# Social Media Discourse during Pandemic

Hackathon for Social Good

# Social Distancing due to COVID19

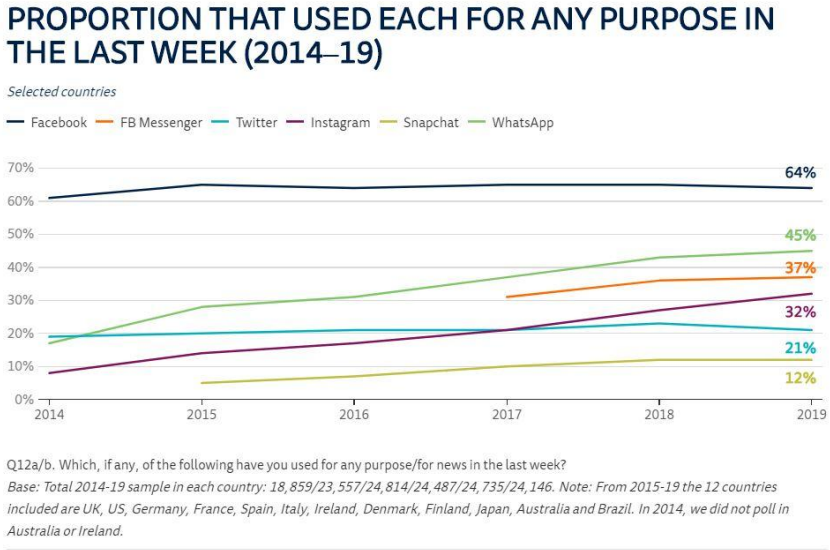
- Due to Covid19, various social distancing measures like
  - Travel bans &
  - Work-From-Home (WFH) policies were adopted.

- The forced quarantines moved people out of public spaces and a lot of conversation moved online to social networks like Twitter, Facebook groups, Reddit channels & messaging platforms.

- We will use this global opportunity to visualize this discourse

# A Note on Mis-Information

- Around 65% of all European youths access media through social networks
- So no one single person can control the information spreading through these networks (egalitarian models)
- But it also has enabled the spread of propaganda, misinformation and influence campaigns at a planetary scale!
- We will first analyse fact-checker tweets that were made during this period

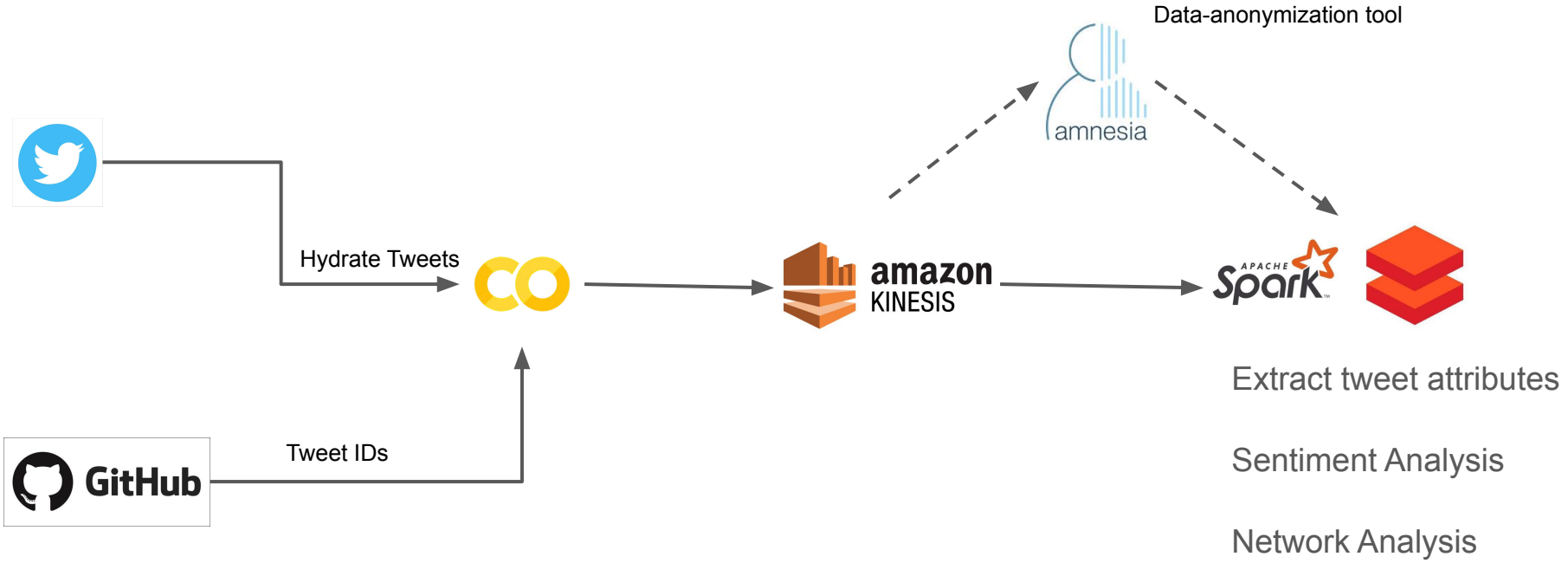



*Source:* Hoaxy a platform for Tracking misinformation
Unveiling co-ordinated groups behind white-helmet disinformation
Digital News Reports - Reuters Institute

