Community Detection on Twitter

Kudsi, M; Stassinopoulos, A; Wang, F

March 14, 2023

1 Abstract

Recent work examined the vast unfolding of communities in large networks, in which it was shown that the Louvain Algorithm was the most effective at identifying and dividing communities into clusters. The growth of social media networks in the modern world nurtures the growth and identification of similarities between groups of people. These similarities between groups can be identified more formally as communities. While the number and types of communities grow, the identification and classification of these communities becomes more challenging. To define the scope of the project, we will be utilizing public data from the social media network Twitter. Specifically, we will look at the followers and followings of users throughout twitter. In this paper, we utilize the Louvain Algorithm to explore communities within the social media platform twitter. The results of the study allowed us to uncover and analyze distinct communities based on our seed account of a modern day rap musician named Dessa (@DessaDarling). Further research is needed to determine the potential applications of these algorithms in the field of community detection in social media since we only used one platform's data.

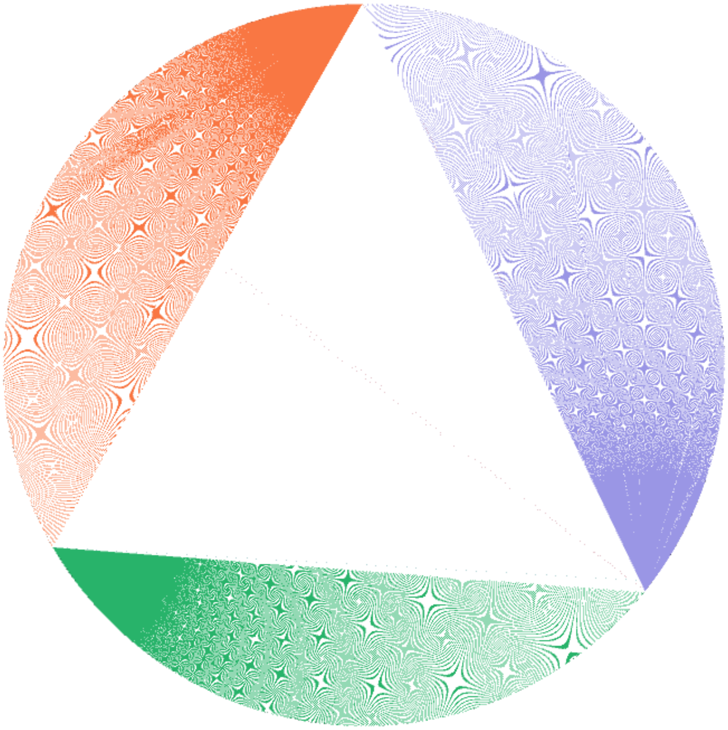

Figure 1: Many of the smaller communities had very high modularity relative to one another.

2 Introduction

Technological innovations during the past few decades, including the rise of computers, the internet, and social media, have accelerated the size and strength of data networks (Kudsi 1). When analyzing the data behind various data networks, communities form naturally within them through connections between individual points of data, or nodes (Kudsi 1). These communities are typically defined by a common variable such as physical location, political alignment, or interest in a public figure (Kudsi 1). However, as more individual nodes of data are added to the data collection, the number of connections between nodes and the number of communities formed to represent these connections grows exponentially, creating difficult problems to overcome when analyzing the data in a timely manner (Kudsi 1).

It's important to note that grouping data has always been a problem that we have been trying to solve, and has been done through clustering algorithms, where using multiple attributes for each data entry can be used to find similarities and differences between them to create "clusters". However, the idea of locating and recovering communities is focused specifically on networks as analysis largely relies on a single attribute type - the edge. This is where the planted clique problem is presented: identifying the subset of nodes in a network that have something in common, all determined by edges. The challenge was constructing an algorithm to do so that could perform in efficient time. Methods to achieve this in polynomial time were introduced in 1995 by Luděk Kučera, and improved upon in 1998 by Alon, Krivelevich and Sudakov (Kudsi 2). Both of which proposed constraints to the size of the planted clique relative to the network, where the planted clique could be found with high probability (Kudsi 2). More recently, the paper "Performance Evaluation of Clustering Algorithms on a Network of Political Blogs" observes that the Louvain algorithm was the most effective at identifying communities and showcasing how a set of nodes with a large number of common neighbors will have a higher probability of being identified as a community, versus a set of nodes with a small number of common neighbors (Kudsi 2).

