Web Science (H)

Network based Social Media Analytics Report

James Park – 2271862p

Source Code

https://github.com/Parklife05/Twitter-Crawler-Web-Science

Sample Data

https://github.com/Parklife05/Twitter-Crawler-Web-Science/blob/master/csv/sample.csv

Introduction

The program I have developed allows for crawling of twitter data collecting tweets using keywords to track tweets. I compiled two python scripts to collect tweet data and both Twitter's Streaming and REST API's were used. My first script [getData.py] was to collect !% of data and was ran on 29/03/2020 at 2pm for 1.5 hours. Tweets were gathered and then classified for word, username frequency and saved to a MongoDB database. I refined my crawler, and selected keywords with my classified data and ran my second script [getAllData.py] on the 30/032020 at 3pm for 1.5 hours. Tweets were again gathered and then classified, this time by through the use of K-Means clustering and saved again to my database. I referenced this site for in compiling my initial crawling scripts:

https://www.toptal.com/python/twitter-data-mining-using-python

I compiled a script to [cluster.py] to extract usernames and hashtags and printed the sorted results to my [output.csv] file. Network interaction graphs were compiled and feature later in this report ,

Data Crawl

The Tweepy library was used for crawling and the scripts mentioned above [getData.py] & [getAllData.py] were compiled for gathering the data. After my initial crawl I compiled a script [wordFreq.py] to parse the database and extract most common word frequencies in order to refine my twitter crawler. After studying the data I felt that the terms below appeared frequently and I based my keywords for my refined crawler on them.

```
coronavirus,2602
@realDonaldTrump,2113
quarantine,1892
```

The keywords I tracked for my second stream in [getData.py] were:

```
keywords = ['Covid-19', 'Coronavirus', 'self-isolation', 'stay at home',
'#flatenthecurve',
            'social distancing', 'pandemic', 'virus', '@RealDonalTrump',
'#StayHomeSaveLives', '@GOVUK',
            '@10DowningStreet', 'Corona Virus', '@BorisJohnson ', 'coronavirus
deaths', 'world health organisation',
            '#COVID_19', '#coronavirus', '@WHO', 'quarantine', '#Quarantine'
'donald', 'trump', 'boris', 'johnson',
            'corona', 'virus', 'test positive covid-19']
```

I also obtained the most frequent usernames from my [wordFreq.py] script and tracked them they were:

```
users = ['25073877', '3131144855', '17481977', '14499829', '14224719']
```

Tweet Grouping

For the next analytical tasks for data gathering I compiled further scripts which include [cluster.py], this script used kMeans clustering using the SKLearn library to analyse the data by extracting and clustering the top 5 Users and Hashtags. Output below:

