Analysis of Twitter Users and Communities during the Russia 2018 and Qatar 2022 FIFA World Cups

Benjamin Richmond

Student ID: 710036960

Abstract

This project investigates how Twitter users form communities and interact during the 2018 and 2022 FIFA World Cups. Using datasets of tweets collected throughout both tournaments, user interaction networks were constructed based on retweets and comments. These networks were filtered using K-core decomposition to focus on the most active users, and community detection was performed using the Louvain algorithm. Each community was then analysed using centrality metrics, language metadata, and hashtag usage to identify key users and underlying themes. The results showed that communities often formed around national teams, languages, and media accounts, with some clusters acting as central hubs and key points of interaction across the network, whilst some were niche and isolated. The 2022 networks also contained more unrelated or bot-like communities, reflecting changes in Twitter usage over time. The findings offer a foundation for understanding how global events shapes user interaction on social media.

Declaration

I acknowledge the following uses of GenAI tools in this assessment:

☑ I have used GenAI tools to:

- ☑ develop ideas.
- ☑ assist with research or gathering information.
- ☑ help me understand key theories and concepts.
- ☑ identify trends and themes as part of my data analysis.
- ☑ suggest a plan or structure for my assessment.
- ☑ give me feedback on a draft.
- ☐ generate images, figures or diagrams.

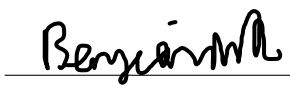
☑ proofread and correct grammar or spelling errors.

☑ generate citations or references.

- ☐ Other: [please specify]

I declare that I have referenced all use of GenAI outputs within my assessment in line with the University referencing guidelines.

I certify that all material in this dissertation which is not my own has been identified.

Signature: 

Contents

1 Introduction

As social networks have grown in popularity, the amount of data users produce has increased, enabling researchers to investigate human online interaction. Among these platforms, Twitter (currently known as X) stands out due to its high velocity and volume of public content and global user base. With approximately 300 million daily users producing over 500 million tweets daily, Twitter provides an excellent opportunity for researchers to observe public reaction and community formation. Its structure, based around hashtags, mentions, replies, and retweets, is especially well-suited for modelling user interactions as networks.

Due to the volume and immediacy of the data, Twitter has been widely used to try to understand how people react to real-world events in real time. An example of real-world events that are ideal to track is major sporting events like the FIFA World Cup, where a large number of global fans interact with each other in response to live matches. This is shown with the 2022 World Cup, which generated 147 billion impressions via the hashtag WC2022 [1]. Prior studies have demonstrated that Twitter activity surges around match events with lots of topic-specific discussions emerging in real time [2, 3]. This scale of participation allows for creating communities that reflect shared interests or opinions.

Despite the large volume of data produced during the World Cup, understanding how users organise themselves into communities and behave online is still complex. As a global platform, Twitter unites hundreds of millions of users, speaking various languages and sharing overlapping interests. However, interaction is not random, with prior research showing that it is often structured by shared interests and topics, leading to the emergence of community clusters [4, 5]. The World Cup is a good example of a truly global event, with 32 participating nations and a vast global viewership, where fans will group around countries and players. This results in the formation of distinct fan communities, a trend observed in previous studies of Twitter during international tournaments [6]. This creates interaction patterns that can be explored using different network science methods.

While major tournaments have been investigated before, with topics such as sentiment analysis and hashtag evolution covered, there is a notable gap in research around the structure of social media networks and the communities within, especially around the 2018 and 2022 World Cups. As Twitter has grown since the past World Cups, there is a lot more data produced during these events compared to previous ones, allowing for a more in-depth analysis of fan behaviour, interaction intensity, and community formation.

This project explores the structure and behaviour of Twitter communities during the 2018 and 2022 FIFA World Cups. The main goals are to understand how these communities form, what the defining theme of each community is, and how these communities interact with others. This will be done by:

- Constructing interaction networks based on user replies and retweets
- Filtering these networks to extract the most engaged users
- Detecting and visualising communities
- Identifying key influential users
- Labelling communities based on key users and other metadata
- Analysing inter-community interaction patterns to reveal bridging groups, isolated clusters, and structural hierarchies

The datasets consist of JSON files containing the tweets made daily over the duration of both World Cups. However, the collection methods differed slightly between tournaments, as shown in Figure 1, with the Qatar dataset collected by tracking common hashtags, countries' flags, and official club accounts. In contrast, the Russia dataset was collected by monitoring common hashtags, and each country's hash flag (a short three-letter hashtag for each country). Each tweet record includes

metadata such as user ID, interaction data related to retweets and replies, timestamp, hashtags, language, and mentioned accounts. This data collection provides a sufficient structure to analyse interaction networks and communities.

Figure 1: The terms used to track tweets for each dataset.

2 Design, Methodology and Implementation

This section details the overall design of the project, the methodological choices that were made to guide the analysis, and the practical implementation of this. It covers pre-processing the datasets, constructing and filtering networks, and extracting user and community features for further analysis. For this, various Python libraries were used to filter and analyse the data, particularly Pandas for data manipulation, alongside Gephi, a graphing software used to analyse and visualise the data.

2.1 Success Criteria

The following criteria were selected to evaluate the success of the following parts of the project:

- **Data Preparation:** If the datasets are converted to PKL without data loss, keeping only the fields wanted, using parallel processing to speed up processing time. This ensures the cleaned datasets are reliable and processable for the following analysis.
- **Network Construction:** If a directed weighted graph is created for both retweets and comments for both tournaments, with the weight representing the number of interactions. This graph should be converted to undirected when in Gephi. This preserves the interaction structure while preparing for community detection.
- **Network Filtering:** If the network nodes are reduced to keep 1% of the most interactive users for all four networks, retaining integrity and structure. This keeps the graphs analysable while reducing noise and allowing analysis in Gephi without computational overload.
- **Community Detection:** If the Louvain algorithm in Gephi produces a set of communities for each network, with a high enough modularity score (above 0.3), to show that the clusters are meaningful and not random.
- **Key User Identification:** If PageRank and the different degree centralities produce a list of influential accounts in all communities, allowing for the identification of the central figures of communities.

- **Community Theme Identification:** If all larger communities (top 5 by number of users), and some smaller communities, can be confidently labelled based on key users, languages and hashtags. This means that communities are meaningfully thematically interpreted.
- **Inter-Community Analysis:** If a directed graph of community-to-community interactions is created to show structure, and the percentages of interactions coming from each community to itself and other communities are produced in a graphic format. This allows the identification of features such as isolated and bridged communities.

2.2 Data Pre-Processing

The dataset for this project was provided by Diogo Pacheco of the University of Exeter, and it consists of tweets collected during the 2018 World Cup in Russia and the 2022 World Cup in Qatar. Initially, the Russia dataset ran from the 11th June to the 24th July, and the Qatar dataset ran from the 10th November to the 2nd January. However, to capture the networks created during the tournament, it was decided to only use days from the first day to the last day of the tournament, cutting the dates for Russia down to the 14th June to the 15th July, and the Qatar dataset down to the 20th November to the 18th December. The reason for this decision was that the aim of the project is to see how these communities form during the World Cup, and days before and after it will only add noise and possibly change the communities.

During initial inspections of the datasets, a similar issue was discovered in both datasets: two days of data had been fused into one, making it appear as if a day was missing. This required splitting these files by analysing the timestamps_ms column and separating them into two new files, both accurately named, allowing for an accurate day-by-day dataset.

Another pressing issue was the size of the files, with many JSON files already over a gigabyte, even though they were already compressed. To combat this, two solutions were drafted: first, cutting out columns that were not relevant to the analysis, and second, converting them to PKL files for faster and more efficient processing. When it came to picking columns to keep for the analysis, a decision was made to select columns that would be needed for this and some that could be required for further study, primarily user identifiers, interaction types, and associated metadata. The columns that were picked are shown in Table 1.

Table 1: The columns that were kept during the conversion to PKL files.

Column	**Description**
`user.id_str`	The ID of the user tweeting
`user.screen_name`	The screen name of the user tweeting
`in_reply_to_user_id_str`	(If a comment) The ID of the user that it is replying to
`in_reply_to_screen_name`	(If a comment) The username of the user that it is replying to
`retweeted_status.user.id_str`	(If a retweet) The ID of the user that is being retweeted
`retweeted_status.user.name`	(If a retweet) The username of the user that is being retweeted
`quoted_status.user.id_str`	(If a quote retweet) The ID of the user that is being quoted
`quoted_status.user.name`	(If a quote retweet) The username of the user that is being quoted
`place_full_name`	The full name of the location of the user
`place_country`	The country code from the user's country
`lang`	Twitter's attempt to classify the language of the tweet
`timestamp_ms`	The timestamp the tweet was posted in ms
`hashtags`	Any Hashtags that are in the tweet

This process was complicated due to the large amounts of data needing to be processed, the JSON files containing lots of nested objects, and data being stored in different columns depending on the tweet length. Two methods were used to deal with the large amount of data needing to be processed when converting. Firstly, to increase processing speed during the conversion of the JSON files, parallelisation was implemented using Python's ProcessPoolExecutor module, which allowed several files to be processed simultaneously. Secondly, due to RAM limitations, each JSON file was processed in chunks of 10,000 to avoid overloading the memory. This value was chosen due to larger chunks causing memory overload. Some data about the actual tweet, such as the hashtags, was stored in a different place depending on the tweet's length, which meant first checking extended_tweet to see if it had data, and if not, moving to entities to get the hashtags.

2.3 Initial Testing

An initial exploratory analysis was conducted to validate and better understand the dataset. This involved creating a dataframe of daily tweet volumes for every country in the World Cup. A user's country was defined using the place_country column. However, most users did not have this data, so only about 2% of tweets were used. However this still left a substantial enough dataset to look for results. Another difficulty in this was dealing with users from the UK, as all users from the UK had their country set as GB (even Northern Ireland), but in the World Cup, countries don't play under the UK, but as their separate countries. This required using place_full_name and parsing the location for the exact country (e.g. Scotland or England). If parsing the place_full_name failed to distinguish the user's home country within the UK, the Python library geopy was used as a fallback geolocation method to identify the country.

Once the datasets were ready with the tweets from all 32 participating nations, a trend was visible. There was a significant increase in tweets from a country on the day they played a game, or when other major matches, such as the final, occurred. Another trend was a drop in tweets after a country was eliminated. To get a better look at this, graphs were generated using the Python library matplotlib to visualise this increase in tweet volume on match days, using the teams with the most tweets and teams that progressed furthest in the tournament. This showed both datasets' relevance and quality for further network analysis. These graphs are shown below in Figure 2 and Figure 3.

