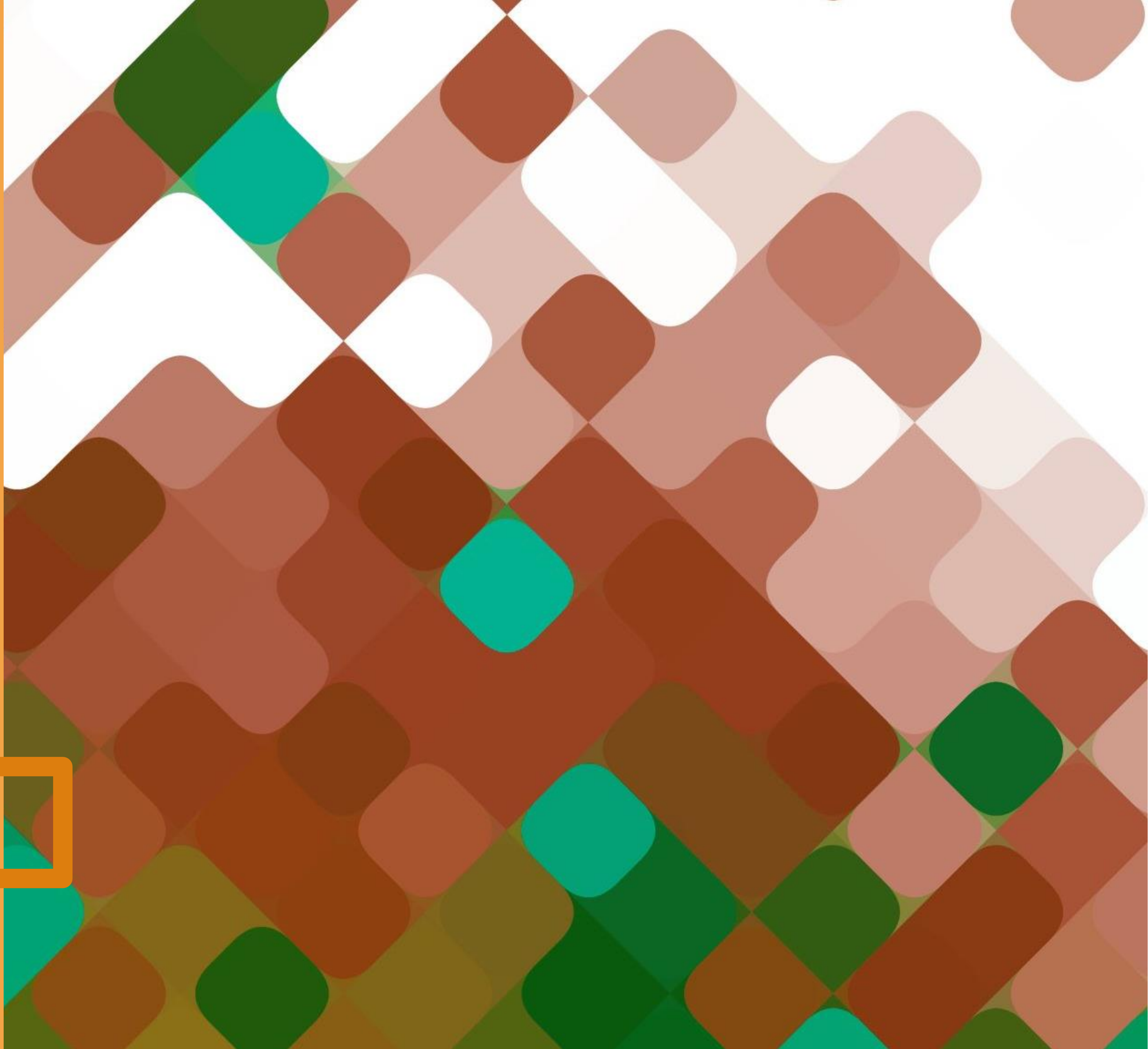
Summary: May-July Research

Alex Nazareth

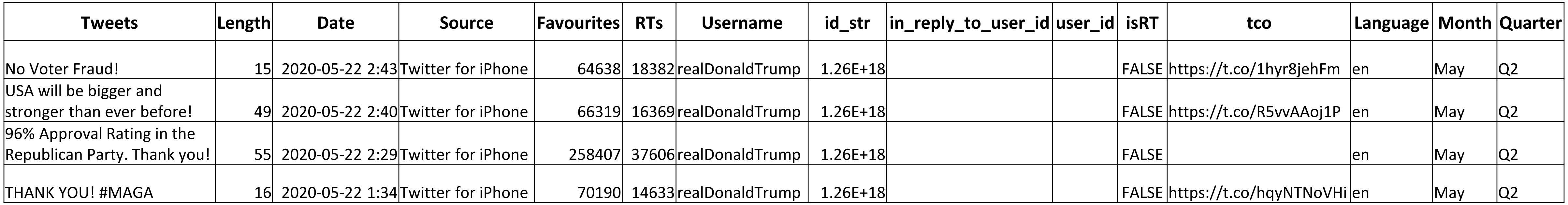
August 2020



Methodology

Data Collection - Python Scraping

- Get @realDonaldTrump tweets from Trump Twitter Archive [1]
- For any other accounts (e.g. @JustinTrudeau, @AOC), use a custom built python application. App uses *tweepy* library [2] to access Twitter API.
    1. Gather tweets from timeline of user between two dates (by tweet id).
    2. Remove retweets and non-English tweets.
    3. Divide tweets into individual words and remove common stopwords.



Tweets	Length	Date	Source	Favourites	RTs	Username	id_str	in_reply_to_user_id	user_id	isRT	tco	Language	Month	Quarter
No Voter Fraud!	15	2020-05-22 2:43	Twitter for iPhone	64638	18382	realDonaldTrump	1.26E+18			FALSE	https://t.co/1hyr8jehFm	en	May	Q2
USA will be bigger and stronger than ever before!	49	2020-05-22 2:40	Twitter for iPhone	66319	16369	realDonaldTrump	1.26E+18			FALSE	https://t.co/R5vvAAoj1P	en	May	Q2
96% Approval Rating in the Republican Party. Thank you!	55	2020-05-22 2:29	Twitter for iPhone	258407	37606	realDonaldTrump	1.26E+18			FALSE		en	May	Q2
THANK YOU! #MAGA	16	2020-05-22 1:34	Twitter for iPhone	70190	14633	realDonaldTrump	1.26E+18			FALSE	https://t.co/hqyNTNoVHi	en	May	Q2

[1] http://www.trumptwitterarchive.com/archive
[2] https://www.tweepy.org/

Keyword Analysis and Community Detection

- Get top 100 most frequent keywords appearing in tweets by month, quarter, and overall.
- Create a 100-by-100 weighted adjacency matrix for each period.
  - This represents a weighted graph where the nodes are keywords, and the edge weight corresponds to the number of co-occurrences of keywords within a tweet.
  - E.g. Add 1 to the weight of the edge between nodes “sleepy” and “joe” whenever both words appear in the same tweet.
- For each period (month, quarter, year-to-date) find the likely number of communities.
  - Use the *python-louvain* package [3] to get the number of communities from a given matrix.
  - Note that the algorithm has an element of randomness, so the community detection is run 100 times to find the best value.