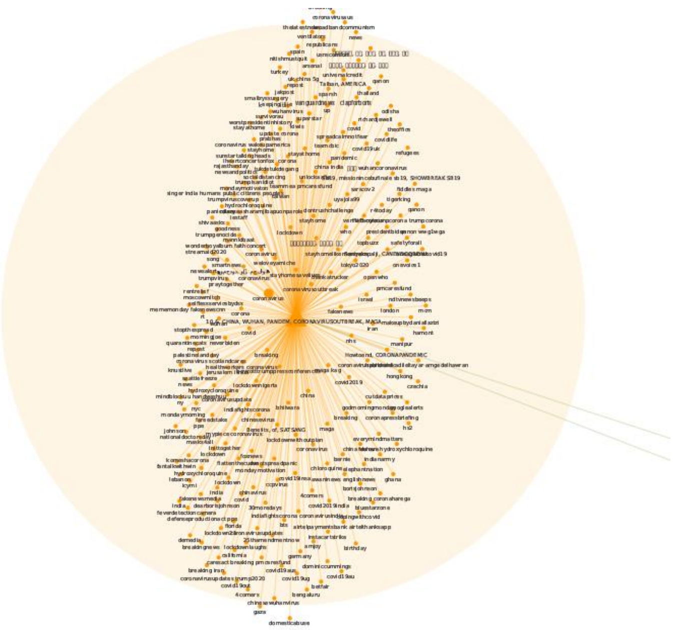
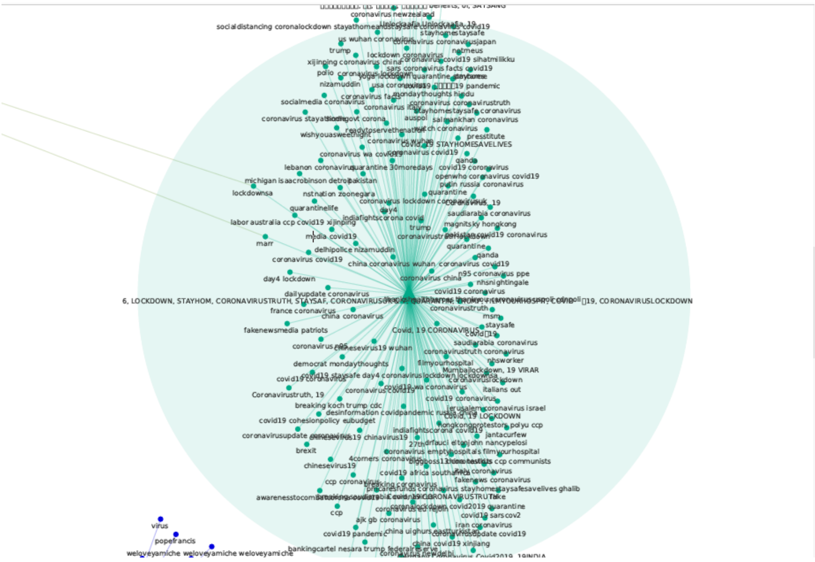
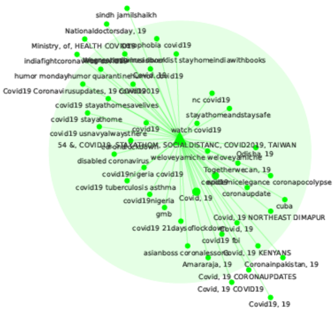
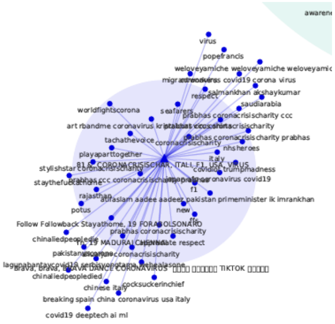
```
=== Preforming KMeans with 10 clusters ===

        === Cluster 0, Size: 354762 ===

=== Usernames ===
[('@realDonaldTrump', 3), ('@joyb37', 2), ('@Maestro_Mathur', 2), ('@BigTony2014', 2),
('@marinasoltan_', 2)]

=== Hashtags ===
[('#singer', 1), ('#India', 1), ('#humans', 1), ('#public', 1), ('#citizens', 1)]

        === Cluster 1, Size: 354762 ===

=== Usernames ===
[('@Mzakal91058', 3), ('@MzalakaMzala2', 2), ('@MzalendoShujaa', 2), ('@sakaHammed11',
1), ('@alakijaofficial', 1)]

=== Hashtags ===
[('#COVID19S', 1)]


        === Cluster 2, Size: 354762 ===

=== Usernames ===
[('@MzalendoShujaa', 3), ('@Mzakal91058', 3), ('@MzalakaMzala2', 2), ('@ouchinagirl', 1),
('@WhiteHouse', 1)]

=== Hashtags ===
[('#COVID19S', 1)]


        === Cluster 3, Size: 354762 ===

=== Usernames ===
[('@MzalendoShujaa', 4), ('@Mzakal91058', 2), ('@MzalakaMzala2', 2), ('@DcdRetblue', 1),
('@KR_KAG', 1)]

=== Hashtags ===
[('#COVID19S', 1)]

        === Cluster 4, Size: 354762 ===

=== Usernames ===
[('@MzalendoShujaa', 3), ('@Mzakal91058', 3), ('@MzalakaMzala2', 2), ('@cmcqueen47', 1),
('@mitchellvii', 1)]

=== Hashtags ===
[('#COVID19S', 1)]

        === Cluster 5, Size: 354762 ===

=== Usernames ===
[('@MzalendoShujaa', 5), ('@MzalakaMzala2', 2), ('@Amethystinia', 1), ('@FT_SriLanka', 1),
('@zzzzzzzkrtls', 1)]

=== Hashtags ===