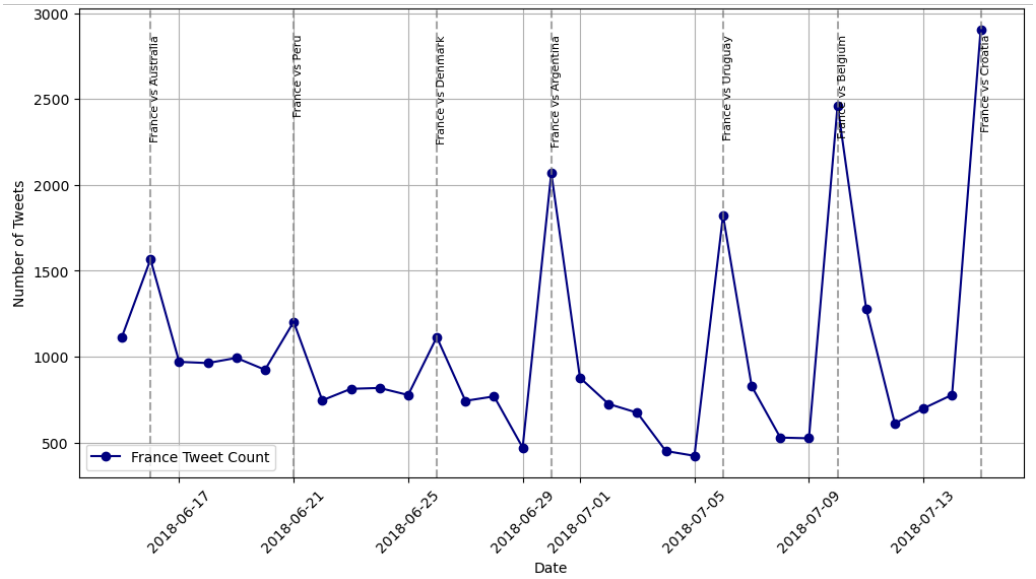


Figure 2: Tweet Progression for France during World Cup 2018

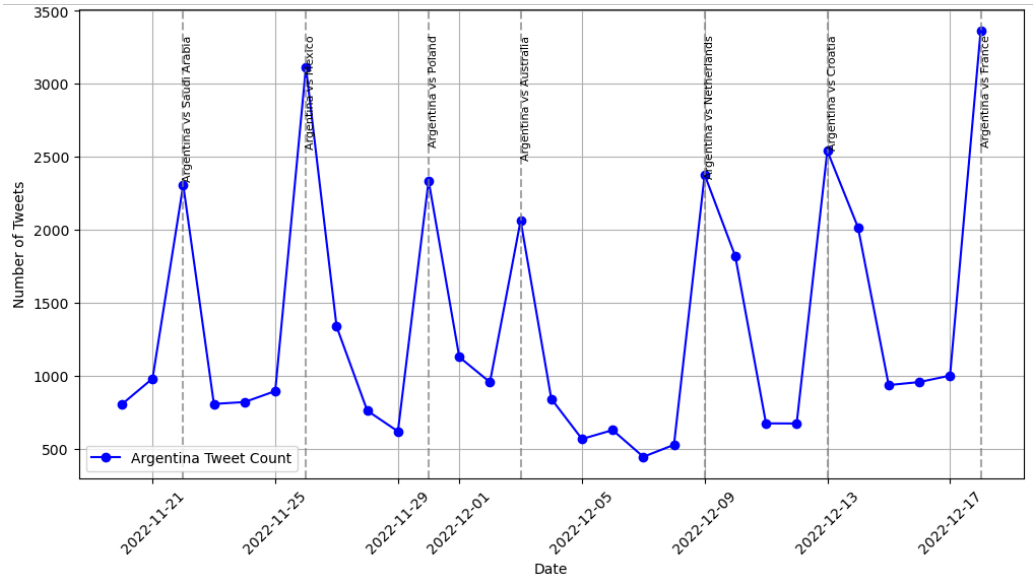


Figure 3: Tweet Progression for Argentina during World Cup 2022

2.4 Network Construction

Four networks were constructed to analyse user interactions during the World Cup, one for retweets and one for comments for each tournament. Each network was modelled as a graph where an edge represents user interaction, and a node represents an individual user. For the retweet networks, an edge was added between two users if one retweeted or quote retweeted the other. For the comment network, an edge was added between two users if one commented on the other's post.

The edges were treated as weighted edges, meaning the weight was incremented each time a connection occurred between two users. This allowed for the network to capture the intensity of interactions between users. While comments and retweets are inherently directional from the user initiating the connection to the user receiving it, the edges were treated as undirected for the visualisation and community detection of the network. This was done to focus on the presence and strength of relationships in the network instead of the direction. However, direction becomes essential later when finding central users in the networks.

To create the networks in Python, the first step was to filter the dataset to include only retweets or comments. After this, the dataset was grouped by the pair of users involved in each interaction, using the initiating user's ID and the receiving user's ID. The number of interactions for each pair was summed, resulting in a weight column recording how many times one user retweeted or commented on another. This step led to a three-column dataframe with the user ID, receiving user ID, and weight. For the actual network, the networkx module was used, and edges were incrementally added to a directed graph, which was then exported to Gephi as a .gexf file. When imported into Gephi, it was then made into an undirected graph.

This led to creating four network graphs, each differing in size. The Qatar networks were larger than their Russian counterparts, and there was a considerably larger count of retweets than comments in both datasets. The final size of the datasets was:

- **Russia 2018 Comment Network** – Nodes: 577,837, Edges: 633,149
- **Qatar 2022 Comment Network** – Nodes: 3,269,163, Edges: 5,099,471
- **Russia 2018 Retweet Network** – Nodes: 7,764,934, Edges: 21,492,946
- **Qatar 2022 Retweet Network** – Nodes: 11,826,682, Edges: 35,741,065

This presented an issue due to the size of the networks. Because all the networks had so many users, with many users barely contributing more than a comment or retweet, it created a lot of noise in the dataset that needed to be filtered out for better analysis.

2.5 Network Filtering

When initially importing into Gephi and attempting community detection, many communities were created, with lots only containing a couple of users stuck in a 1-to-1 relationship with each other. These accounts contributed very little to the structural understanding of the network. It also took a lot of computational power, with runtimes spanning a considerable time, and frequent crashes occurring within Gephi. This made it hard to graph and analyse, so further filtering was required to make the community-level analysis possible.

When addressing this issue, several possible ways of cutting down users were tried. The first approach tried was cutting these large networks down before being imported into Gephi. For this, the method used was filtering by every user's in/out/total degree. This was done by removing users with fewer than a certain number of directed/undirected connections with other users. However, this was flawed because it favoured users who's in/out degrees were significant, including accounts that merely interact with one user repeatedly, and users in a 1-to-1 relationship with another user in which they both interacted frequently, instead of users whose engagements span a wide range of users, which helped build interconnected communities.

After researching different methods of filtering, the one that applied most to the situation was using K-Core decomposition to filter [7], due to it being able to capture a subset of users that have a minimum amount of connections in the network, therefore filtering out users who do not interact with many users. K-Core creates a subgraph of the original network in which every node has at least K connections within that subgraph [8]. This helps create a locally cohesive network, retaining users in tight-knit interaction clusters.

When using K-core decomposition to filter down these users, the aim was to keep 1% of users in each network, an amount high enough to retain network structure but low enough for efficient analysis. The network was loaded into Gephi, and the built-in K-Core filter was applied. Aiming for 1% of users meant experimenting with different values of K using trial and error until this target was met. This led to the following number of edges and nodes within the now trimmed-down networks:

- **Russia 2018 Comment Network** – K-core used: 6, Nodes: 5,903, Edges: 47,648
- **Qatar 2022 Comment Network** – K-core used: 10 Nodes: 32,266, Edges: 511,058
- **Russia 2018 Retweet Network** – K-core used: 24, Nodes: 75,292, Edges: 2,642,305
- **Qatar 2022 Retweet Network** – K-core used: 29, Nodes: 115,597, Edges: 5,106,657

The filtering process resulted in four networks of varying size, making it possible to observe the impact of network size on the number of communities formed.

2.6 Community Detection

After filtering the networks to retain the most connected users, the next step is to understand how users interact by identifying communities within the networks. These communities are users who interact more with each other than with the broader network. In the context of Twitter during the FIFA World Cup, discovering these communities within the dataset allows for understanding how users possibly group around nations, teams, and media.

This identification was performed using Gephi's built-in community detection, which implements the Louvain method, which is widely used for unsupervised community detection in large networks. The Louvain method works by optimising the modularity of communities, with modularity measuring

the density of links inside communities compared to links between them. This method is advantageous as it can discover both large and small clusters within the network [9].

Using Gephi's community detection algorithm, the network was split into communities, with each community assigned an ID, ranging sequentially from 0 upwards. Every user in the network was assigned this ID to show the user's community. Another output was the modularity, which indicated how well each network was partitioned into communities.

To visualise these networks in Gephi, each community was given a colour, and each node was then assigned the colour of this community. These nodes were then arranged using the Force Atlas 2 algorithm in Gephi. Force Atlas 2 is a layout method that is force-directed and specially designed for network visualisation. It is suitable for large social graphs, making it preferable over other layouts. Nodes repel each other like charged particles, while the edges attract the nodes towards each other. The result of this is a graph where nodes with stronger interaction are positioned closer together, while less connected nodes are pushed further apart [10].

Once the data had been visualised, the next step was to analyse it back in Python. This data was exported as a CSV, which was then converted into a dataframe in Python with the user and community for each network. This data frame was then merged with the edges for the respective network, and the user's ID was replaced with each user's username, giving us a new primary dataset with every row being an edge containing two users, their respective communities, and the weight. Unlike in Gephi, this dataset was directed.

After this dataset was created, the next step was to identify the themes of each community. This was done by analysing the key influencers, languages, and hashtags used in each community.

2.7 Identifying Community Themes

Now that the datasets for analysis are constructed, the next step involves identifying key users within each community to gain insight into the potential themes of the generated communities. Communities were expected to form around key themes such as nationality or media affiliation, so the users at the centre of the community are a guide to labelling these communities. To identify central users in the network, four centrality measures were used:

- **Out-degree:** This measures the number of outgoing interactions a user initiates. This is important for finding central users in groups where the main accounts initiate interactions instead of receiving them. The weakness of using this, however, is that accounts that have many of these may be accounts that interact a lot with other users and do not receive anything, therefore not necessarily being a central part of the community.