[3] https://python-louvain.readthedocs.io/en/latest/

Tweets of Politicians

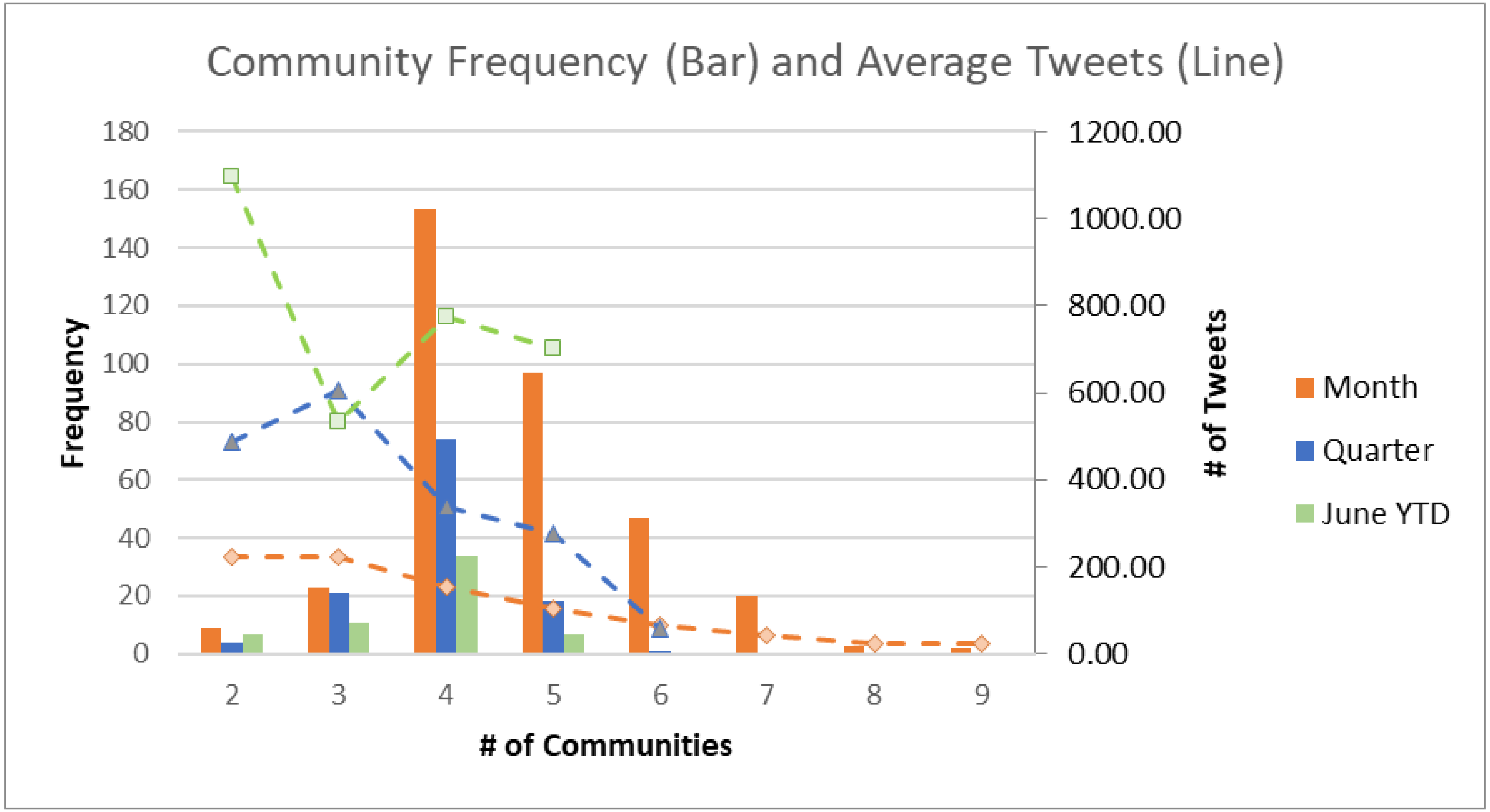
All tweets compiled in *all_users_clean.csv*

Results found in *num_communities.xlsx* sheet *num_communities*

Accounts Gathered

- All governors (incl. D.C. Mayor Muriel Bowser) <u>except</u>:
  - John Bel Edwards (LA), Charlie Baker (MA), Gretchen Whitmer (MI), Mike DeWine (OH), Phil Murphy (NJ)
- Toronto City Councillors:
  - Joe Cressy, Josh Matlow, Kristyn Wong-Tam
- Other US Political Figures:
  - Donald Trump, Alexandria Ocasio-Cortez, Bernie Sanders, Lindsay Graham, Nancy Pelosi, Mitch McConnell, Bill DeBlasio
- Other Canadian Political Figures:
  - Justin Trudeau, Doug Ford, Jagmeet Singh, Andrew Scheer

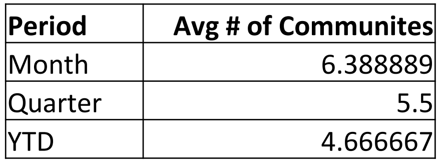
Results from Political Tweets



I can also split this into 2 graphs if you want, I think this one may be unclear.