[('#lka', 1), ('#COVID19S', 1)]

        === Cluster 6, Size: 354762 ===

=== Usernames ===
[('@Mzakal91058', 3), ('@MzalakaMzala2', 2), ('@MzalendoShujaa', 2), ('@pretendasaur', 1),
('@Lowkey0nline', 1)]

=== Hashtags ===
[('#COVID19S', 1)]

        === Cluster 7, Size: 354762 ===
=== Usernames ===
[('@Mzakal91058', 4), ('@MzalakaMzala2', 2), ('@aqila_yayah', 1), ('@thisiswafiy', 1),
('@zzzzzzzkrtls', 1)]

=== Hashtags ===
[('#COVID19S', 1)]

        === Cluster 8, Size: 354762 ===

=== Usernames ===
[('@MzalendoShujaa', 3), ('@Mzakal91058', 2), ('@MzalakaMzala2', 2), ('@FelonyHarlem11',
1), ('@zzzzzzzkrtls', 1)]

=== Hashtags ===
[('#COVID19', 1), ('#StaySafe', 1), ('#Day4', 1), ('#CoronaVirusLockdown', 1),
('#LockdownSA', 1)]

        === Cluster 9, Size: 354762 ===

=== Usernames ===
[('@MzalendoShujaa', 4), ('@Mzakal91058', 3), ('@MzalakaMzala2', 2), ('@theannuarya', 1),
('@duttsanjay', 1)]

=== Hashtags ===
[('#COVID19', 1), ('#StayAtHomeSaveLives', 1), ('#Corona', 1), ('#COVID19S', 1)]
[['#singer', '#India', '#humans', '#public', '#citizens', '#peoples', '#coronavi', '#PMCaresFunds',
'#CoronaVirus', '#StayHomeStaySafeSaveLives', '#Ghalib', '#everyone', '#bad', '#around',
'#world', '#virus', '#should', '#TheStep'], ['#COVID19S'], ['#COVID19S'], ['#COVID19S'],
['#COVID19S'], ['#lka', '#COVID19S'], ['#COVID19S'], ['#COVID19S'], ['#COVID19',
'#StaySafe', '#Day4', '#CoronaVirusLockdown', '#LockdownSA', '#COVID19S'], ['#COVID19',
'#StayAtHomeSaveLives', '#Corona', '#COVID19S']]
```