- **In-Degree:** This measures the number of incoming interactions a user receives. This is important for finding central users in groups where the main accounts receive many interactions, which is more common than the reverse.

- **Total Degree:** This is a combination of a user's in-degree and out-degree and indicates how many interactions a user participates in, either receiving or initiating. It shows how active this user is within the community, whether replying to/retweeting another user or being retweeted/replied to by a user.

- **PageRank:** Provides a measure of influence within the community, where each account has a score, and an initiation from an account with a high PageRank score will give you a higher score than an account with a low PageRank score. This means that accounts that receive lots of interactions from low-influence accounts will not get the same boost as those that receive interactions from high-influence accounts.

These metrics ranked users in each community. Each community was treated as an isolated network, so only interactions from within the community were counted for calculating degrees. As In-Degree and PageRank were the best popularity indicators, showing off users that are interacted with,

instead of users with who could spam interactions, the top 10 for each community were considered, while for Out-Degree and Total Degree, the top 5 of each were considered.

Different techniques in Python were used to calculate these centrality measures. For all three degrees, a new dataframe was created from the primary one, which was filtered down to rows where both users were in the same community. A new dataframe was created from there: one column was the user, and the other was the sum of all weighted incoming interactions, making the In-Degree. Similarly, we created a dataframe in which one column was the interacting user, and the other was the sum of all weighted outgoing interactions, giving us the Out-Degree. These were then merged on the user column, with both degree columns summed for every row, giving us another column with the total degree. To calculate the PageRank of every user, every community was created into a network using the networkx module, which has a built-in PageRank function, which was then used for every network to give each user a PageRank score. These were all added to a dataframe as user, community and PageRank score columns. This dataframe was then merged on the user column with the dataframe containing the degrees, giving a new dataframe with user, community, and a score for all four centrality measures.

The next method used to help identify the theme of networks was to analyse data about the tweets sent by users in each community. The two parts of the tweet that were used were the language of the tweet and any hashtags used by a user in the tweet. For every tweet, Twitter tries to classify the language of the tweet so that it can be translated on Twitter by using machine language detection algorithms [11]. This metadata is formatted as a BCP 47 language tag, a standard code to identify human language. Using this language is incredibly useful for a network where many communities are possibly based around countries. Alongside this, all the hashtags in every tweet are stored in an array, which gives a great indication of the themes of these communities.

To implement these two additions, a new dataframe was created, containing every tweet by a user within a community. This was done by doing an inner join on the CSV file generated by Gephi after community detection, as well as a dataframe containing the user ID, language, and hashtags. Before analysing the dataframe, a step required was converting the language tags into the actual language, which was done using the langcodes library. With the data ready, each community was iterated through, with the top 5 languages used as a percentage, and the top 5 most used hashtags outputted.

Using the lists of top users ranked by PageRank and Degrees for each community and the top 5 hashtags and languages, the communities were manually inspected to assign thematic labels. For each community, several methods were used. When analysing the influential users, the focus was on identifying key types of accounts: official national team accounts, internationally recognised players, major news and sports media outlets. Also, if still available, as a majority were, the accounts were reviewed on Twitter to gauge an idea of the theme of their tweets. Alongside this, the main languages used were important to classify communities as either from one country or language, or being more global, with many languages used. The hashtags were used to try to refine the topic that each community was based on.

Communities were then manually thematically labelled based on the dominant patterns observed among these top users, languages, and hashtags. Due to their small size and lack of information, some communities did not have a clear theme, so they were labelled as niche.

With each community now labelled based on the features of its most central users, the next step involved analysing the patterns of interaction between different communities to allow for a broader understanding of the network's structure.

2.8 Inter-Community Analysis

After identifying and labelling each community's internal themes, the next phase focused more on its external structure or how different communities interact. This analysis aimed to understand the broader structural dynamics of Twitter during the World Cup, such as whether communities are isolated or more a sub-community of a larger group.

- **Internal Interaction Rate:** This is the percentage of interactions from users in a community where the target of the interaction is also in the same community. This gives a good idea of whether communities are isolated or whether they interact more with other communities.
- **Cross-Community Interaction Mapping:** This shows interactions from users in the community, where the target does not reside within the same community. From this data, we can look at which communities are linked with each other, getting a sense of whether there are groups of communities in a larger community.

To do this, the primary dataframe containing both users, their communities, and their weights was used. It was grouped by both communities, leaving a community dataframe with the directed interactions between communities. From this, a directed networkx graph was created with the nodes being the communities, and connections as an edge with the weight. This graph only included interactions between users from different communities, which were filtered out, meaning that it captures external connections. This was then imported into Gephi to be visualised using Force Atlas 2, to get a special idea of the communities.

To explore the balance between internal and external engagement, the community dataframe was made into a matrix showing the percentage of interactions of each community. This was then graphed using Seaborn in two ways to show off the internal and cross-community interaction rates. First, a bar chart was created to show the percentage of users' interactions in each community where the target is in the same community, showing the internal interaction rate. Then, a heatmap of the matrix was created to show the percentage of interactions from each community to every other community, to understand which communities strongly interact with each other, and which ones are more isolated. By looking vertically at this, you gain an idea of which communities receive lots of interactions from different communities, and horizontally, you gain an idea of which communities send lots of interactions to other communities.

3 Results

This section presents the results of the network analysis performed on the 2018 and 2022 FIFA World Cup Datasets. These datasets produced two networks each, one based on Comment interactions and one based on Retweet interactions. For each of the four networks, the following results are presented:

- **Community Structure:** This includes the modularity score, number of communities, and the visualisation of the network layout.
- **Community Analysis:** Identifying the most influential accounts by centrality measures, and analysing the main languages and hashtags used within communities. Assigning labels to define the themes of major communities based on the key users, hashtags, and languages.
- **Inter-Community Interactions:** Examining the level of internal vs external interactions between communities.

Because of the large amount of data and communities, only a summary of significant and interesting communities will be discussed. However, the appendices provide complete extended data, such as detailed community breakdowns.

3.1 Russia 2018 Comment Network

Overview, Modularity and Visualisation

The Russia 2018 comment network, constructed from comment interactions, contained 5903 nodes and 47,648 edges after filtering using K-core decomposition.

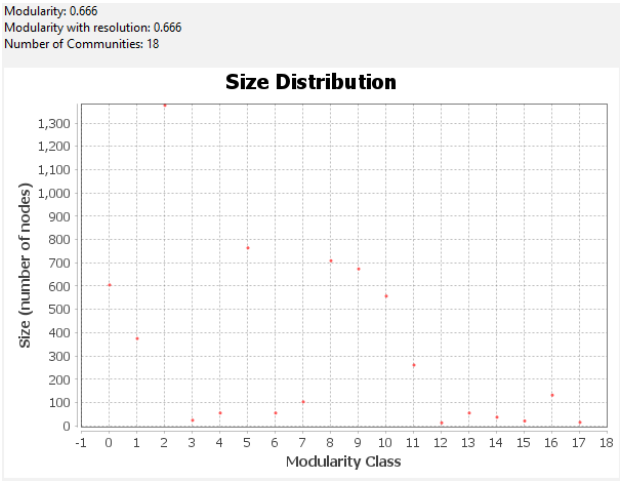


Figure 4: Modularity result showing 18 communities detected with modularity score of 0.666.

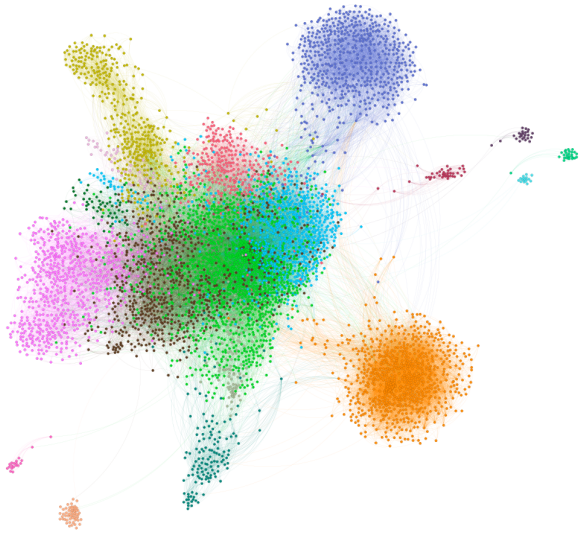


Figure 5: Gephi visualisation of the Russia 2018 comment network with community colouring.

As shown in Figure 4, community detection via the Louvain algorithm gave a modularity score of 0.666, resulting in 18 communities, the largest containing 1441 users and the smallest containing only 17. When visualising with Gephi's ForceAtlas2 layout and colouring by community, the result shows a central hub of communities, with other communities of various sizes surrounding it, as shown in figure 5.

Community Analysis

These are some of the significant and interesting communities generated for the Russia Comment Network, with the theme identified based on key users, languages, and hashtags:

Community 2: The largest community within the network, containing 1441 users. Top users in this included the official @FIFAWorldCup account, @TwitterSports, @brfootball, a popular football content account, and several national team accounts such as Argentina, Nigeria, and Germany. This community is mainly in English (81.09%), with a small amount of German and Spanish. Prominent hashtags included #WorldCup and hashflags for England, France, Croatia, and Belgium, the four semi-finalists in the tournament. Based on this data, this community is likely a broad global fanbase containing main media accounts, primarily in English, discussing the World Cup throughout the tournament.

Community 5: A large community with 797 users. Top users include @BergerPaintsInd, @TecnoMobileInd, @OnePlus_IN, and @Vivo_India, all Indian technology brands. This is a primarily English-speaking community(92.86%), with popular hashtags relating to teams that did well in the World Cup and #WinWithTecno. Based on this data, this community is likely full of Indian commercial brands, likely ones that do giveaways related to the World Cup.

Community 8: This is another large community with 720 users. Top users include @BBCMOTD and @BBCSport, English ex-players and pundits @GaryLineker, @AlanShearer, and @IanWright0. This is a primarily English-speaking community (93.61%) with top hashtags including #ENG and #ThreeLions. Based on this data, this is the fanbase of the English national team, containing English media and pundits.

Community 9: A community containing 692 users. Top users include accounts of national teams participating in the tournament, such as England, Brazil, France, and Belgium, as well as famous players such as Cristiano Ronaldo, Harry Kane, Kylian Mbappe, and Eden Hazard. A variety of languages are spoken in this community with half the tweets in English, a quarter in Spanish,

and fifteen percent in French. Top hashtags include #WorldCup, #FIFA, and #Rusia2018 (Spanish spelling). This data implies that this community is a global football group with no specific team with much support, based around national teams and key players.