# Twitter Developer Labs - Covid19 streaming endpoint

- Twitter has exposed a streaming endpoint for covid-19 dialogues
- It delivers free, full-fidelity data in real-time on the COVID-19 conversation
- The filtered stream contains only 1% of the conversation that matches the filtering criteria
- But it provides a good sample of the dialogues in real-time

# Data flow A - Stream Processing



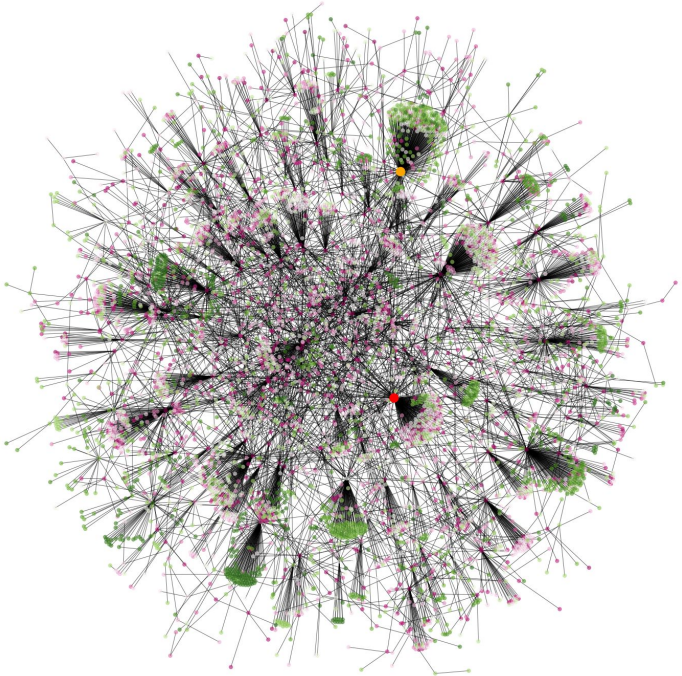

*Notebook Links:*
https://github.com/CoronaWhy/Hackathon-For-Social-Good/blob/master/Hydrating_Streaming_AWS_Kinesis.ipynb
https://github.com/CoronaWhy/Hackathon-For-Social-Good/blob/master/Covid19_Tweets_Streaming_Analysis_from_AWS_Kinesis.ipynb
https://github.com/CoronaWhy/Hackathon-For-Social-Good/blob/master/Analyzing_Tweets_PySpark_Batch_Analytics.ipynb

# Hydrating of tweets

- While analyzing large amounts of data from Twitter, exporting all the related tweets is time consuming and cumbersome. Tweet id, a unique integer representation for every tweet, comes handy to the researchers.
- Each tweet is associated with a unique tweet id, which is linked to the tweet text, user and various other details that can be used for the analysis.
- There are various tools available for hydrating the tweets such as Docnow, which is a GUI tool.
- In this project, twarc - a command line tool and python library for archiving twitter JSON data, is used

Note: During this process some IDs might be ignored as the original tweet is deleted from the database.

# Network analysis & visualization



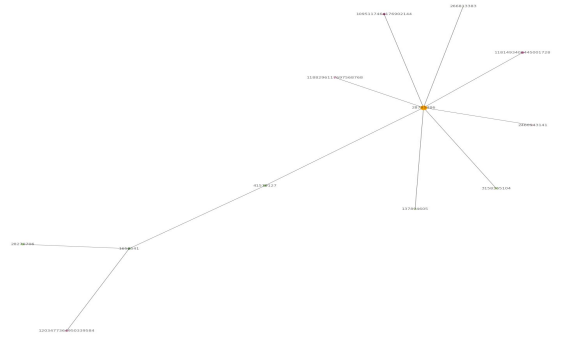
Network visualized from a sample of tweets from 23rd Jan 2020

| | |
|---|---|
| Average degree of nodes in graph | 1.90 |
| Number of node | 5,927 |
| Number of Edges | 6,724 |
| Connected components in the graph | 1,260 |
| Avg distance between two nodes | 6.56 |
| | |