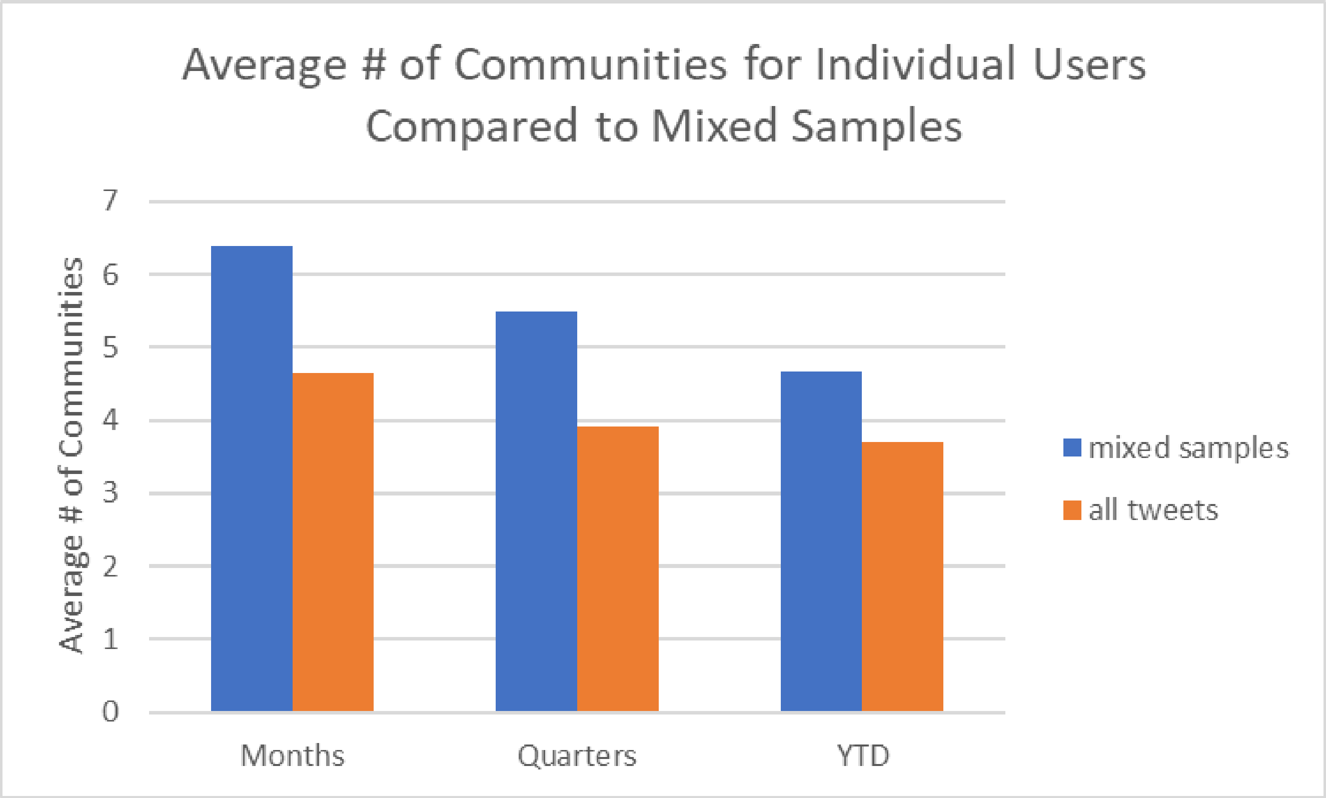
Results from Political Tweets - 2

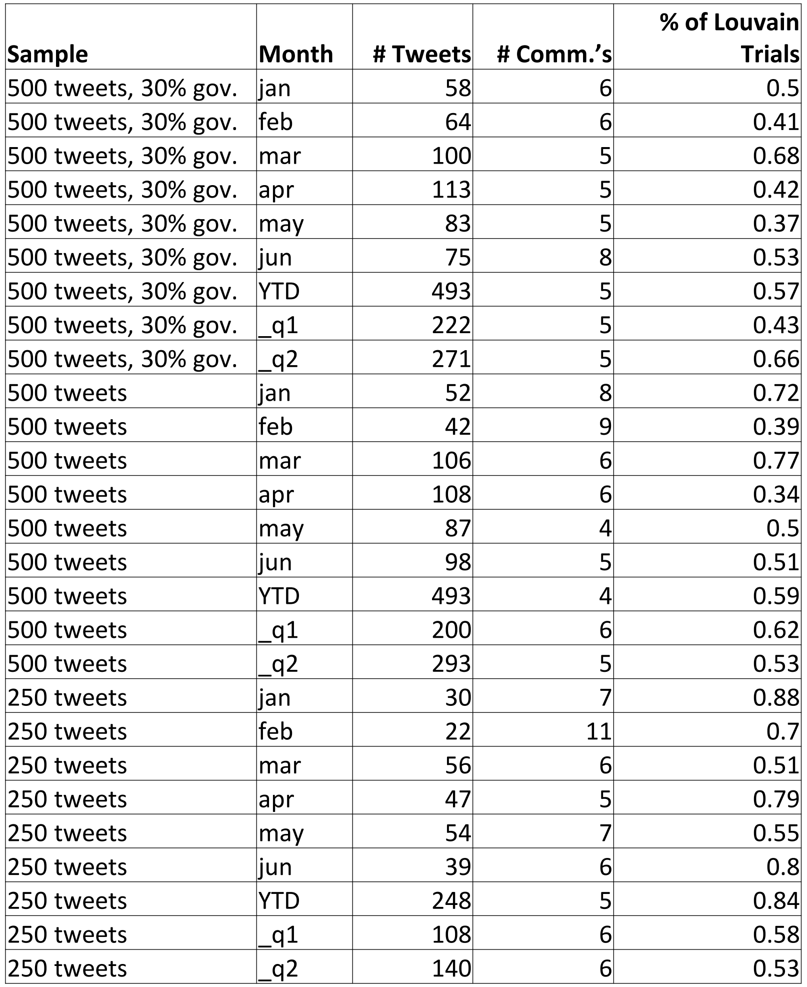
- Expected larger number of communities due to many different users (different language/word choice).
- Results were slightly higher in 3 samples analyzed:
  - Taken from all users mixed together.
  - 500 tweets, 250 tweets, and 500 tweets with only 30% US governors
  - (The 3rd sample had a specific proportion because I suspected the governors might be too self-similar, though it didn't appear to have any effect in the end)



