BonVangUFO	Surly01	Nosenatorsson1	AdoreUSAalways	MaryLis98915532
GlobalUnion3	fahimmoledina7	DclareDiane	sjdemas	NatObserver	bbwlover2019	TheEdgePB	TwitchyTeam

After							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
RobotChange	MsLisaWilliams	yportbill	shihtzzu	HarwoodEdu	LynnSaxton6	DenisPetit2233	DenisPetit2233
TinTincognito	mlc11580	deeth_jim	KiplingIfby	DanAlbas	TylerPrize	deeth_jim	copper90000
GailWaldby	gupdiver	mzee26	ShaktiviryatY	eduCCateGlobal	CBCQueensPark	euronewsgreen	realTuckFrumper
HarwoodEdu	DavidGe96918347	MoeChanda	boppinmule	highcountrynews	mmyer1018	lensfocus	skjayarajskjay1
Surly01	NewaiGreen	Roark_Architect	RecentLatestVia	EnvHamilton	JeffreyGeorgeR1	CGTNEurope	deeth_jim
eduCCateGlobal	SoniCubed	DavidGe96918347	trsmiami	JM_Coppede	StarCdnPoli	cgtnamerica	Mrkalman
CCLSVN	taichinow	JohnNor19663795	Mrkalman	nprworld	athleteswalk	yportbill	PRiMOhui
GlobalUnion3	bruce_schlink	Arthur59611540	MoeChanda	checkupcbc	nprstations	DclareDiane	RecentLatestVia
CelloMomOnCars	FargoTundra	britho	360CNN	democracynow	AbbasM	GoodKindHappy	ScienceNotDogma
highcountrynews	TinTincognito	Knewz_Currently	Peter87214766	DavidWa59907969	AndreaLearned	bruce_schlink	klausammann

Table A.4: Top 10 most important actors for each action relationship, for each stratum. This data was obtained through the TE influence graph for the COP26 dataset.

Before							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
HBreen2	BurningClock	-	-	BurningClock	Reuters	-	-
tdzarnick	-	-	-	Reuters	BurningClock	-	-
guardian	-	-	-	HBreen2	HBreen2	-	-
EarthAccounting	-	-	-	axios	guardian	-	-
EveningStandard	-	-	-	WashTimes	axios	-	-
IEyeOfTheStorm	-	-	-	ConversationEDU	ConversationEDU	-	-
indy_climate	-	-	-	tdzarnick	EveningStandard	-	-
LordGittins	-	-	-	guardian	ScotNational	-	-
PoliticsKulture	-	-	-	EarthAccounting	WashTimes	-	-
bullshitjobs	-	-	-	EveningStandard	bullshitjobs	-	-

During							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
Vastuullisuus	PoliticsKulture	-	-	auto_news_feed	auto_news_feed	ArabNewsBiz	ArabNewsBiz
auto_news_feed	rpujolvives	-	-	Vastuullisuus	Telegraph	-	-
jftaveira1993	-	-	-	ECIU_UK	Vastuullisuus	-	-
dicklibertyshow	-	-	-	Daily_Express	TheCanaryUK	-	-
TheEconomist	-	-	-	PoliticsKulture	TopClimateNews	-	-
BurningClock	-	-	-	Telegraph	ECIU_UK	-	-
Daily_Express	-	-	-	TheEconomist	Daily_Express	-	-
PoliticsKulture	-	-	-	TheCanaryUK	TheEconomist	-	-
Telegraph	-	-	-	BurningClock	dicklibertyshow	-	-
NahidAlaei	-	-	-	dicklibertyshow	PoliticsKulture	-	-

After							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
BullardCenter	-	-	-	SafetyPinDaily	SafetyPinDaily	FriendsOScience	-
PhydellaLL	-	-	-	jftaveira1993	PhydellaLL	-	-
-	-	-	-	DrBobBullard	jftaveira1993	-	-
-	-	-	-	AP_Climate	AP_Climate	-	-
-	-	-	-	TIME	TIME	-	-
-	-	-	-	indy_climate	indy_climate	-	-
-	-	-	-	k__ei	BullardCenter	-	-
-	-	-	-	BullardCenter	DrBobBullard	-	-
-	-	-	-	CCLSVN	AandNoa	-	-
-	-	-	-	GlobalUnion3	CCLSVN	-	-

Table A.5: Top 10 most important actors for each action relationship, for each stratum. This data was obtained through the JDD influence graph for the COP27 dataset.