Community 0: A community containing 620 users. Key users in this include Spanish content creator @2010MisterChip, the Spanish FIFA World Cup account, @FOXSportARG, and the Mexican and Peruvian national teams. The language used within this community is Spanish(87.14%), and top hashtags include #Rusia2018 and the hashflags for Mexico, France and Croatia. Using this data, we can infer that this community is likely full of Spanish-speaking people, mainly from Latin American countries such as Peru, Mexico, and Argentina, who are following the World Cup. There is more of an aspect of following it due to the appearance of France and Croatia, the finalists, in the hashtags. However, support for Mexico is also there due to the hashtag and appearance of the national team account.

Other Communities: Communities based around national teams or countries included: Community 10 (562 users), focused on English commercial brands and giveaways; Community 1 (384), French national team; Community 16 (138), Nigerian national team; Community 7 (114), South African fans and brand engagement; Community 4 (69), Turkish betting accounts; Community 13 (62), Brazilian national team; Community 6 (60), Saudi Arabian national team; and Community 14 (41), Spanish national team. Other identified communities included: Community 11 (268), focused on global news and politics; Community 3 (28), centred around Brazilian artist Marília Mendonça; and Community 12 (17), based on a comment thread of "Anipal" (pet-themed) accounts discussing the World Cup.

Inter-Community Interaction

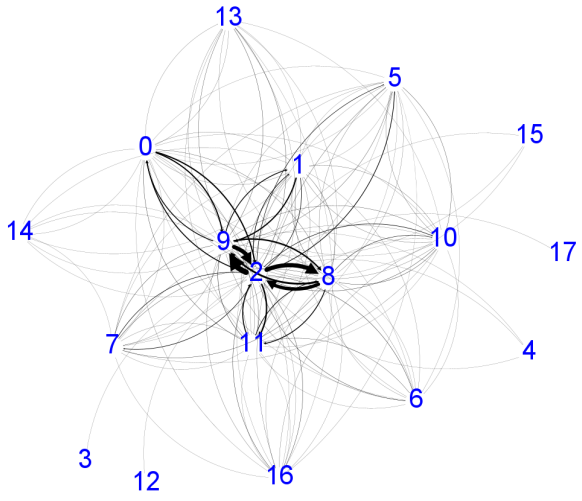


Figure 6: Gephi layout of the Russia 2018 comment network, showing the amount of interactions between communities.

Comm No.	IIR %
0	83.84
1	86.69
2	72.27
3	99.93
4	99.72
5	98.71
6	82.38
7	83.01
8	63.03
9	65.69
10	95.14
11	72.99
12	99.60
13	74.74
14	96.61
15	99.19
16	87.83
17	99.33

Table 2: The Internal-interaction rates (IIR %) for each Community (Comm No.)

Internal cohesion varied throughout the network, with communities displaying different levels of internal and external interactions. A shown in Table 2, nearly half the communities had internal interaction rates of over 90%, with 3, 4, 12, 15, and 17 all having rates over 99%. These groups functioned as isolated clusters, typically smaller in size and often centred on niche topics or content unrelated to the core World Cup discourse.

In contrast to these high interaction rates, several communities such as 2, 8 and 9 showed lower

levels of internal interaction. Community 2, the largest in the network and composed of a broad global football audience, had an internal interaction rate of 72.27%, but received many interactions from other networks, with it being the highest receiver for all but two of the other communities. Similarly, Community 9 which was a global football fan base containing players and team accounts, had an interaction rate of 65.69%, however, it was among the top three most interacted-with communities for 13 other groups. This shows that these two communities are central hubs for tournament discourse in the network. Community 8, composed largely of UK football media and pundits, also showed a relatively low internal rate of 63.03%, indicating its role in bridging discussions between British fans and broader tournament narratives.

The ForceAtlas2 network layout (Figure 6) reflects these dynamics, with Community 2 positioned centrally and visibly linked to numerous surrounding communities, while more isolated groups appear on the network's periphery with few outgoing connections.

3.2 Russia 2018 Retweet Network

Overview, Modularity and Visualisation

The Russia 2018 retweet network, constructed from retweet interactions, consisted of 75,292 nodes and 2,642,305 edges following filtering through K-core decomposition.

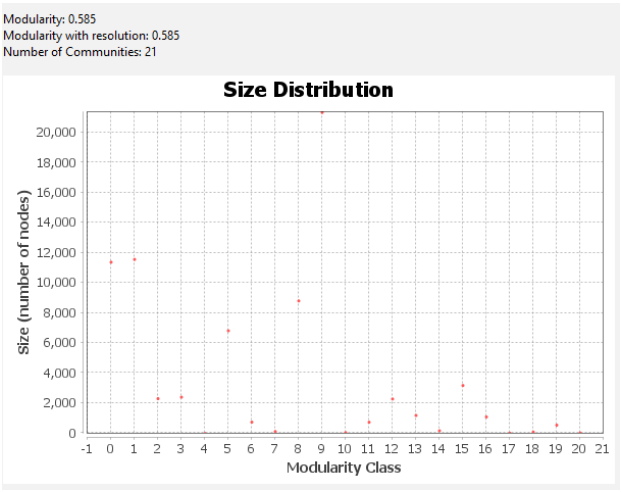


Figure 7: Modularity result showing 21 communities with a modularity score of 0.585.

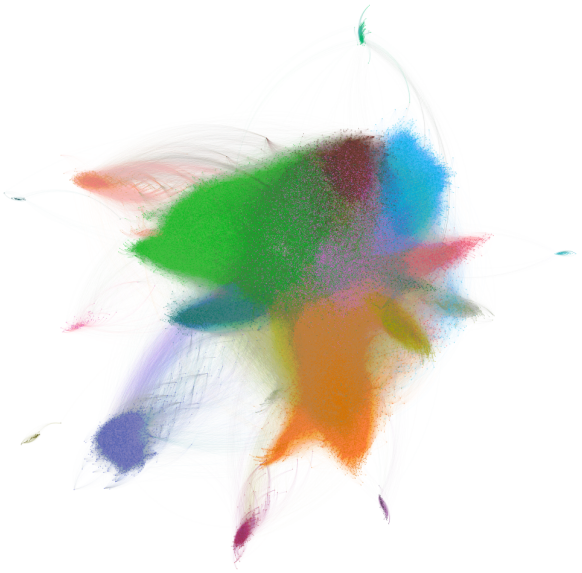


Figure 8: Gephi visualisation of the Russia 2018 retweet network with community colouring.

As shown in Figure 7, community detection via the Louvain algorithm gave a modularity score of 0.585, resulting in 21 different communities being formed, the largest containing 36,026 users and the smallest containing only 63. When visualising with Gephi's ForceAtlas2 layout and colouring by community, as shown in Figure 8, the result shows a central hub of with a large number of communities, with other communities that look smaller outside this hub.

Community Analysis

These are some of the most interesting and influential communities identified within the Russia Retweet Network, based on key users, hashtag trends, and dominant languages:

Community 9: The largest community within the network, containing 22,342 users. Top users include several @OptaJoe accounts in English, Dutch, Portuguese, Spanish, and two @Squawka accounts. These are all stats and analytics accounts which post fun facts throughout the tournament.

There are also several English Premier League football teams. This community is overwhelmingly English-speaking (92.83%), with a small amount of Indonesian and Spanish; the prominent hashtags are hashflags for France, England, Belgium and Nigeria. Based on this information, this community could have a global fanbase of the World Cup, which contains popular media accounts that would discuss the World Cup throughout the tournament.

Community 1: A large community containing 12,132 users. Top users include @equipedefrance (the French national team account), @fifaworldcup_fr (the French World Cup account), and @Opta-Jean, the French version of Opta. This is a mainly French-speaking community (88.49%), with some minor representation of Spanish and Arabic, and the hashtags include #CM2018 (Coupe de Monde – World Cup in French) and the hashflags for France, Belgium, and Senegal (all French-speaking countries). This shows that this cluster is French-speaking, primarily supporting the French national team alongside other French-speaking countries.

Community 0: A community containing 12,014 users. Top users include the @FIFAWorldCup account, national team accounts for Belgium (@BelRedRevils), Germany (@DFB_Team), and Croatia (@HNS_CFF), and several content accounts based on football like @WorldCupUpdates. This is a primarily English-speaking community (72.41%) with hints of Spanish, French, German and Portuguese. The prominent hashtags contain the hashflags for the four semi-finalists, concluding that this is likely a global fan base based around following the World Cup.

Community 8: This community contains 9,045 users. Key users include @fifaworldcup_es (the Spanish FIFA World Cup account) and several Spanish-speaking content accounts such as @2010MisterChip. There is also a presence of Mexican media, with @ESPNmx and @FOXSportsMX. The dominant language is Spanish (83.13%), with some minor use of English, Portuguese, and Catalan. Top hashtags include #Rusia2018 and hashflags for Mexico, France, and Croatia. This indicates a Spanish-speaking fanbase, primarily from Latin America, engaging with the tournament from a Spanish-language media and commentary perspective.

Community 5: This community contains 7,105 users. The top users in the cluster are all British media accounts like @BBCSport and @itvfootball, along with the @England national team account and Harry Kane (@HKane), a popular English player. This is a majority English-speaking community (91.8%), with the popular hashtags related to the English national team (#ENG and #ThreeLions). Based on the above, this community is likely around English football fans supporting the English national team, with key users being media and pundits.

Other Communities: Communities based around national teams or countries included: Community 15 (size 2471), Japanese national team support; Community 2 (2454), Arabic fanbase; Community 12 (2428), Thai fans; Community 13 (1233), Russian coverage; Community 16 (1127), British brands; Community 6 (828), Portuguese national team support; Community 11 (788), Indonesian fans; Community 7 (148), Turkish sports betting; Community 18 (107), Nigerian political messaging. Other identified communities included: Community 15 (3363), Fans of the K-pop group EXO.

Inter-Community Interaction

Communities within the Russia 2018 retweet network showed various internal and external interaction patterns. As shown in Table 3, of the 21 communities, 8 had internal reaction rates above 90%, with communities 10, 15, 17, 19 and 20 showing rates of over 97%. These isolated groups were usually small without easily identifiable topics, except for 15, the large community based around the K-pop group Exo.