Period	Avg # of Communites
Month	6.388889
Quarter	5.5
YTD	4.666667

Results from Political Tweets - 3





Sample	Month	# Tweets	# Comm.'s	% of Louvain Trials
500 tweets, 30% gov.	jan	58	6	0.5
500 tweets, 30% gov.	feb	64	6	0.41
500 tweets, 30% gov.	mar	100	5	0.68
500 tweets, 30% gov.	apr	113	5	0.42
500 tweets, 30% gov.	may	83	5	0.37
500 tweets, 30% gov.	jun	75	8	0.53
500 tweets, 30% gov.	YTD	493	5	0.57
500 tweets, 30% gov.	_q1	222	5	0.43
500 tweets, 30% gov.	_q2	271	5	0.66
500 tweets	jan	52	8	0.72
500 tweets	feb	42	9	0.39
500 tweets	mar	106	6	0.77
500 tweets	apr	108	6	0.34
500 tweets	may	87	4	0.5
500 tweets	jun	98	5	0.51
500 tweets	YTD	493	4	0.59
500 tweets	_q1	200	6	0.62
500 tweets	_q2	293	5	0.53
250 tweets	jan	30	7	0.88
250 tweets	feb	22	11	0.7
250 tweets	mar	56	6	0.51
250 tweets	apr	47	5	0.79
250 tweets	may	54	7	0.55
250 tweets	jun	39	6	0.8
250 tweets	YTD	248	5	0.84
250 tweets	_q1	108	6	0.58
250 tweets	_q2	140	6	0.53

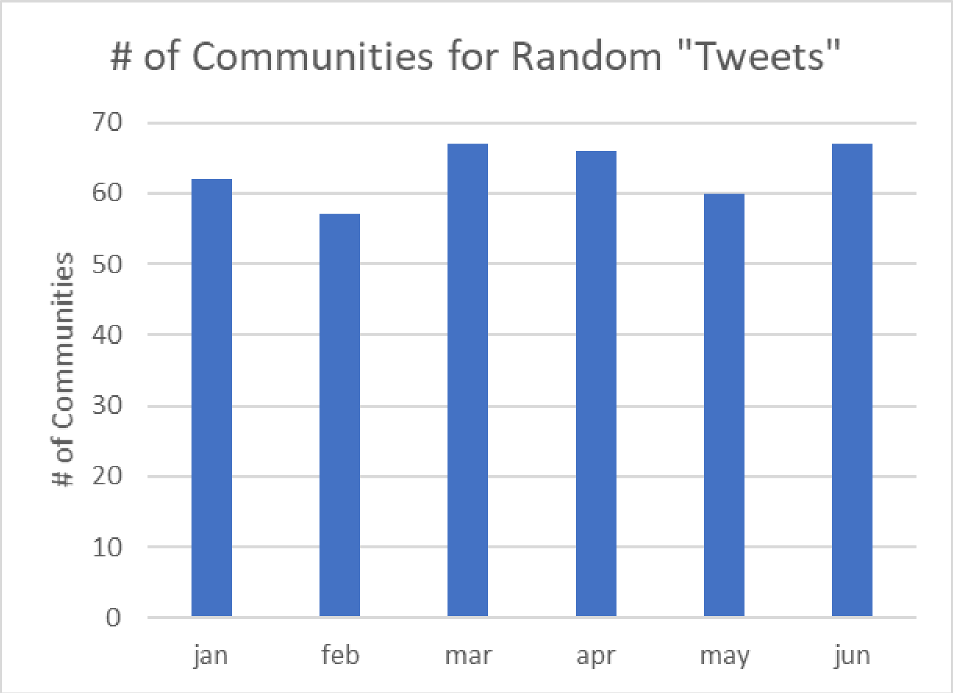
Other Analysis for Comparison

Random English tweets

GPT-2

Random English Words

- Using a list of English words found online [4], generate 600 “tweets” (strings of maximum 280 characters)
- Run analysis on this set of “tweets”, expectation is a very large number of communities due to huge vocabulary compared to common language used by human tweeters.