Before							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
RobotChange	AnitaTr47909916	DclareDiane	WordpeckerUSA	RobotChange	TheAusInstitute	EvrekaCo	Libertas3127
Surly01	rdgresident	sagcast452	Libertas3127	Surly01	Breakingviews	TheConWom	GoogeliArt
Reuters	WeLnever	cleverativity	pablothehat	7adair	FranklinB51	SputnikInt	WordpeckerUSA
TopClimateNews	RA40489851	brenda_spiller	SocMedBoost	ECIU_UK	VenusianAndroid	Tel22730304	buffaloon
GlobalUnion3	ILuvCO2	sovereigntyeire	plentyus	margreis9	hannah_poetin	Terry93632337	CGTNGraphics
GISP_Tweets	Unionbuster	johnaddams2022	Adyb82176144	newscientist	katereggadennis	billm9	CarbonBubble
danspena	Hokahey10	Carol38553	EnviroEdgeNews	GISP_Tweets	AntoniaJuhasz	stockypig	SocMedBoost
joincurby	MC_00_	Mikeoflondon	CGTNEurope	mommom_dayton	ScotNational	JavierC10170911	latimeralder
AAClearinghouse	Zayphar	FriendsOScience	Hedenberg	seekhopeact	artworksplus	MsLisaWilliams	sagcast452
indy_climate	abraxas1954	CestAlain	suehen4941	ClimateRadio	rosahhmaria1	latimeralder	sharonkgilbert

During							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
RobotChange	TaraYarla	psworldwide2023	AsliBasakYildiz	RobotChange	AmyMacKinnon	MsLisaWilliams	AsliBasakYildiz
TopClimateNews	GarretLebois	DclareDiane	eampact	bernieT36	fourzeesmom	ArabNewsBiz	BillSparow1
GlobalUnion3	AsliBasakYildiz	empyreanprotoc1	MarkoGreen80	GlobalUnion3	tess_woolfenden	PatriciaHoldin2	FunnyGu31492803
joincurby	bobhillbrain	coastriskcanada	SANDALIOCARMONA	empyreanprotoc1	LBCNews	Opportu727272	cappelletti_n8
FreshElecSolar	psworldwide2023	bluwndzdancing	GraviolaDOTfi	SocMedBoost	ZSchneeweiss	MelanieAlex62	MarkoGreen80
empyreanprotoc1	NewaiGreen	BigScuba99	FFF_Jour	AltayErgun	andy_webbo	WayneGabler	MelanieAlex62
danspena	cappelletti_n8	AsliBasakYildiz	Michael27174882	TopClimateNews	dw_environment	empyreanprotoc1	Michael27174882
Robert76907841	CX3PSocial	SyedaShabanaAsh	Opportu727272	JunkScience	heather09353201	AsliBasakYildiz	kcjw33
drogon_dracarys	Hotel14f	gupdiver	RossSilverstar	anamafalda1992	Andy_Olsen	GrabienMedia	CCGevirtz
MrMatthewTodd	BigScuba99	bobhillbrain	InkICan	FreshElecSolar	ClimateTreaty	arabnews	Confederate2014

After							
betweenness T-T	betweenness T-U	betweenness U-T	betweenness U-U	outdegree T-T	outdegree T-U	outdegree U-T	outdegree U-U
RobotChange	yportbill	yportbill	queenb_wiov	CAROL11959252	Construonetwork	brave0nft	cappelletti_n8
TopClimateNews	Bob12151959Bob	FloridaIssues	Gjallarhornet	RobotChange	HeatherAnne524	OrwellsRevenge	UsBurning
ClubAdaptation	brenda_spiller	Bigmoe16574013	NoahJamesBangs	empyreanprotoc1	leonpui_	KieckCarl	Bob12151959Bob
foodandwater	empyreanprotoc1	_financiallaws	Thorne75658284	BizSustainably	CAROL11959252	arizman2	Dlw20161950
BizSustainably	RealKSridharan	realTuckFrumper	MelanesianWomen	EIA_News	CleanAirUK	FriendsOScience	Thorne75658284
danspena	AlanDix44633097	RealKSridharan	mannpersonalas1	natty4bumpo	Franco931723205	ntesdorf	jujuone12
empyreanprotoc1	Taslim_Reza1	gc22gc	Bob12151959Bob	BMDD10	Sophiakeijzer	peta	worldnetdaily
GlobalUnion3	_OfficialECI	Juliett59778255	Chris_1791	BernThemAll	Taslim_Reza1	cgtnamerica	Bigmoe16574013
PhydellaLL	heather_cynical	OrwellsRevenge	LarryNeufeldSK	CaplanComms	ZimmerMar68	BruceNo40418166	CaffeineGuy1
CleanAirMoms	EarthAccounting	TopClimateNews	AllthingsWW2Kg	TopClimateNews	cambskeptics	mikekirbyone	ClimateSt

