COMMUNITY DETECTION ON TWITTER

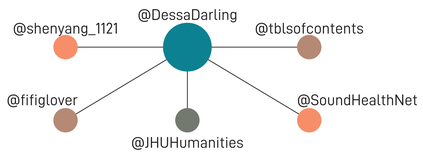
kudsi, m.
stassinopoulos, a.
wang, f.
U.C. SAN DIEGO

Summary: We scraped a large amount of community data on Twitter, starting from a seed account, and analyzed it to see what communities we could find.

BACKGROUND: COMMUNITY DETECTION

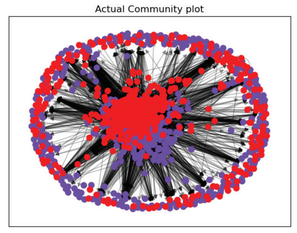
- Recent work shows that the Louvain Algorithm is the most effective at identifying and dividing communities into clusters in large networks.
- Social media networks promote the growth and identification of similarities between groups of people, which can be identified as communities.
- As the number and types of communities grow, identifying and classifying them becomes more challenging.
- The scope of the project is to utilize public data from Twitter, specifically followers and followings of users, to explore communities within the platform.
- The Louvain Algorithm will be used to identify and analyze distinct communities based on a seed account of a modern day rap musician named Dessa (@dessadarling).
- The results of the study will provide insights into the community structure within Twitter.

On the right: A visualization of some users directly connected to Dessa.



OUR QUARTER 1 PROJECT

- For our Q1 Project, we investigated community algorithms.
- We measured performance on a dataset of political blogs.
- Predicted whether a blog was part of community A or community B (left- or right-wing blog).
- Our best results were from the Louvain Method, which was able to predict the community of a given point with 95.9% accuracy.



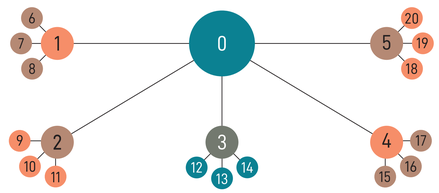
check out our website:

capstone.ari-s.com

HOW WE GOT OUR DATA

- Data was gathered from Twitter using the Followers/Following API through a breadth-first search.
- A "seed" user was chosen, and their followers and users they follow were scraped.
- Mutual followers were identified and added to the graph.
- An edge was added between the seed user and the mutual followers, and the mutual followers were added to the end of the queue.
- Once all of a user's mutual followers were scraped, the user at the front of the queue was processed next.

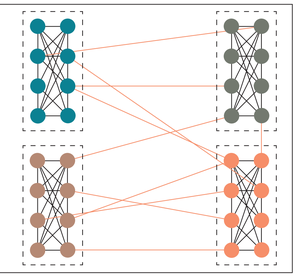
On the right: the order nodes were traversed in our breadth-first traversal.



HOW WE DETECTED COMMUNITIES

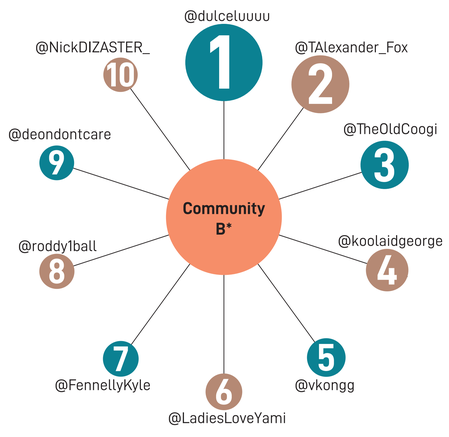
- The Louvain community detection algorithm is based on a modularity approach.
- The algorithm compares the actual number of edges in a community to the expected number of edges in a community.
- The algorithm uses a recursive format to achieve maximum accuracy of community detection.
- The Louvain algorithm was applied to the dataset to identify all communities within it.
- After running the Louvain algorithm, top communities were identified.
- The identified communities were manually researched from their Twitter usernames and bios to identify similarities within their characteristics.

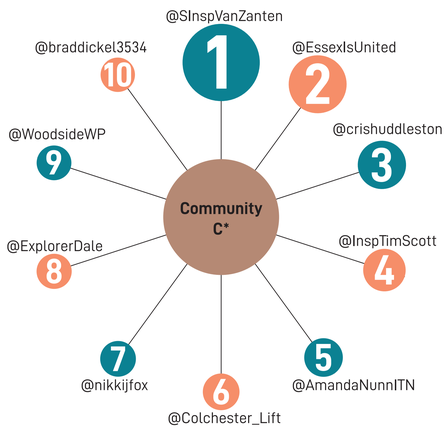
On the right: A visualization of communities detected through modularity; black edges are in-group edges and salmon edges are out-of-group



SAMPLE COMMUNITIES AND DESCRIPTIONS

Below: The three communities we found most interesting, with the top ten most connected users in each





Community A

- A community centered around "geek" hobbies.
- Consists of tabletop games, podcasts, and comics.
- Completely interest-centered, as opposed to professional/geographical.
- Great target audience for shared marketing-type applications.

Community B

- A community of rappers, producers, and other music-adjacent areas.
- Accounts are used primarily in a personal/informal manner.
- Lack of biography information makes them difficult to identify.

Community C

- Based around a county in Eastern England, Essex.
- Consists of various community figures; police inspectors, community centers and the like.
- Geographically centered, covers a variety of people in one area.

SUMMARY OF FINDINGS

- Louvain algorithm successfully identified 10 interesting communities in the Twitter dataset.
- Two distinct features connected accounts in each community: geographical location and interest/profession.
- The findings suggest that undetected communities exist within social media networks and can be accurately uncovered by the Louvain algorithm.
- Future research should implement findings and processes on various social networks to identify possible applicabilities
- Analyzing a social network's API capabilities and identifying creative solutions is important for collecting meaningful data
  - Twitter's API key has limitations that challenged data gathering capabilities.
  - We utilized top 3000 followers/followings of users of each user.
  - We were able to still extract meaningful results with our API limitations.
- Further research and implementation of learnings in the community detection space is essential.
  - Advertisers can target social media users using less sensitive data, benefits both the consumer/account owner and advertisers.
  - Social media account recommendations can be based upon common interests among individuals, in contrast to current mutual followers/following.

mentor: Arya Mazumdar | amazumdar@ucsd.edu group members: Marwan Kudsi | mkudsi@ucsd.edu, Aristotle Stassinopoulos | aristotle@ucsd.edu, Freddy Wang | f5wang@ucsd.edu