[4] https://github.com/dwyl/english-words

GPT-2 Trump Generation

- GPT-2 is an algorithm designed by OpenAI to predict the next word (repeatedly) in an English sentence [5].
- Using the *gpt-2-simple* python package [6], finetune the GPT-2 learning model on all collected tweets from President Trump.
- Generate 50 files of 600 tweets each, and run analysis on each file.

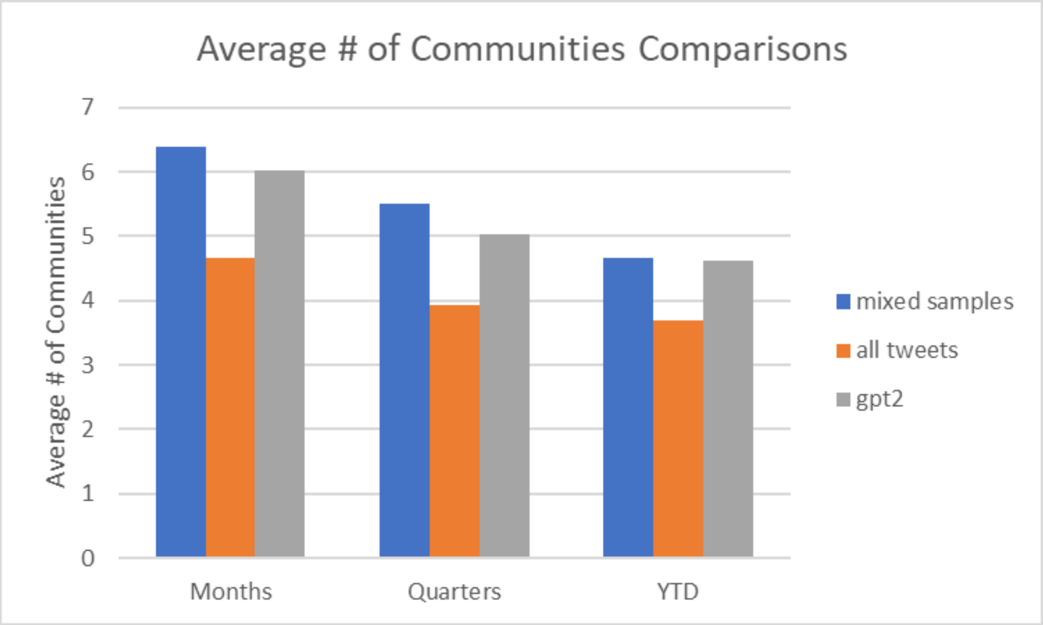
[5] https://openai.com/blog/better-language-models/
[6] https://github.com/minimaxir/gpt-2-simple

GPT-2 Trump – Example Generated Tweets

- (Note that they don’t always make sense, but the vocabulary seems accurate for Trump.)
- *“Venezuelan President Nicolas Maduro on Wednesday accused the United States of trying to overthrow his government and called on other Latin American countries to do the same.*
  We're not going to be bullied!”
- *“A great coach and a great friend. My fond memories of him are the many trips to his house and the many sit-down dinners he set up. He was a true American hero!*
  Thank you @SenatorCory”
- *“Our Republic is not dead. It is on its way back to life. This is a very exciting time in our Nation's history. The Resistance is slowly but surely winning the day. We are winning in every conceivable way.”*

GPT-2 Trump Generation Results

- Generated tweets are generally very high quality in terms of word choice and tone, so expectation is that number of communities is low.
- Results are low and similar to that of mixed sample from all users.



Conclusion

Summary

- Individual users have their own vocabulary and their tweets tend to form 4 communities using Louvain.
- When users tweets are analyzed mixed together, tweets form 5 communities using Louvain.
- GPT-2 does a pretty good job of faking user tweets, but not as focused as individual users, leading to more communities.