Analysis of Social Graphs

Data Science Homework #3
Sean Herman, Daniel Reidler, Haiwei Su

GDELT Analysis

We utilized BigQuery and NetworkX to perform a basic analysis of various world governments' relationship with the business community. To retrieve relevant data from the GDELT dataset, we limited our search to events with the 05 and 11 EventCode prefixes. Based on details covered in the "CAMEO Event and Actor Codebook", the 05 prefix covers "ENGAGE IN DIPLOMATIC COOPERATION" (e.g., 051 "Praise or endorse") and the 11 prefix covers "DISAPPROVE" (e.g., 111 "Criticize or denounce") type events. This was intended to more limit the results to events involving supportive language and events involving critical language.

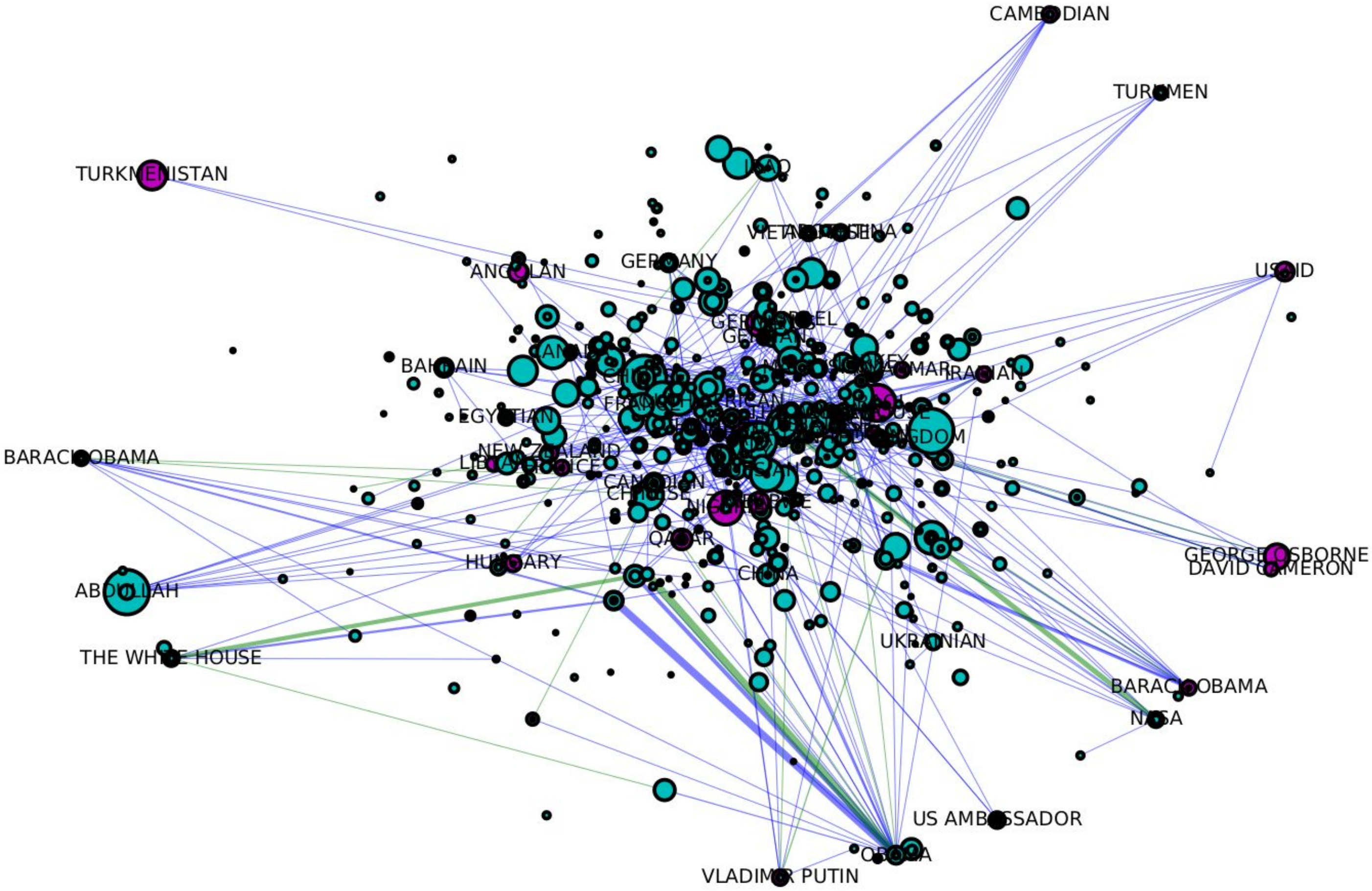
```
WHERE (REGEXP_MATCH(EventCode, '^05.*') OR REGEXP_MATCH(EventCode, '^11.*'))
```