2.1 How we gathered our data

We gathered our data from Twitter by doing a breadth-first search using the Followers/Following API. We began by picking a "seed" user, described in figure 2 below as node 0. After scraping this user's followers, and the users following them, we picked out mutual followers to add to the graph, represented by different sized and colored nodes in Figure 2. For each of the mutual followers, we added an edge between the seed user and the mutual to the graph data, and added the mutual to the end of the queue of our API data scraper. Once we had scraped all of a user's mutuals[1], we moved on to the user at the front of the queue.

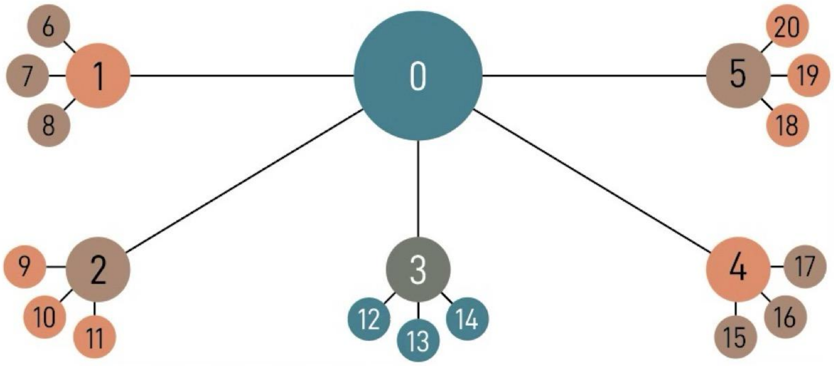

Figure 2: An illustration of the order Twitter users were traversed using our breadth-first search algorithm

[1] We capped each individual user at 3000 followers and 3000 following users retrieved due to Twitter rate limits.

To keep the process as efficient as possible, we also kept a set in memory of all users whose data we had scraped and a set in memory of all users who were in the queue so that we could avoid adding them to the queue a second time or creating duplicate edges.

2.2 Overview of Dataset

The names dataset[1] we collected contains a table of twitter account names and assigned user ids while scraping. These accounts were scraped from the social media website Twitter which houses millions of individuals accounts.

The graphs dataset[2] we curated contains a table of connections between Twitter accounts, scraped from the Twitter API utilizing the described breadth-first-search approach in section 2.2. The dataset consists of two columns representing the link between two twitter accounts that we treat as nodes. The first column 'id_a' contains Twitter user ids, every row in this column we classify as "node a". The second column 'id_b' contains Twitter user ids as well, every row in this column we classify as "node b". Every row in this dataset represents a link between node a and node b (two twitter accounts). The node values can be classified as follows.

Column	**Row Node Classification**	**Row User Classification**
id_a	Node a	User a
id_b	Node b	User b

These recorded links between node a and node b catalog the graph and path of the Louvain algorithm while scraping Twitter and finding links between accounts. Column 'id_a' represents all the nodes (account id's) utilized by the breadth-first-search approach to discover links to new accounts. Column 'id_b' represents all the accounts discovered from the links of corresponding node a from the same row index.

The graph data contains a total of 1,048,576 accounts discovered utilizing the breadth-first-search approach. The number of distinct Nodes A in column id_a are 523,722. Therefore to discover the total 1,048,576 accounts only 523,722 distinct accounts were utilized.

2.3 How we detected communities

The Louvain community detection algorithm is based upon 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 two step process assigns nodes to communities, then iteratively using a modularity approach to evaluate whether moving a node to another community has a positive increase in accuracy. If the algorithm detects a higher accuracy ("gain"), then the node is kept in its new community and the algorithm moves onto the next iterative node. However, if the algorithm detects a lower accuracy ("gain"), then the node is kept in its current community (Louvain).

This process repeats recursively until the accuracy ("gain") of the model is no longer improved per recursive iteration.

We will apply the Louvain algorithm on the dataset, and identify all communities within our dataset. The results will be reported in the paper, along with a discussion of the identified communities and an in-depth analysis of 10 unique communities. We hope that our work will provide insights into the hidden communities within the twitter social media platform, and will help guide future research on community detection within social media platforms.

After running the Louvain algorithm on the dataset and identified possible communities, we manually researched each community's twitter usernames and bios to identify comparisons within their characteristics.