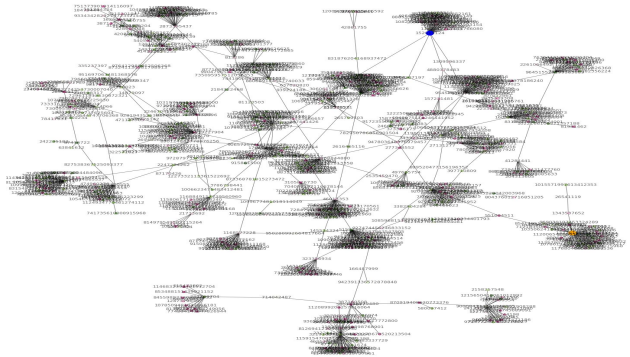
*Source: Analysis & visualization done using networkX*

*Notebook Link:*
https://github.com/CoronaWhy/Hackathon-For-Social-Good/blob/master/COVID19_NetworkX_Analysis.ipynb

# Network Visualization



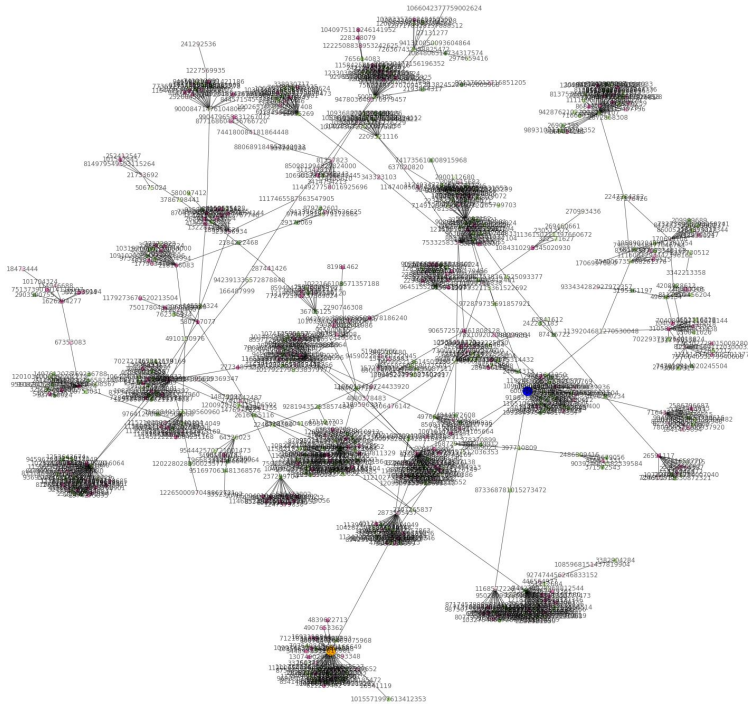
Network captured for 21 &22, Jan



Network captured on 3rd of Feb

- The prima influence during the period 21 & 22, Jan is a News channel that tweets about the outbreak
- The main influencing tweet in the start of Feb is on the panic situation among people

# Network visualization



Network captured on 18th Mar

- The most influential tweet is by a common man who blames humankind for all the destruction caused

- As it can be seen the number of people talking about COVID-19 increased rapidly from Jan, indicating how the social media is playing an important role as a form of communication and information transfer

# Analysis of Fact Checker Tweets

1. We have collected fact-checker tweets made by IFCN affiliated organization.
2. 44 user IDs and more than 85k+ tweets were collected
3. Tweets were collected from 1st Dec 2019 - 2nd June 2020
4. From the hash-tag & cluster analysis, it's evident that the pandemic triggered an avalanche of fake information
5. List of top twitter handles tweeting during this period. The second column has the number of tweets made during this period.

| | |
|---|---|
| observadorpt | 19714 |
| lemondefr | 15486 |
| Newtral | 7255 |
| snopes | 5746 |
| boomlive_in | 4190 |

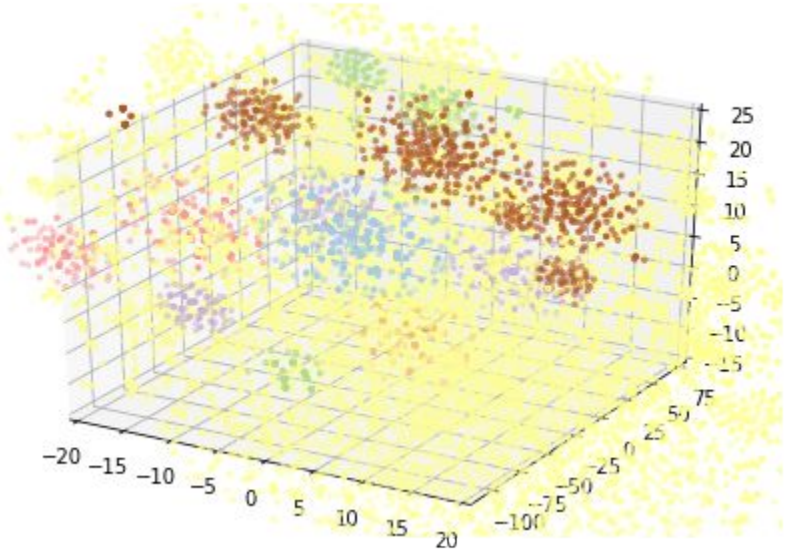
*Notebook Link:*
https://github.com/CoronaWhy/Hackathon-For-Social-Good/blob/master/Analysis_of_Fact_Checking_Tweets.ipynb