The data is further filtered to government type Actors1 and business type Actors2.

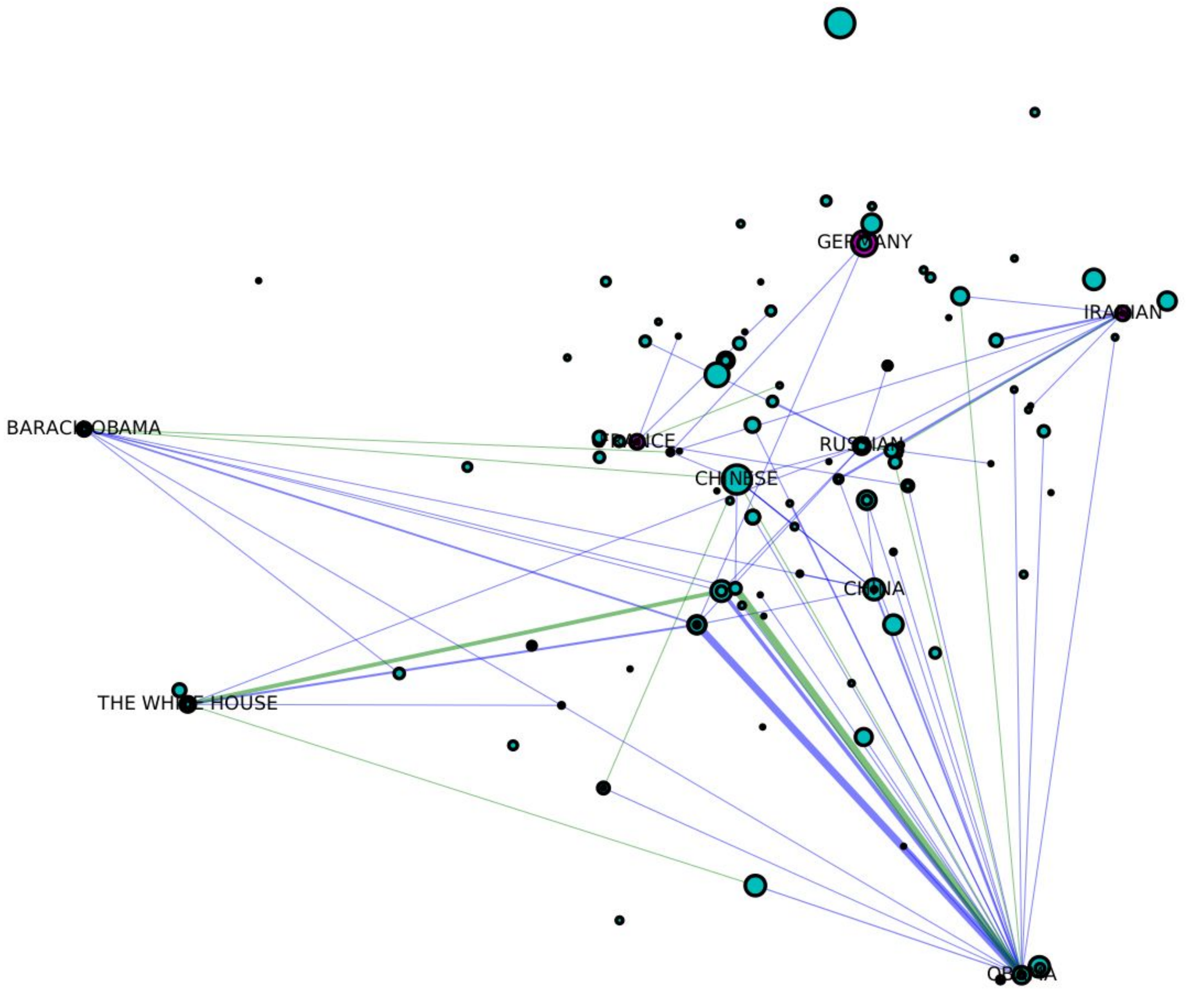
```
AND Actor1Type1Code = 'GOV'
AND (Actor2Type1Code = 'BUS' OR Actor2Type2Code = 'BUS')
```