3 Community Analysis

3.1 Community 1 - UK Journalists

Account Handle	**Characteristics**
daniellejourno_	Researcher at BBC Wales
knobblymonsters	Account that posts about British tabloids
charlie_crispy	Research and interviewer at Cosmopolitan UK
OliviaCrellin	Former BBC world editor now editor at Journalism News
Shamaan_SkyNews	Correspondent at Sky News UK
oliviaotigbah	Journalist at BBC News UK
james_lewer	Broadcaster at Sky News UK
ThatDavidHarper	Journalist and Presenter at BBC News UK and LBC News
Sprosto	Journalist and content producer for sports broadcasts in the UK
MeirionTweets	Editor at TBIJ Media News

3.1.1 Community 1 Analysis

This community was centrally focused on Journalism and news professionals in the country United Kingdom. Particularly a majority of the accounts were individuals working at the BBC organization. This is an example of a community that is bounded both by profession and geographical location.

3.2 Community 2 - Medical Professionals

Account Handle	**Characteristics**
conscious_tlab	Cognitive scientist and researcher working at labs in Australia
DrSystemsPsych	Physician and Psychiatrist utilizing data informatics
ty_renshaw	Associate Professor of Psychology at the University of Utah
MaraNievesCabr1	Medical Nuclear HCSC Professor in Spain
DrJamesBooth	Professor of Psychology at Vanderbilt university
jamesrachal3	Chair of department of psychiatry at Atrium health in North Carolina
DulayMario	PHD in Neurology
head_like_egg	Clinical psychologist at Queens University in Canada
NaoTsuchiya	Student studying consciousness in Australia and posting youtube videos on the topic
LauraStroudPhD	PHD and Professor at Brown University Medicine and Director at Miriam Hospital

3.2.1 Community 2 Analysis

This community centers around the medical professionals in the industry and education sector of the medical industry. Specifically, all medical accounts are in the cognitive sector of Psychology and Psychiatry. You also find a mini subcommunity of PHD students. It is interesting that there is only one account that isn't a medical professional or educator which was the account “NaoTsuchiya” that was solely a student and youtube video maker about the cognitive phenomena of consciousness.

3.3 Community 3 - Writers

Account Handle	**Characteristics**
OttoKolbl	Researcher on health issues at the University of Lausanne in Switzerland
Dipesh_Nepal_	Book reviews and literature translator
Rosenkrantz	Hispanic Literature translator
teacup_media	Chinese History Podcast, Media company
DoraMalech	Associate professor of writing at John Hopkins university, editor at The Hopkins Review
Tomwilk0	Historian researcher and PHD at University of Melbourne in Australia
Wm_McKenna	Producer/Shooter/Editor at the BBC World Organization
LucasWMann	Professor at Umass Dartmouth and author of several books
peepaltreepress	Publishing house in the UK for Caribbean and Black British Fiction
AidenHeung	Chinese existentialist poet

3.3.1 Community 3 Analysis

This community has multiple subcommunities which was quite interesting, however overall all accounts were related to writers, reviewers and creators. The first interesting subcommunity was book reviewers from different countries, including one that was also a book translator. The second interesting subcommunity was university professors and researchers from across the world. The last subcommunity was digital media accounts spanning from podcasts to producers at the BBC. This is an example of a community that is bounded solely by profession.

3.4 Community 4 - UK Screenplay Writers

Account Handle	**Characteristics**
SychoticComedy	Stand up comedy and sketches/pictures of comedy
athenastevens	Actress writer and creative entrepreneur that posts youtube videos and book reviews
thesecondshelf	Writer in the UK
Sian_Rowland	Playwriter and Content Writer in London, UK
joecbrownn	Associate Producer of Spongebob musical in London, UK
xymyorkrapper	Cocreater and writer for Sky TV in the UK
KathrynBond	Comedy Writer in London UK
SiBeckwith	Comedian in Newcastle UK
johnharrigan	Award winning writer and director of TV shows in UK
Carolyounghusba	Writer for film, tv and radio in the UK

3.4.1 Community 4 Analysis

This community contained individuals that were writers for either theater productions, television shows, or comedy shows. All the accounts were also based in the United Kingdom. It was interesting how there were three distinct sub communities in the writing of theater, television and comedy. Lastly, there were also 2-3 accounts that were award winning writers which is an interesting sub community within writers. This is an example of a community that is bounded both by profession and geographical location.