# Top most common #hash-tags with counts

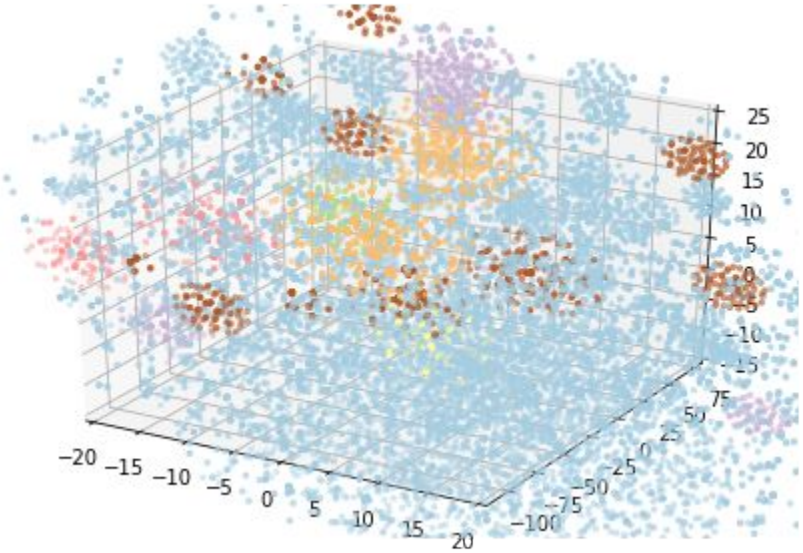
| | |
|---|---|
| #coronavirus | 1624 |
| #fakenews | 1284 |
| #covid19 | 1158 |
| #boomfactcheck | 922 |
| #coronavirusfacts | 759 |
| #faktencheck | 445 |
| #facebook | 430 |
| #coronavirusitalia | 199 |
| #covid19italia | 175 |
| #corona | 165 |

| | |
|---|---|
| #whatsapp | 156 |
| #thema | 151 |
| #fake | 145 |
| #covid_19 | 142 |
| #video | 121 |
| #coronavirusoutbreak | 120 |
| #factcheck | 111 |
| #factchecking | 108 |
| #lockdown | 107 |
| #datoscoronavirus | 107 |

# tSNE 3D visualization




k-means




K-means normalized

*Notebook Link:*
*https://github.com/CoronaWhy/Hackathon-For-Social-Good/blob/master/Fake_News_Countering.ipynb*

# Team Members

We are part of CoronaWhy.org, it's global community of volunteers from diverse backgrounds. We have come together to find solutions to problems raised by the pandemic. As such we had 5 members of our community who had volunteered for this project. We have split ourselves into two teams.

Team 1: Data extraction, pipelines & visualization

1. Aakash Gupta
2. Nithin Krishna K S
3. Ali Haider Bangash

Team 2: Modeling of fake news

1. Pranjalya Tiwari
2. Li Xueqi

# References


1. Chen E, Lerman K, Ferrara E Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set JMIR Public Health Surveill 2020;6(2):e19273 DOI: 10.2196/19273 PMID: 32427106 (https://github.com/echen102/COVID-19-TweetIDs)
2. Hoaxy services (https://hoaxy.iuni.iu.edu/)
3. Amnesia a data anonymization tool (https://amnesia.openaire.eu/index.html)
4. Twitter Developer Covid19 Streaming end-point (https://twittercommunity.com/t/new-covid-19-stream-endpoint-available-in-twitter-developer-labs/135540)
5. Reuters Institute Digital Research 2019 (http://www.digitalnewsreport.org/)
6. Misinformation during a Pandemic (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3580487)
7. Misinformation Containment in Social Networks (https://papers.nips.cc/paper/7317-on-misinformation-containment-in-online-social-networks.pdf)
8. Kaggle Dataset of Fact checks (Dec 2019 - June 2020) (https://www.kaggle.com/skylord/fact-checker-tweets)