Table A.6: Top 10 most important actors for each action relationship, for each stratum. This data was obtained through the TE influence graph for the COP27 dataset.

A.3 Miscellaneous

climate query	("climate change" OR #climatechange OR #climate_change OR "climate crisis" OR #climatecrisis OR #climate_crisis OR "climate emergency" OR #climateemergency OR #climate_emergency OR "global warming" OR #globalwarming OR #global_warming OR "climate action" OR #climateaction OR #climate_action) has:links lang:en
Skripal query	(skripal OR #skripal OR novichok OR #novichok) has:links lang:en

Table A.7: Exact queries made to extract data. Note that the words are case and accent insensitive. Different part of a query without explicit logical operator are linked with logical AND (e.g. "has:links lang:en" is interpreted as "has:links AND lang:en")

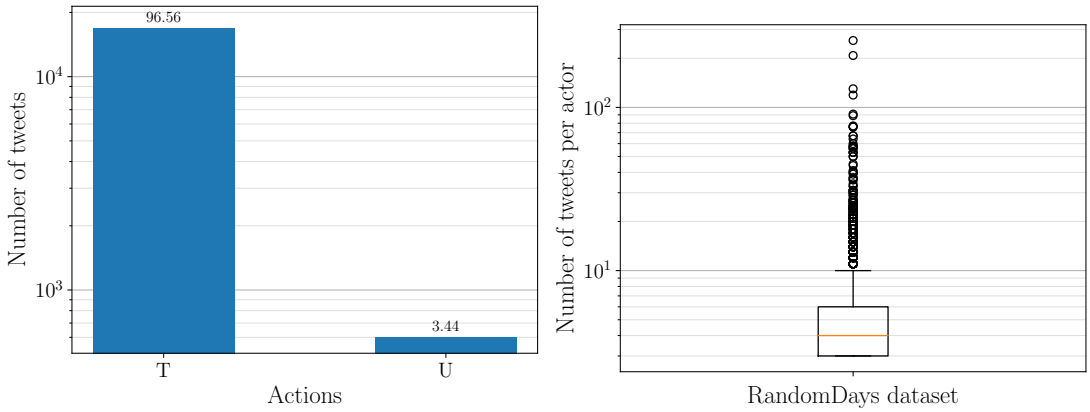


Figure A.4: (a) Actions distribution (the proportion is indicated as a number above the bar) and (b) distribution of the number of tweets per actor. These statistics describe the control set (RandomDays dataset).

news outlet	article url
globaltimes.cn	https://www.globaltimes.cn/page/202111/1238053.shtml
	https://www.globaltimes.cn/page/202111/1238795.shtml
cgtn.com	https://news.cgtn.com/news/2021-11-03/China-tells-U-S-to-take-actions-not-empty-words-on-climate-change-14TuduNq58Y/index.html
	https://news.cgtn.com/news/2021-11-02/China-highlights-arduous-efforts-it-has-made-to-fight-climate-change-14Rz4e4rRVS/index.html
breitbart.com	https://www.breitbart.com/environment/2021/11/01/xi-jinping-scolds-world-on-climate-change-while-china-keeps-polluting/
	https://www.breitbart.com/politics/2021/11/01/watch-sleepy-joe-biden-struggles-to-stay-awake-during-climate-change-summit/
thegatewaypundit.com	https://www.thegatewaypundit.com/2021/11/bidens-marxist-treasury-nominee-says-quiet-part-loud-fossil-fuel-industry-want-go-bankrupt-want-tackle-climate-change-video/
zerohedge.com	https://www.zerohedge.com/geopolitical/watch-al-gores-latest-solution-climate-change-mass-surveillance

Table A.8: Replication of Table 4.12 with entire urls.