In contrast, several large and structurally central communities had significantly lower internal interaction rates. Community 0, the largest group in the network, had an internal interaction rate of just 59.03%, with substantial inbound and outbound activity, showing how it is used as a bridge in the middle of the network. Similarly, Community 5, consisting mainly of UK-based sports media and official team accounts, had an internal rate of 52.89% and interacted heavily with Communities 9 and 0. The Indonesian fanbase in Community 11 also exhibited low cohesion (57.12%) and had a large number of interactions with Communities 0 and 9.

Similarly to Community 0, Community 9 played a central role in tournament discourse, with an

internal rate of 72.50%, but lots of inbound activity. It featured high-profile player accounts and statistics pages and appeared in many other communities' top three interaction destinations.

These dynamics are shown below in Figure 9, where communities 0 and 8 are at the centre of the network with many inbound and outbound connections, while more niche and isolated communities are on the edge with few connections.

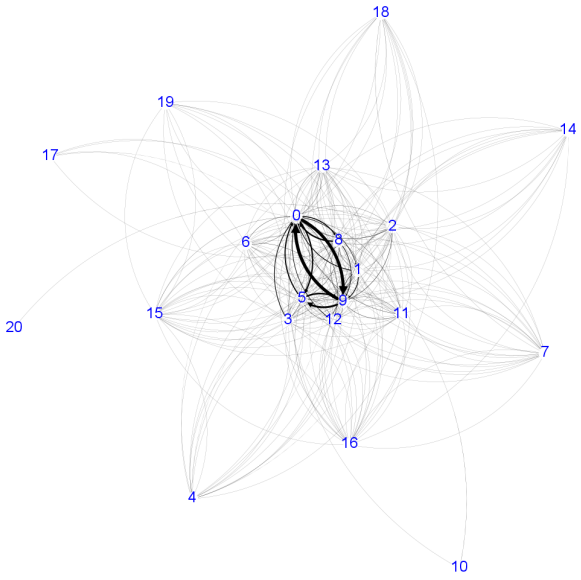


Figure 9: Gephi layout of the Russia 2018 retweet network, showing the amount of interactions between communities.

Comm No.	IIR %	Comm No.	IIR %
0	59.03	10	99.89
1	88.28	11	57.12
2	82.75	12	63.78
3	82.52	13	85.22
4	72.10	14	98.95
5	52.89	15	97.49
6	70.73	16	91.05
7	84.28	17	99.38
8	84.03	18	94.56
9	72.50	19	97.90
		20	99.87

Table 3: The Internal-interaction rates (IIR %) for each community (Comm No.) in the Russia 2018 retweet network.

3.3 Qatar 2022 Comment Network

Overview, Modularity and Visualisation

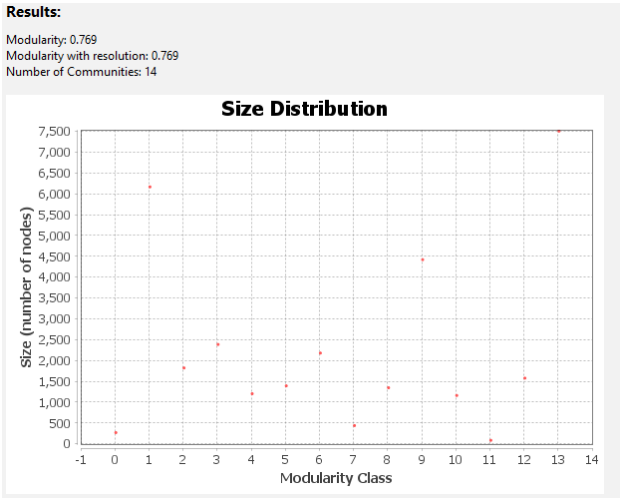


Figure 10: Modularity result showing 14 communities detected with modularity score of 0.769.

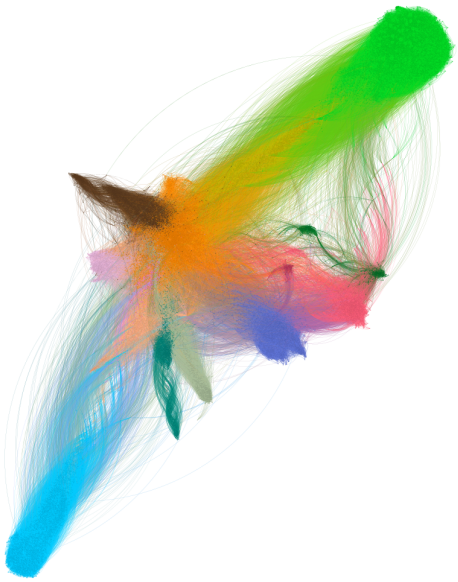


Figure 11: Gephi visualisation of the Qatar 2022 comment network with community colouring.

The Qatar 2022 comment network was constructed from reply-based interactions during the tournament. After filtering via K-core decomposition, the final network contained 32,266 nodes and 511,058 edges.

Community detection using the Louvain algorithm revealed a highly modular structure, achieving a modularity score of 0.769 and identifying 14 communities. As shown in Figure 10, the largest community contained 7,874 users, while the smallest had 108 users. The Gephi visualisation in Figure 11, produced using the ForceAtlas2 layout and coloured by community ID, shows a striking linear alignment of communities, with multiple mid-sized clusters connected along a central spine.

Community Analysis

The following communities represent a selection of the most prominent and recognisable groups within the Qatar Comment Network, identified through central users, language, and topic:

Community 13: This was the largest community in the network, consisting of 7,874 users. Top accounts include brands such as @MobilyPay, @Marsalqatar, @AppMrsool, and @ZainKSA. The language is overwhelmingly Arabic (88.77%), with secondary use of Japanese and English. Prominent hashtags included #FIFAWorldCup and Arabic equivalents of "World Cup" and "Qatar 2022". Based on this, the community appears to represent Arabic-speaking commercial accounts and brands, likely centred around World Cup promotions and Saudi Arabian team support.

Community 1: This large community of 6,483 users included many official national team accounts such as England, France, Croatia, Australia, Brazil, Netherlands, Portugal, and Morocco. It also included other large accounts such as the official @FIFAWorldCup account, and @Cristiano (Ronaldo). It was primarily English-speaking (73.97%), with minor French and Spanish representation. Top hashtags include #Qatar2022, #FIFAWorldCup, and #WorldCup2022. Given the accounts and language profile, this community likely represents a general international audience following the tournament, especially English-speaking fans and official channels from major national teams.

Community 9: This is a large community with 4,637 users, with top users being Brazilian political accounts such as @jairbolsonaro, @pauloguedesfc, and @michellefcs2. The main language in this community was Portuguese (89.85%), with a slight amount of Spanish and English. The most common hashtags were world cup themed one such as #Qatar2022 and #FIFAWorldCup, and political one such as #SOSFFAASalvemOBrasil (A hashtag aimed around calling up the military to 'save the country'. This is likely a far-right Brazilian political group based around the president at the time Jair Bolsenaro, who left his role right after the World Cup after being voted out in October 2022. The use of the hashtags indicates a desire for the military to intervene due to the election results not going Bolsonaro's way and an attempt to latch onto the world cup hashtags to raise interactions.

Community 3: This medium-sized community with 2,518 users had top users like the official Spanish World Cup account @fifaworldcup_es, and several national team accounts such as Argentina, Mexico and Uruguay and Spain. Language use was Spanish (85.47%) with minor English and Portuguese. Popular hashtags included #Argentina, #Qatar2022, and #ESPNQatarEnStarPlus. The data suggests this community consists of Spanish-speaking, primarily Latin American fans, particularly those supporting Argentina, Mexico, and Uruguay, with a subsection of Spanish fans as well.

Community 6: This smaller community with 2,292 users was dominated by cryptocurrency accounts, such as @cryptojack, @binance, and @AltCryptoGems. The language was primarily English (96.81%). Hashtags included tournament related ones like #FIFAWorldCup, but also crypto-specific ones such as #Crypto, #BSC, and #bnb. This community clearly represents crypto enthusiasts and promotional accounts, using the World Cup as a trending topic to boost visibility.

Other Communities: Communities based around national teams or countries included: Community 2 (1900 users), Saudi Arabia national team fanbase; Community 8 (1534), US and Iranian fans with protests around their match; Community 7 (467), Indian and South African brands; Community 0 (313), Brazilian national team fanbase. Other identified communities included: Community 12 (1758), Crypto giveaways; Community 5 (1467), Fanbase of K-pop star Jungkook and his World Cup song Dreamers; Community 4 (1258), US political commentary with a right-wing element.

Inter-Community Interaction

Internal and external interaction rates across the Qatar 2022 comment network varied considerably between communities. As shown in Table 4, half of the communities in this network had an internal interaction rate of over 90%, including communities 5, 9, and 13, which had a rate over 97%. These were generally communities where the sole focus was not the World Cup; for example, Community 9 was Brazilian politics, and Community 5 was fans of K-Pop star Jungkook.

In contrast, some of the more football-focused communities exhibited lower internal interaction rates, with communities 0 and 3, the Brazilian and Spanish-speaking fanbases, featuring in the lowest three rates and initiating lots of interactions with Community 1. The central point within the community was Community 1, with a high self-interaction rate of 85.66 %. This community is a global audience of football fans watching the World Cup, containing International Teams. This cluster was the most interacted with community from every other football-based community, showing it as a bridge for users worldwide to talk about the World Cup.

These dynamics are reflected in the ForceAtlas2 layout (Figure 12), where Community 1 occupies a central position with numerous outward links. In contrast, highly self-contained communities such as 5, 11, and 7 appear on the network's edge, reflecting their minimal external engagement.

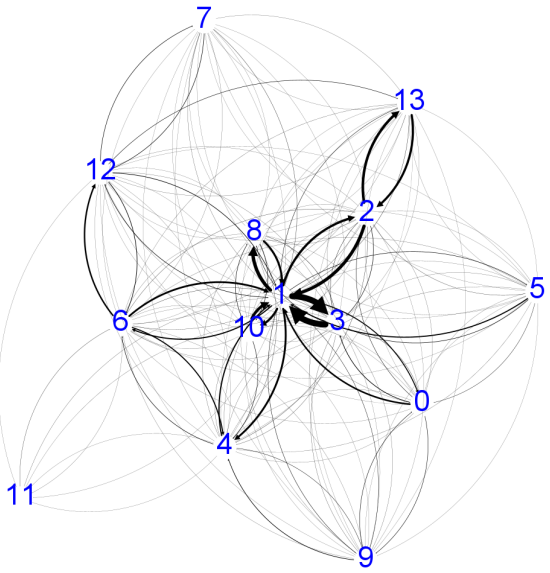


Figure 12: Gephi layout of the Qatar 2022 comment network, showing the amount of interactions between communities.