3.5 Community 5 - Nature

Account Handle	**Characteristics**
JimBair62221006	International climate activist in Canada
Naturevolve	Magazine showcasing nature and its science within art, based in the UK
Mohamma64508589	River researcher and geopolitical analyst in Bangladesh
EliGreenbaumPhD	Professor and National Geographic Explorer
Anthropolitan_	University of London City Anthropology department blog and magazine about twitter
GenerationCo2	Professor of sustainability at University of Exeter in UK
timechols	Georgia public service commissioner
jmollins	Climate journalist in Canada
GenoWorldview	Media Consultant around the world and nature photographer
d_giovannelli	Microbiologist researching climate extremes in Italy

3.5.1 Community 5 Analysis

This community was centered around nature and climate change. The first largest subcommunity was nature and climate media. The second subcommunity was researchers professional and independent on climate change. This is an example of a community that is bounded solely by profession.

3.6 Community 6 - Leaders

Account Handle	**Characteristics**
ChrisWilko	Cryptocurrency endorser
Abc_brentwood	Insolvency practitioners
CDO_Insights	Digital trends aggregator
MarnieGrundman	Trauma therapist
MendyYButler	Self-care advocate
Sciz	Advisory firm CMO
eProducer	Entertainment investing CEO
SpecsImprov	Improv company
LibDem_News	Liberal/Democrat news aggregator
ShellieDeringer	Homemaker/blogger

3.6.1 Community 6 Analysis

A community of corporate leaders with little to no overlap otherwise. It’s interesting that despite being in very different fields, that they are part of the same community due to simply being leaders in their field. This is an example of a community that is bounded solely by profession.

3.7 Community 7 - Producer/Rappers

Account Handle	**Characteristics**
Dulceluuuu	Unidentifiable
TAlexander_Fox	Music journalist
TheOldCoogi	Hip hop blog owner
Koolaidgeorge	Photographer
Vkongg	Unidentifiable
LadiesLoveYami	Rapper
FennellyKyle	Unidentifiable
Roddy1ball	Dennis Rodman parody account
Deondontcare	Unidentifiable
NickDIZASTER_	Music producer

3.7.1 Community 7 Analysis

A music producer/rapper community, it was difficult to identify many of the accounts and their purpose, likely due to the informal nature of the community. Many accounts were missing their bios, and were operated more like personal accounts rather than a public-facing channel. It's an interesting quirk of the community subculture.

3.8 Community 8 - Political Activists

Account Handle	**Characteristics**
Mahmoudobaida2	Palestinian charity officer
Mingall63	Trade unionist/former Labour leader
Enamhaque31	Manchester doctor
ShareenIdu	Surgeon
PPHRtweets	Human rights organization
SBakerWatch	Anti-Steve Baker
GroomB	Scottish writer/editor
Aaron_Kiely	Political activist
Big007_big	Roman gypsy community figure
JamesPSVine	British political commenter

3.8.1 Community 8 Analysis

A community of political activists, centered geographically around the United Kingdom, and around the Labour party of UK politics. Scattered in the mix are several accounts in adjacent spaces such as human rights and charities. This is an example of a community that is bounded largely by profession, and a small correlation to geographical location.

3.9 Community 9 - Game/Comic Enthusiasts

Account Handle	**Characteristics**
StrayBasilisk	Independent game developers
ThisMightBeAPod	Music fan podcast
Stanagstevlfan	Show fanclub
Zacnaoum	Dungeons and Dragons host
ComicsVerse	Comics/social change podcast
Sarahjeanious	Midwestern LA transplant
Ad_magic	Custom printer for designers
NCootalot	Game character
Witchwordsmith	Game designer
GamePawn	Dungeons and Dragons host

3.9.1 Community 9 Analysis

This seems to be a community centered around tabletop games, comics, and other “geek” type hobbies. These include a huge variety of content producers in the scene such as designers, podcasters, and fanclubs. It’s interesting that despite there not exactly being a clear overlap in some of these hobbies, that we can still identify them as sharing an audience and being in the same community. This is an example of a community that is bounded solely by profession.