Once retrieved and downloaded, we import this GDELT data into 2 separate directed graphs. The first graph covers all results with the 05 EventCode prefix ("cooperate" graph) and the second covers all results with the 11 EventCode prefix ("disapprove" graph).

When drawing the 2 graphs, we filtered the graph to include only "government" nodes with an out degree of at least 1. To further enhance the visibility of the graph, we color "government" nodes magenta, and color "business" nodes cyan. Further, edges with a positive tone are colored blue, while edges with a negative tone are colored green. Finally, the nodes are sized by their respective out (gov) and in (business) degree. The edge widths reflect the relative count of events covered by the edge.



05* / Cooperate graph



11* / Disapprove graph

PageRank Analysis

We use built in pagerank function of NetworkX. We tried to find pagerank of nodes of graph in both cooperate graph and disapprove graph.
From the pagerank data generated, we observed that in both graphs, the value of pagerank is quite small, this indicates that nodes in graph have little page referenced to them. We also found that there’s a hierarchical structure revealed in pagerank data. For example, in cooperate graph, the pagerank score of node named “ Abdullah” is equal of the sum of pagerank scores nodes it pointed to. Similarly, this hierarchical structure happened in disapprove graph as well.

In cooperate graph, the most influential actor is the node named “ Philip Hammond”. In disapprove graph, node “ Philip Hammond” is also the most influential actor.

Extra Credit: Reddit Comments Analysis

Here is a link to the ipython notebook (“Reddit Data.ipyn”) showing our workflow to analyze Reddit Comments.

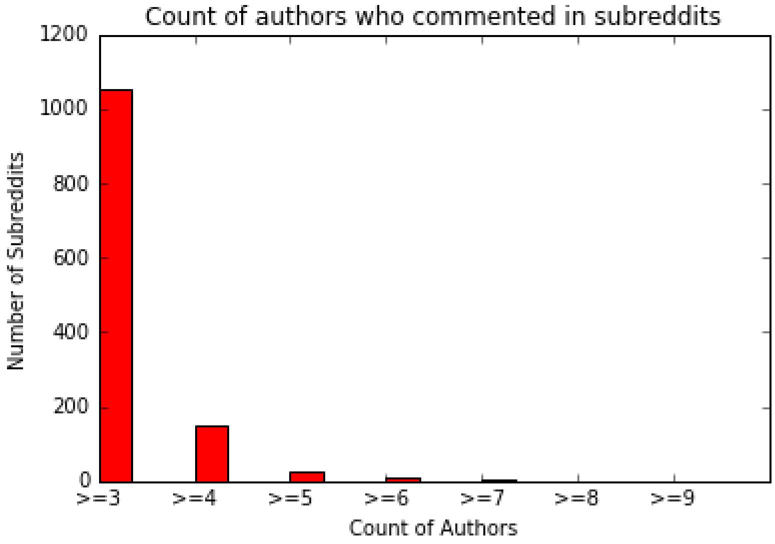
In short, we look at 10 randomly selected SubReddit groups with the following IDS:

subreddit	subreddit_id	unique authors
InternetIsBeautiful	t5_2ul7u	15199
GetMotivated	t5_2rmfx	9833
UpliftingNews	t5_2u3ta	9705
tf2	t5_2qka0	9089
techsupport	t5_2qioo	10821
Minecraft	t5_2r05i	11254
iamverysmart	t5_2yuej	7274
unitedkingdom	t5_2qhqb	10944
Smite	t5_2stl8	9760
history	t5_2qh53	10078

We proceed to analyze the number of authors who have commented on more than 2 of the 10 subreddit groups. There are 1,052 authors which makes the data more manageable.

author	numb_subreddits
CCV21	3
I*Hate_Idiots*	3
GlennDrexler	3
Fausthor	3
-Stupendous-Man-	3
....	....