Comm No.	IIR %
0	67.07
1	85.66
2	81.95
3	80.80
4	83.46
5	97.62
6	93.69
7	95.80
8	77.79
9	98.53
10	84.82
11	95.17
12	91.57
13	97.96

Table 4: The Internal-interaction rates (IIR %) for each Community (Comm No.)

3.4 Qatar 2022 Retweet Network

Overview, Modularity and Visualisation

The Qatar 2022 retweet network was constructed from retweet interactions between users over the course of the tournament. After filtering using K-core decomposition, the resulting network contained 115,597 nodes and 5,106,657 edges, representing the largest network in this study.

Community detection resulted in a modularity score of 0.769, indicating a well-clustered structure with strong internal connectivity. As shown in Figure 13, 37 communities were identified, with the largest containing 24,924 users and the smallest just two. Figure 14 highlights the diversity and scale of the network's structure, with a lot of large communities with some overlap, but less of a clear central part when compared to the others.

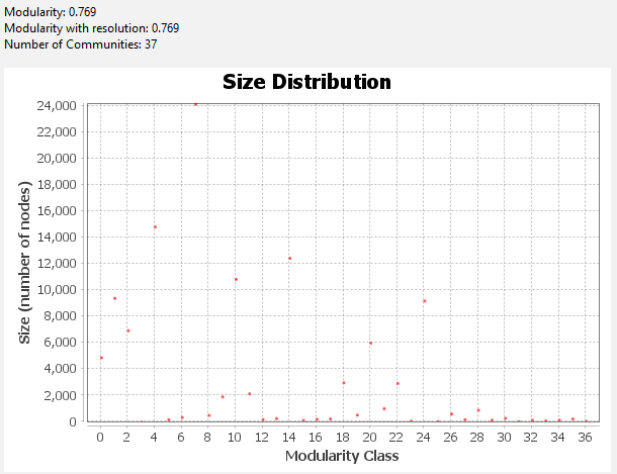


Figure 13: Modularity result showing 37 communities detected with modularity score of 0.769.

Figure 14: Gephi visualisation of the Qatar 2022 retweet network with community colouring.

Community Analysis

Below are examples of the most significant and unified communities from the Qatar Retweet Network, with themes inferred from key accounts, common hashtags, and language use:

Community 7: This was the largest community in the network, containing 24,924 users. Top users include official FIFA-affiliated accounts such as @FIFAWorldCup, @FIFAWWC, and @FIFAcom, online content creators such as @FabrizioRomano. This group is primarily English-speaking (81.32%), and top hashtags such as #Qatar2022 and #WorldCup relate to general tournament coverage, along with #BlackStars and #TeamGhana. Given the dominance of official accounts and general content distribution, this community likely represents the central media hub for the World Cup, distributing official highlights, scores, and tournament updates to a global audience.

Community 24: A large community of 21,569 users. Leading accounts include @FootballApeFC, @FomoSportsnft, and @bitsportgaming, which are associated with NFTs and blockchain-based football content. The language is primarily English (95.22%), and key hashtags are all related to the World Cup such as #fifa and #World Cup. The presence of terms to do with NFTs in usernames strongly suggests this group was using World Cup visibility to promote crypto projects. Based on this, the community likely represents a Web3/NFT promotional cluster, using World Cup hashtags to increase exposure.

Community 4: This community contained 15,178 users and was mostly composed of Portuguese-speaking users (89.12%). The top accounts included @vozdopovobrazil, @NikoIasFerreira, and @maicon-sulivanbr, all associated with Brazilian nationalist or right-wing political content. Hashtags mixed tournament coverage (#FIFAWorldCup, #Qatar2022) with political messages like #BrazilianSpring and #SOSFFAA. Based on the user and hashtag patterns, this is likely a far-right Brazilian political community using the World Cup to amplify their messaging, particularly around military involvement and post-election unrest in late 2022.

Community 14: This was a mid-sized community with 12,715 users. The top accounts were Spanish-speaking media and official football accounts, including @fifaworldcup_es, and national teams from Spain, Mexico and Argentina. The dominant language was Spanish (82.69%), and the most used hashtags included #ARG, #Qatar2022, and #CopaMundialFIFA. Based on this, the community likely represents Spanish-speaking Latin American fans engaging with official content and media, with a clear large group focused on the tournament winners Argentina.

Community 10: This community contained 11,281 users, with top accounts including official accounts like @fifaworldcup_pt, @CBF_Futebol (Brazilian national team), and several major BTS fan accounts such as @BTSChartsDailyx and @btschartsnews. It was primarily English-speaking (61.58%), with some Portuguese, Spanish, Korean, and Japanese. Hashtags like #Jungkook, #Dreamers2022, and #FIFAWorldCup suggest this group represents a crossover between Brazilian football fans and BTS supporters, centred around Jungkook's World Cup performance.

Other Communities: Communities based around national teams or countries included: Community 1 (9,753 users), French national team support; Community 0 (5012), Saudi national team support; Community 9 (1959), Japanese national team support; Community 21 (1090), Nigerian betting; Community 26 (649), Indonesian fans; Community 6 (367), Polish national team support. Other identified communities included: Community 2 (7325), Iranian political activism; Community 11 (2177), French far right politics; Community 19 (564), Tigray war.

Inter-Community Interaction

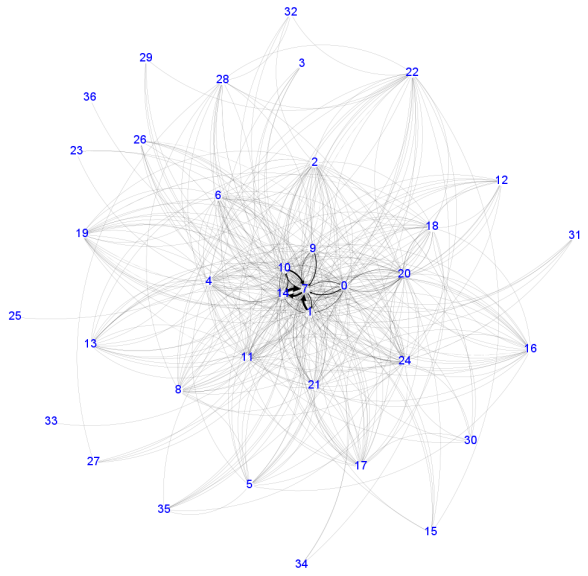


Figure 15: Gephi layout of the Qatar 2022 retweet network, showing the amount of interactions between communities.

Comm No.	IIR %
0	80.60
1	82.16
2	99.15
3	88.19
4	98.75
5	89.83
6	79.96
7	85.87
8	91.91
9	82.11
10	88.48
11	90.15
12	95.96
13	93.03
14	82.69
15	99.61
16	94.71
17	85.83

Comm No.	IIR %
18	96.55
19	99.53
20	92.79
21	89.11
22	99.88
23	99.29
24	98.21
25	99.99
26	99.64
27	99.98
28	99.71
29	99.93
30	95.42
31	99.64
32	99.79
33	99.99
34	99.87
35	98.24
36	99.95

Table 5: The Internal-interaction rates (IIR %) for each Community (Comm No.)

The Qatar 2022 retweet network showed a wide range of internal cohesion across its 37 communities, as shown in Table 5. Unlike the other networks, all 37 communities in this network have an internal interaction rate of over 75%, with the lowest being community 6 with 79.96%. This shows that this network has a lot of communities with high cohesion. 27 out of the 37 communities have internal interaction rates of over 90%, with 14 being over 99%, meaning a nearly fully isolated community. These communities were commonly more complex to identify, with no large recognisable users.

In contrast, several large or prominent communities displayed much lower internal interaction rates, suggesting outward-facing engagement. Community 7, the largest in the network and based around a global World Cup fanbase, had a self-interaction rate of 85.87%, directing a proportion of its interactions to Communities 14 and 1, other football communities, but focused on different languages. Community 1, composed mainly of French-language media outlets and national team coverage, similarly interacted widely with Communities 7 and 14, with a moderately high internal rate of 82.16

These structural roles are visible in the ForceAtlas2 layout (Figure 15), where Communities 7, 1, 14, 10, 9, and 0 form a dense central core connected to many surrounding clusters. These communities

are clearly connected to the World Cup, most linked to a specific country or language. Meanwhile, communities with high internal rates, such as Communities 20, 24, and 26, occupy peripheral regions of the graph with minimal external links.

4 Discussion

4.1 Discussion of Results

This project investigated how Twitter users clustered into communities and interacted during the 2018 and 2022 FIFA World Cups. By analysing interaction networks based on retweets and comments, distinct structural and behavioural patterns were observed across the two tournament years and interaction types. This discussion critically examines the significance of these patterns, reflecting on their broader implications.

A key difference between networks was the size of the networks generated. Comment networks had far fewer nodes and edges, but higher interaction per user, averaging just over one comment, compared to around three retweets per user. This suggests that retweeting played a far more prominent role in how users engaged with World Cup content, showing its use a tool for amplification and reach.

A consistent finding across all four networks was the emergence of thematically coherent communities. These were largely structured around national teams, language groups, media accounts, and high-profile players. This reflects how online spaces tend to mirror offline identities, particularly during globally unifying events like the World Cup. Language appeared to be a strong factor in how communities formed, with clear examples of Arabic, Spanish, and Portuguese-speaking clusters, especially in the comment networks.

A noticeable difference between the two World Cups was the rise in communities that had little or no connection to the tournament itself, which were more prominent in the Qatar 2022 datasets. This could be an example of how Twitter has evolved in the four years between 2018 and 2022, with a larger section of users now using generic popular hashtags, which in this case are related to the World Cup. This could also show how the amount of bot users has increased, with many internally active but themeless communities, lacking key users or hashtags.

Visualisations using the ForceAtlas2 layout support these findings. In each case, communities with high interactivity and broader relevance occupied the centre of the graph, while isolated or self-contained groups appeared on the outer edges. This spatial arrangement reflects the functional roles these communities played in tournament conversations, whether as hubs of interaction or isolated clusters. These dynamics were also echoed in interaction patterns between communities.

Another important pattern was the variation in how communities interacted with one another. Some groups, particularly those based around more niche topics, or smaller communities, maintained strong internal interaction levels but limited external interactions. In contrast, clusters such as the central global football hubs engaged widely across the network. This contrast reveals the structure of online conversation: isolated communities tended to reflect niche or highly localised interests, while highly interactive communities often acted as hubs of general tournament discourse. This structure reflects ideas in network theory, where high-connectivity nodes are critical for spreading information and sustaining connections between user groups.