3.10 Community 10 - Essex, UK

Handle	**Characteristic**
SInspVanZanten	Essex Police Inspector
EssexIsUnited	Essex county community
Crishuddleston	Human rights organization director
InspTimScott	Police Inspector
AmandaNunnITN	News editor/producer, interest in crime
Colchester_Life	Colchester promotional twitter
Nikkijfox	BBC Health correspondent
ExplorerDale	Geologist
WoodsideWP	Park promotion twitter
Braddickel3534	Essex Police Inspector

3.10.1 Community 10 Analysis

A community centered geographically around Essex, a county in the UK. Interestingly, many of the top account handles seem to be police inspectors. That doesn't necessarily mean that this is a police-centered community however, it could simply be that police inspectors tend to have a larger outreach, and that our usage of a node's degree as the ranking causes them to show up higher. This is an example of a community that is bounded solely by geographical location.

4 Conclusion

4.1 Summary of Findings

After utilizing the louvain algorithm to identify communities on the twitter dataset we have identified 10 interesting communities. Characterizing these communities and analyzing their commonalities we found two distinct features that connected the accounts in each community. The first feature being the geographical location of the twitter accounts, for example community 10 consisted of accounts solely based in Essex county, in the United Kingdom. The second feature was the interest and professions of the twitter accounts, for example community 2 consisted of accounts of individuals in medical professions and PHD medical students. Throughout our analysis we also identified communities that were connected based upon the merging of both geographical location and interest. One distinct example can be found within community 5 which consisted of accounts that were all based in the UK and all had a focus on scriptwriting, ranging from tv shows to theatrical productions.

Overall these findings suggest that undetected communities exist within social media networks and can be uncovered with high accuracy by the Louvain algorithm. It is important to carefully test the performance of the louvain algorithm individually on every social media platform to determine whether these communities can be extracted. More specifically, it is important to utilize sampling and random testing to manually assess extracted communities that may contain accounts that are not correlated. An example of this can be seen in community 7 where 4/10 user accounts had unidentifiable similarities to the community of producers/rappers, showcasing a possible case of low accuracy community detection. As previously stated, further research is required to determine the Louvain algorithm's applicability in the field of community detection within a variety of social media networks that structure their accounts and networks differently.

4.2 Recommendations and Implications for Further Research

Looking to future research, it's essential to take our findings and processes and implement them on various social networks to identify possible applicabilities.

Furthermore, it is recommended to make use of API access on social media networks productively, as we learned throughout our research. Twitter's API key gave us access to only scrape 1 account or 1000 results per request (whichever is less), equating to 15 requests per 15 minutes per api key. With 3 api keys, a minimum of 15 accounts per 15 minutes or up to 45 accounts per 15 minutes depending on a user's followers/following. This limitation challenged our data gathering capabilities with a database that houses millions of users. Through analysis we discovered that by sampling only the top 3000 followers/following of each user, we were able to still extract meaningful results with our mode. Therefore, analyzing a social network's API capabilities and identifying creative solutions to collect meaningful data.

Utilizing our findings it would be interesting to research the application of community detection on advertisement recommendations. Specifically, through our identification of community interests and locations using less sensitive data, social networks can allow advertisers to target accounts accurately without the release of highly sensitive account data that is currently utilized. This benefits the consumer/account owner by keeping their personal data safe, while also giving advertisers similar advertisement turnover accuracy. This is becoming particularly important in a world where more of our private data is shared on a daily basis.

Furthermore, it would be interesting to apply the louvain algorithm's community detection capabilities in group and account recommendations on social media platforms. Currently, on many platforms such as Instagram or facebook, account recommendations are based upon the people you follow and their connections. With the uncovering of hidden communities of common interests, platforms may be able to implement account recommendations based upon common interests among individuals.

We have only surfaced the tip of the iceberg to identify communities within social media platforms, it is now in the hands of the social media network giants and scientific community to both further research and implement learnings in this space.

Appendix

[1] https://github.com/stassinopoulosari/dsc180b-wi23-a15-2-data/blob/main/names.csv

[2] https://github.com/stassinopoulosari/dsc180b-wi23-a15-2-data/blob/main/graph.csv

References

Kudsi, M; Li, Y; Nguyen, J; Rayalu, V; Stassinopoulos, A; Wang, F "Performance Evaluation of Clustering Algorithms on a Network of Political Blogs", 2022

Louvain. Neo4j Graph Data Platform. (n.d.). Retrieved December 4, 2022, from https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/