Method for Capturing & Organising User and hashtag information

User interaction and hashtag information was used to compile interaction graphs using the packages, NetworkX. MatPlotLib and web based network tool at:

https://www.cortext.net/

The Hashtag data was extracted from the database using the compiled script [hashtags.py] and the output was saved, uploaded to the above site and a network mapped. The result can be seen in the [Hashtags_network.pdf] file in the root folder of this project and in the figures below.


coronavirusus
thailand
news
republicans
spain
turkey
Taliban, AMERICA
qanon
thailand
odisha
superstar
staysafe
stayathome
coronavirus
pandemic
refugees
lockdown
who
nhs
china
covid
covid19
trump
maga
Howtoend, CORONAVIRUSPANDEMIC
coronavirus
breaking
hongkong
covid2019
china
3, CHINA, WUHAN, PANDEMIC, CORONAVIRUSOUTBREAK, MAGA
israel
london
iran
manipur
fakenews
florida
ghana
germany
birthday
dominiccummings
covid19uk
bengaluru
domesticabuse
gates
chinesewuhanvirus
breakingnews
newyork
covid19sa
indiafightscorona
lockdown
staysafe
4corners
covid2019india
borisjohnson

coronavirus newzealand
socialdistancing coronalockdown stayathome
us wuhan coronavirus
trump
lockdown coronavirus
xijinping coronavirus china
sars coronavirus facts covid19
polio
nizamuddin
usa coronavirus
covid19 pandemic
mondaythoughts hindu
socialmedia coronavirus
coronavirus coronavirustruth
coronavirus stayathome
stayhomestaysafe coronavirus
wishyouasweetnight
sanctuary
presstitute
coronavirus wa covid19
COVID19 STAYHOMESAVELIVES
lebanon coronavirus
quarantine 30moredays
covid19 coronavirus
michigan isaacrobinson detroit pakistan
openwho coronavirus covid19
putin russia coronavirus
lockdownsa
quarantine
nstnation zoonegara
quarantinelife
coronavirus lockdown coronavirus
saudiarabia coronavirus
labor australia ccp covid19 xijinping
indiafightscorona covid
magnitsky hongkong
trump
media covid19
coronavirustruth
marr
delhipolice nizamuddin
quarantine
coronavirus covid19
qanda
china coronavirus wuhan
coronavirus covid19
day4 lockdown
coronavirus china
n95 coronavirus ppe
nhsnightingale
dailyupdate coronavirus
covid19 coronavirus
6, LOCKDOWN, STAYHOM, CORONAVIRUSTRUTH, STAYSAF, CORONAVIRUS
19, CORONAVIRUSLOCKDOWN
france coronavirus
coronavirustruth
china coronavirus
msm
fakenewsmedia patriots
Covid, 19 CORONAVIRUS
staysafe
coronavirus n95
chinesevirus19 wuhan
saudiarabia coronavirus
coronavirustruth coronavirus
democrat mondaythoughts
nhsworker
filmyourhospital
covid19 staysafe day4 coronaviruslockdown
Mumbailockdown, 19 VIRAR
covid19 coronavirus
coronaviruslockdown
covid19 wa coronavirus
coronaviruscovid19
italians out
Coronavirustruth, 19
covid19 coronavirus
breaking koch trump cdc
jerusalem coronavirus israel
desinformation covidpandemic russia
Covid, 19 LOCKDOWN
covid19 cohesionpolicy eubudget
hongkongprotestors polyu ccp
indiafightscorona covid19
coronavirusupdate coronavirus
chinesevirus19 chinavirus19
jantacurfew
27th
drfauci eltaljohn nancypelosi
brexit
coronavirus emptyhospitals filmyourhospital
4corners coronavirus
chinesevirus19
ccp communists
italy coronavirus
covid19 africa southafrica
ccp coronavirus
breaking coronavirus
fakenews coronavirus
coronavirus stayhomestaysafesavelives ghalib
awarenesstocombatcoronavirus
Covid, 19 CORONAVIRUSTRUTH
fake
coronavirus covid2019 quarantine
ccp
covid19 sarscov2
ajk gb coronavirus
iran coronavirus
covid19 pandemic
china uighurs easterturkistan
coronavirusupdate covid19
china covid19 xinjiang
bankingcartel nesara trump federalreserve
virus
popefrancis
weloveyamiche weloveyamiche weloveyamiche

sindh jamilshaikh
Nationaldoctorsday, 19
Ministry, of, HEALTH COVID19
xenophobia covid19
indiafightcorona
stayhomeindiawithbooks
humor mondayhumor quarantinelife
Covid, 19
Covid19 Coronavirusupdates, 19 COVID2019
covid19 stayathomesavelives
nc covid19
stayathomeandstaysafe
covid19 stayathome
covid19 usnavyalwaysthere
watch covid19
covid19
54 &, COVID19, STAYATHOM, SOCIALDISTANC, COVID2019, TAIWAN
disabled coronavirus
weloveyamiche weloveyamiche
Odisha, 19
Togetherwecan, 19
covid19nigeria covid19
epidemicelegance coronapocolypse
covid19 tuberculosis asthma
coronaupdate
covid19nigeria
Covid, 19
cuba
gmb
Covid, 19 NORTHEAST DIMAPUR
covid19 21daysoflockdown
Covid, 19
covid19 fbi
asianboss coronalessons
Covid, 19 KENYANS
Amararaja, 19
Coronainpakistan, 19
Covid, 19 CORONAUPDATES
Covid, 19 COVID19
Covid19, 19

awarene
virus
popefrancis
weloveyamiche weloveyamiche weloveyamic
migrantworkers covid19 corona virus
salmankhan akshaykumar
respect
saudiarabia
worldfightscorona
seafarers
prabhas coronacrisischarity ccc
art rbandme coronavirus
coronacrisischarity
tachathevoice coronacrisischarity
prabhas coronacrisischarity prabhas
nhsheroes
playaparttogether
italy
stylishstar coronacrisischarity
coronavirus trumpmadness
coronavirus covid19
staythefuckathome
f1
rajasthan
assalam aadee aadeez pakistan primeminister ik imrankhan
potus
new
Follow Followback Stayathome, 19 FORABOLSONARO
chinaliedpeopledied
prabhas coronacrisischarity
19 MADURAI Candidate respect
pakistan
coronacrisischarity
chinaliedpeopledied
cocksuckerinchief
chinese italy
breaking spain china coronavirus usa italy
covid19 deeptech ai ml

Another script was compiled [interactions.py] which classifies the tweets and constructs network interaction graphs:

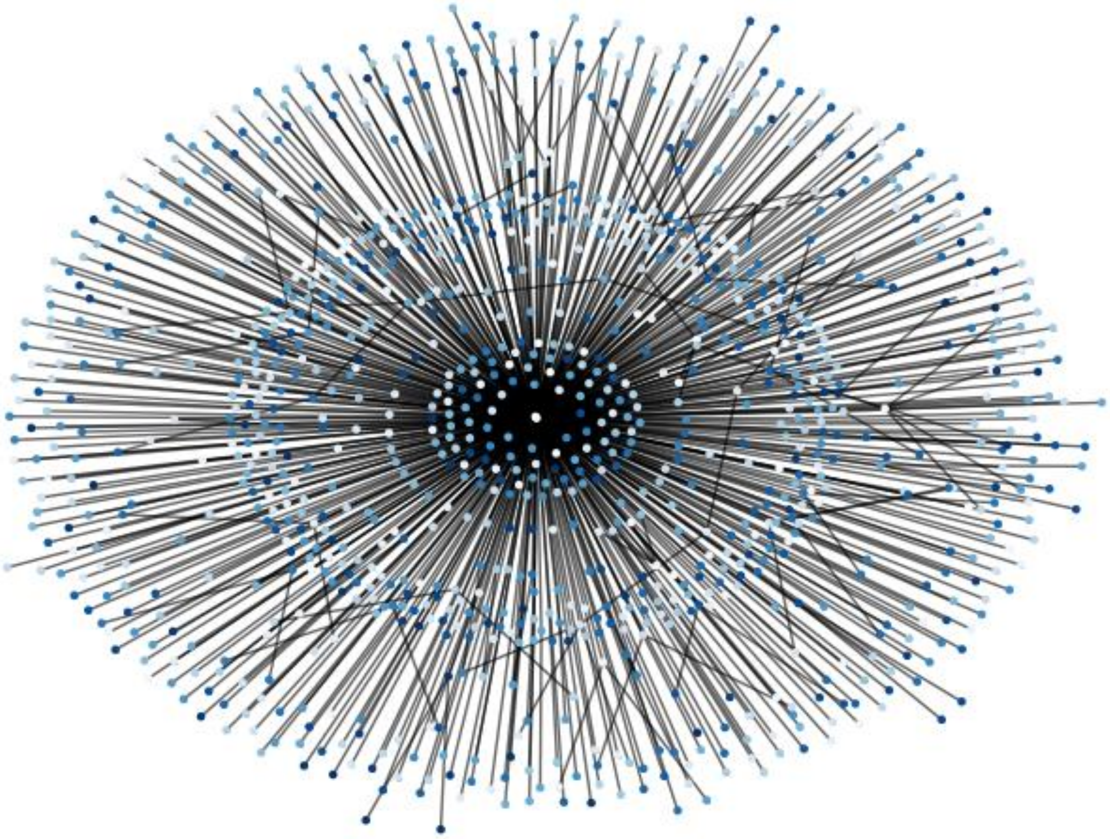
- Replies
- Retweets
- Mentions
- Quoted text

Pandas data frame package are used to organise the data and the NetworkX package is used to graph the interactions. A sample size was used for these functions of 7000, this was due to time constraints, sample size could be increased and factored into any future work.

- Nodes → Users
- Edges → Interactions

Network Analysis

Replies:

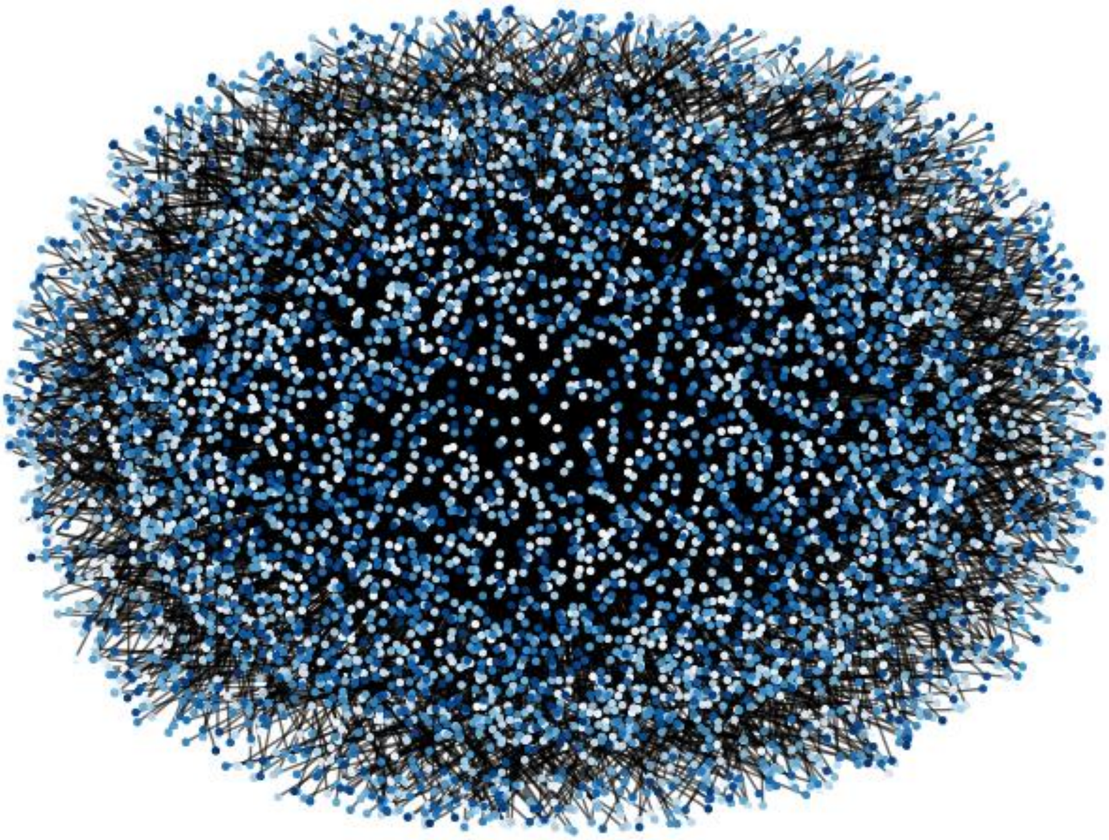


Graph Data

- 1069 nodes

- 431 edges
- maximum degree 596
- minimum degree 1
- average degree 2.3
- most frequent degree 2

Retweets:

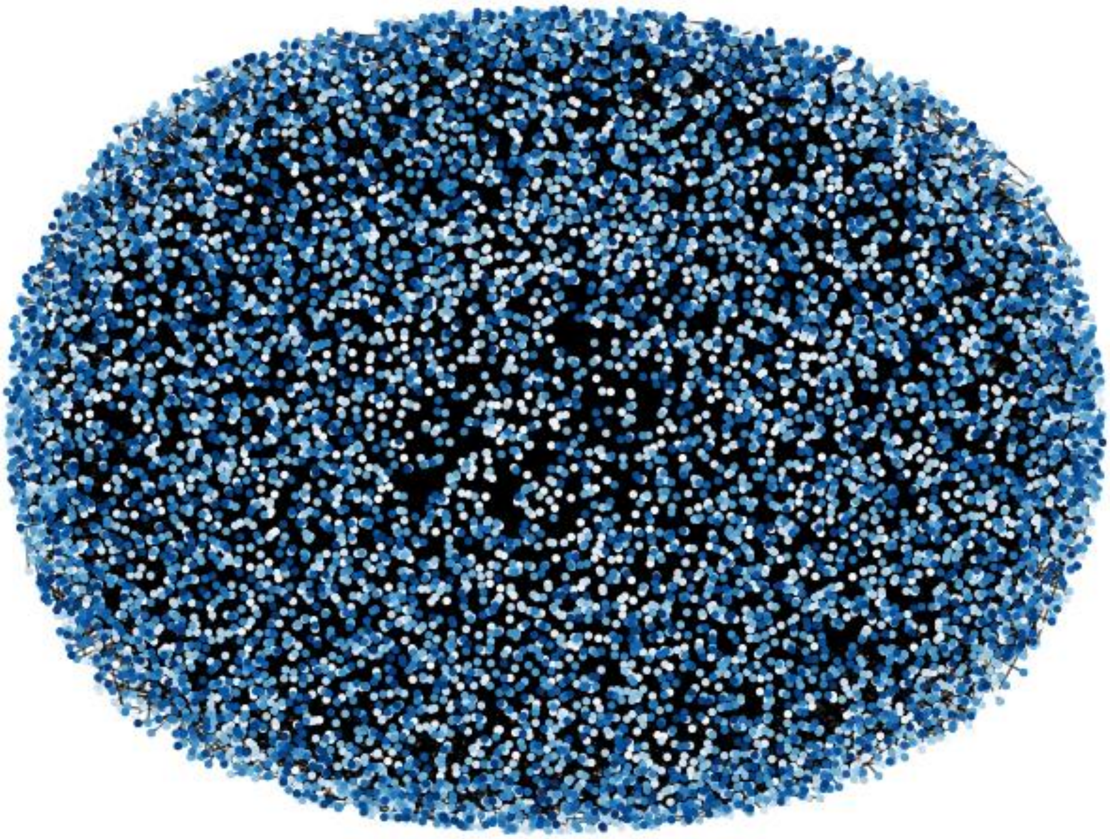


Graph Data

- 7760 nodes
- 6935 edges

- maximum degree 4609
- minimum degree 1
- average degree 2.7
- most frequent degree 2

Mentions:

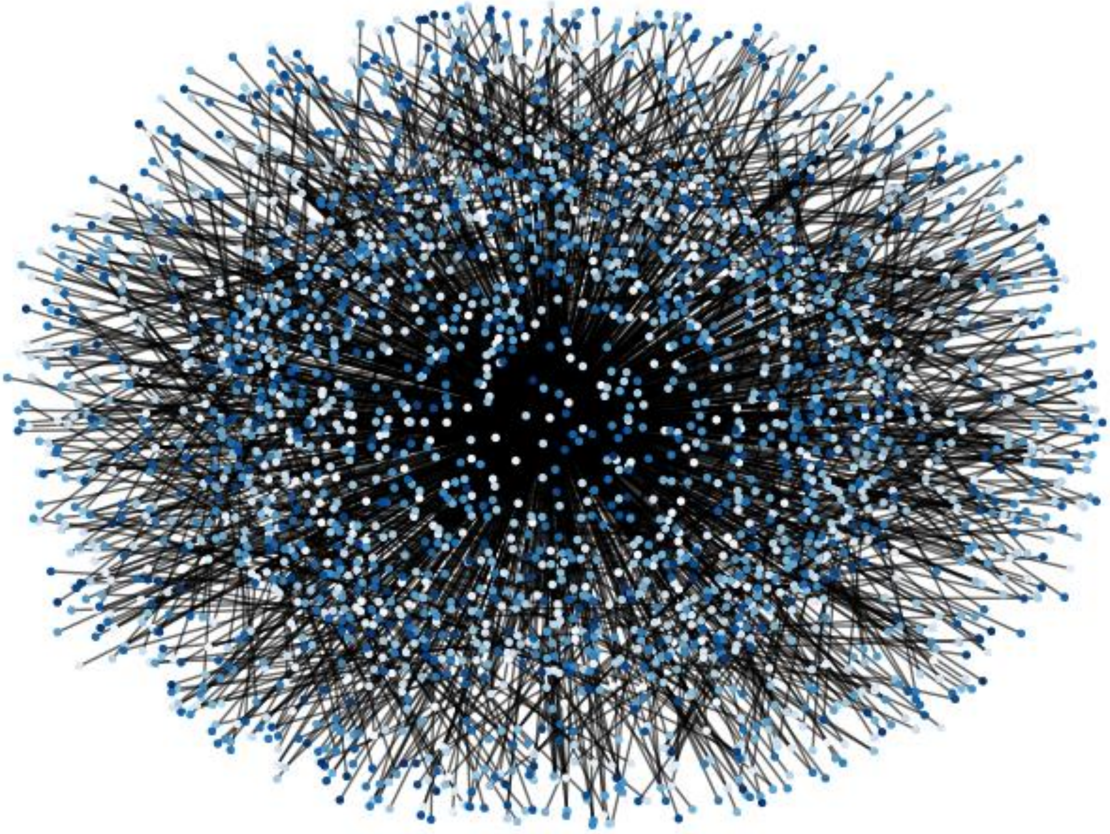


Graph Data

- 10020 nodes

- 9945 edges
- maximum degree 6239
- minimum degree 1
- average degree 2.6
- most frequent degree 2

Quotes:



Graph Data

- 2589 nodes
- 1345 edges

- maximum degree 1267
- minimum degree 1
- average degree 2.7
- most frequent degree 1