A recurring theme across all four networks was the presence of a large, centralised community representing a global football audience. These clusters consistently included official accounts such as @FIFAWorldCup, various national team profiles, and prominent media or football content creators. Typically English-speaking, these communities also featured multilingual engagement, reflecting their broad international reach. Structurally, they acted as hubs, receiving interactions from a wide range of peripheral communities and often serving as the primary point of contact between otherwise disconnected groups. Additionally, consistent national fanbases, particularly Spanish, French, and Portuguese-speaking communities, appeared prominently in both World Cups, often forming well-defined clusters.

4.2 Limitations

While the analysis produced meaningful insights into Twitter community structures during the FIFA World Cups, several limitations should be acknowledged.

One major limitation was the need to significantly reduce the size of the original networks. The raw datasets for both tournaments contained millions of users and interactions, particularly in the retweet networks. Due to computational constraints, these had to be filtered using K-core decomposition to retain only the most connected 1% of users. While this made analysis feasible and allowed for community detection to run reliably in Gephi, it inevitably removed large numbers of users who may still have played meaningful roles within the network. As a result, the findings are skewed toward more active and visible users, and smaller or less connected groups may have been excluded entirely.

Another limitation lies in the inconsistency between the data collection methods for the two World Cups. The 2018 Russia dataset was collected using hashtags and hashflags for national teams, whereas the 2022 Qatar dataset included country flags and official account mentions. These differences may have introduced bias into the types of tweets and users captured in each dataset. For example, the Qatar collection pulled in more off-topic content, which might explain the increased number of unrelated communities in the 2022 networks.

Finally, the project relied on structural network data and metadata such as hashtags, usernames, and tweet languages to understand what each community was about. Because the actual tweet content was not analysed, the themes of communities had to be infered using only this data. This approach worked well for identifying broad topics, like which country a community was based around, but made it harder to understand more detailed behaviours or discussions within smaller communities.

4.3 Future Work

There are several ways that this project could be improved in future research. One key area is analysing two datasets collected using the same methods and search terms. This would make comparisons more reliable, as different terms would not be a possible reason for differences in results.

Another area would be including more of the original network in the analysis. With a more powerful computer, Gephi would have been run with more nodes in each network, allowing for a more in-depth exploration, instead of only analysing 1% of all users.

While this project has focused just on two World Cups, incorporating past World Cups would give a better idea of trends in user interaction over time on Twitter. Looking at other global events, such as the Olympics, could also provide more information about user interaction during global events.

Lastly, analysing the actual content of tweets could help strengthen how communities are identified and labelled. For example, using natural language processing (NLP) techniques to extract common topics or sentiment within each community would provide a more detailed picture of what each group is discussing.

5 Conclusion

This project explored how Twitter users interacted and formed communities during the 2018 and 2022 FIFA World Cups by building and analysing networks based on retweets and comments. Across all networks, communities tended to form around shared languages, national teams, and significant media or football accounts. Although some limitations, such as reduced network size, affected the depth of analysis, the findings offer a foundation for understanding how global events shape user interaction on social media.

The project also found that specific communities consistently acted as central hubs for tournament discussion, while others remained more isolated or niche. The appearance of unrelated communities, especially in the 2022 dataset, highlighted changes in the nature of Twitter users, possibly influenced by bot activity.

References

[1] B. Piovesan, "worldcup on twitter: The g.o.a.t." 2022, official Twitter blog.

[2] G. Yan, N. M. Watanabe, S. L. Shapiro, M. L. Naraine, and K. Hull, "Unfolding the twitter scene of the 2017 uefa champions league final: Social media networks and power dynamics," *European Sport Management Quarterly*, vol. 19, no. 4, pp. 419–436, 2019.

[3] A. Zmudzinska and B. Wietczak, "Analysis of twitter activity by country during competitive events," 2018.

[4] M. A. Smith, L. Rainie, B. Shneiderman, and I. Himelboim, "Mapping twitter topic networks: From polarized crowds to community clusters," Pew Research Center, Tech. Rep., February 2014.

[5] A. Logan, P. LaCasse, and B. Lunday, "Social network analysis of twitter interactions: a directed multilayer network approach," *Social Network and Analysis Mining*, vol. 13, no. 65, 2023.

[6] D. Pacheco, F. B. de Lima Neto, L. Moyano, and R. Menezes, "Football conversations: What twitter reveals about the 2014 world cup," in *Anais do IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM).* Sociedade Brasileira de Computação, 2015.

[7] N. Govind and R. P. Lal, "Evaluating user influence in social networks using k-core," in *Proceedings of the International Conference on Innovative Computing and Communications*, ser. Advances in Intelligent Systems and Computing, D. Gupta *et al.*, Eds. Springer, 2021, vol. 1166, pp. 11–18.

[8] S. B. Seidman, "Network structure and minimum degree," *Social Networks*, vol. 5, no. 3, pp. 269–287, 1983.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[10] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PLoS One*, vol. 9, no. 6, p. e98679, 2014.

[11] Twitter, "Introducing new metadata for tweets," https://blog.x.com/developer/en$_us$/a/2013/$introducing-new-metadata-for-tweets$, 2013.

Appendix A

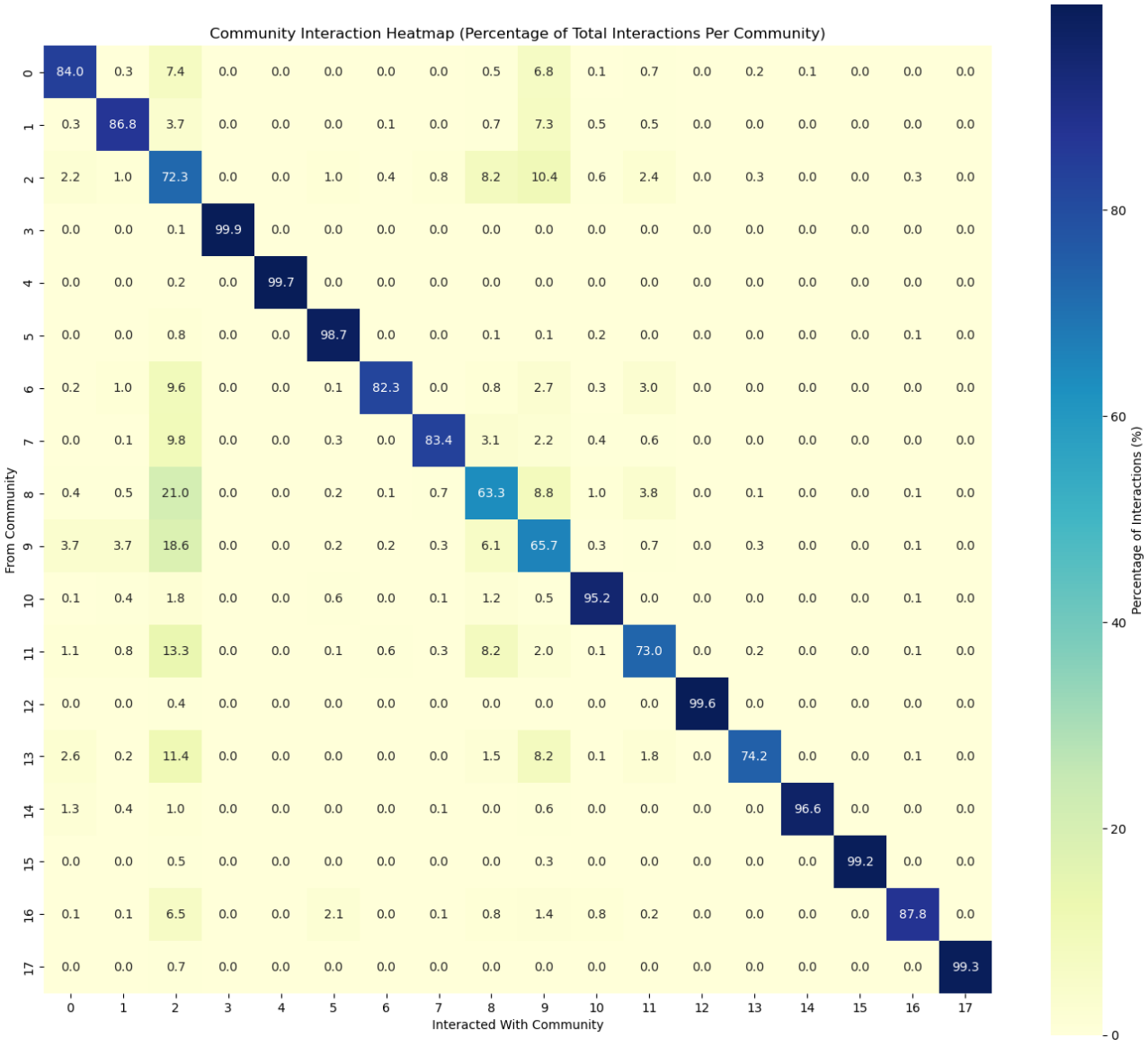


Figure 16: Heatmap for intercommunity interactions in Russia 2018 Comment network