TotesMessenger	7
JoeBidenBot	7
autowikibot	7



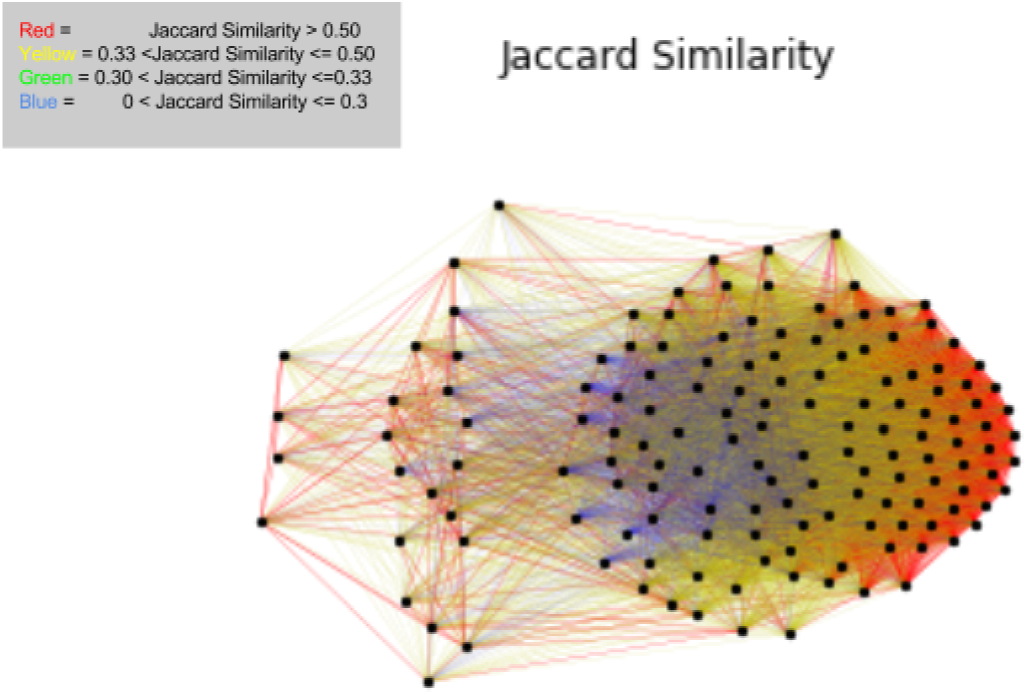
Next, we query to find out how many common subreddits authors have commented. (e.g. GlennDrexler and JoeBidenBot commented in Minecraft & techsupport).

We constricted the data to authors > 3 subreddits. (So, we are analyzing the relationship of 150+ authors).

See sample results below.

Author1	Author2	Jaccard Similarity
Dad_Jokes_Inbound	crysisnotaverted	0.333333
autowikibot	crysisnotaverted	0.375
JoeBidenBot	crysisnotaverted	0.571429
Coffeechipmunk	crysisnotaverted	0.333333
LittleHelperRobot	crysisnotaverted	0.285714
antdude	crysisnotaverted	0.285714
TweetsInCommentsBot	crysisnotaverted	0.571429

BHOP_TO_NEUROFUNK	crysisnotaverted	0.6
TotesMessenger	crysisnotaverted	0.375


Red = Jaccard Similarity > 0.50
Yellow = 0.33 <Jaccard Similarity <= 0.50
Green = 0.30 < Jaccard Similarity <=0.33
Blue = 0 < Jaccard Similarity <= 0.3
Jaccard Similarity