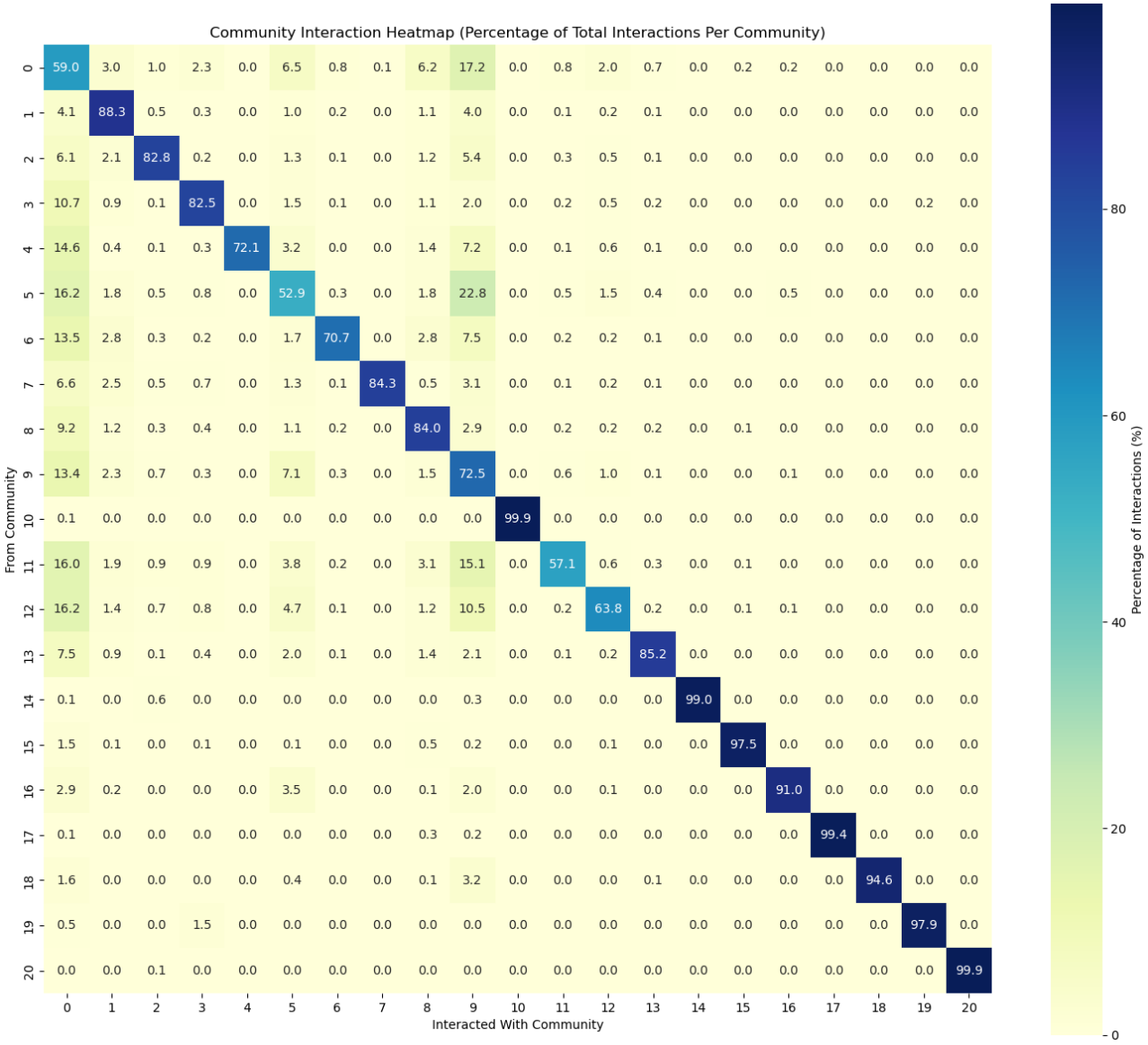


Figure 17: Heatmap for intercommunity interactions in Russia 2018 Retweet network

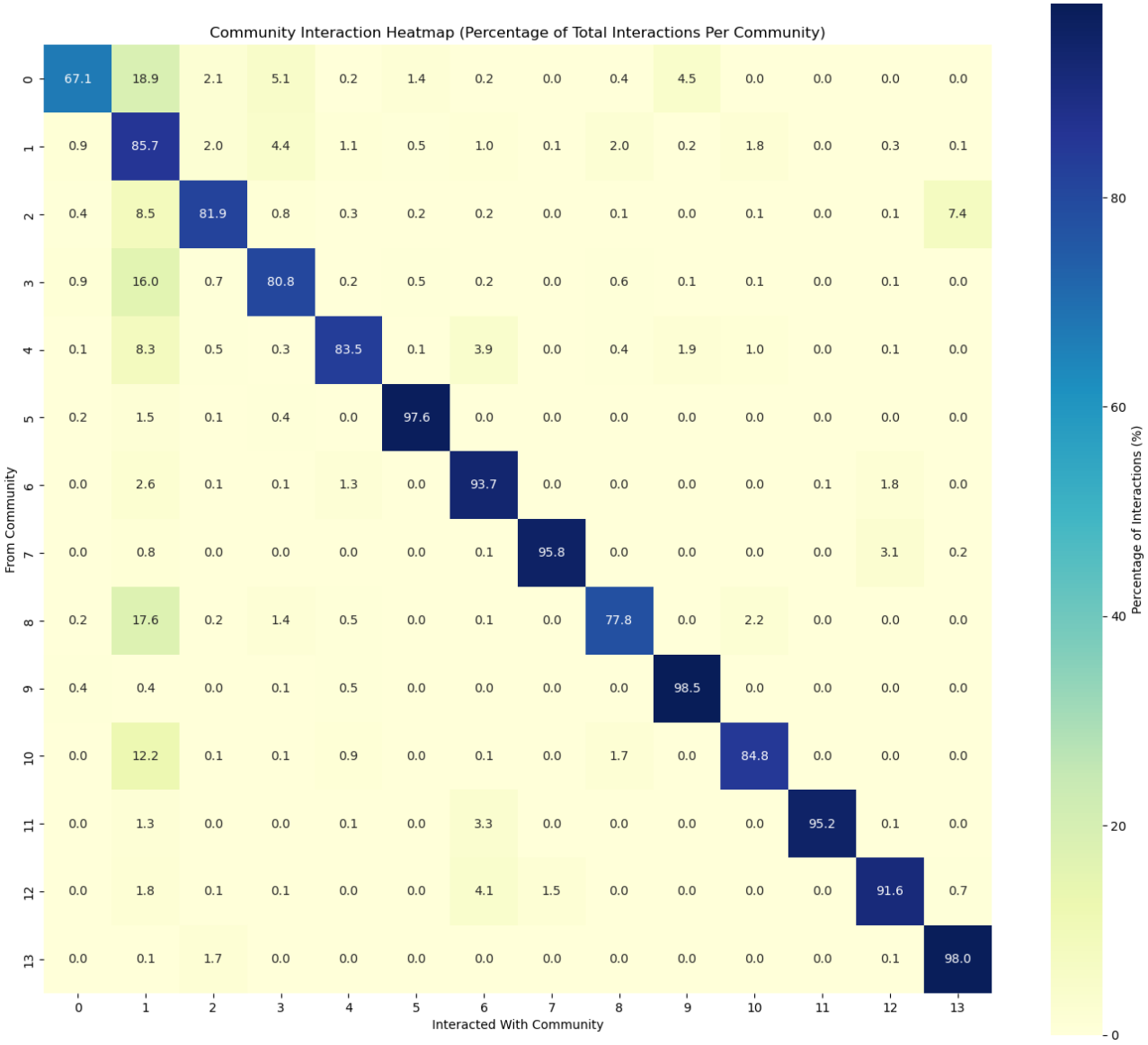


Figure 18: Heatmap for intercommunity interactions in Qatar 2022 Comment network

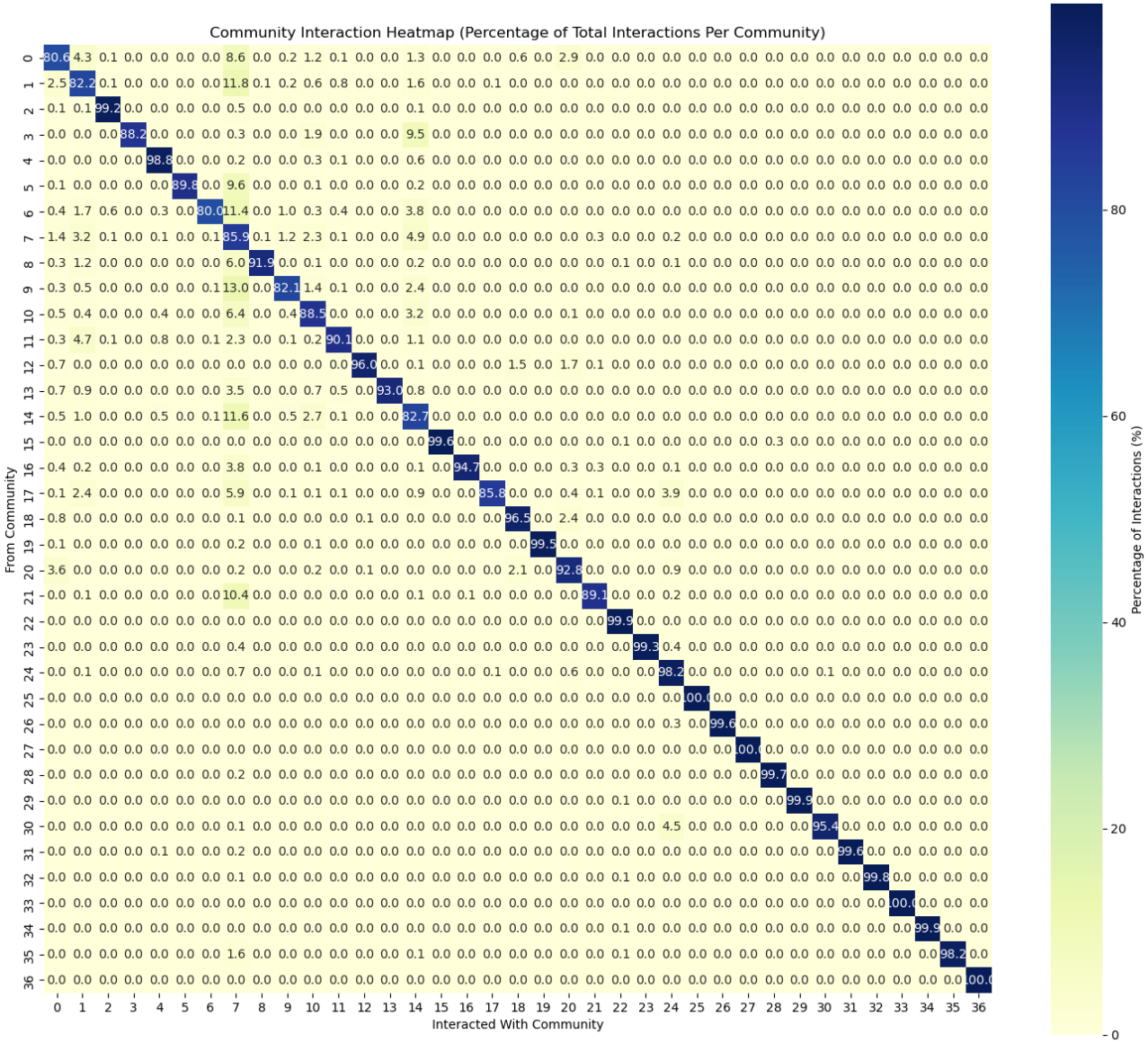


Figure 19: Heatmap for intercommunity interactions in Qatar 2022 Retweet network