

Twitch Project Paper

Team 13B

Fazil Ahmed

Wei-Chao Chen

Ethan Le

Jiayi Yin

Ziyue Zhang

BANA 277

March 16, 2024

**Table of Contents**

[**1. Introduction 2**](#_2vv60yajiw7j)

[**2. Data Explanation 2**](#_3k6r24cxkzik)

[**3. Exploratory Data Analysis 3**](#_g8ja8swwx9s)

[3.1 Views 4](#_gnjm7xt3f8w)

[3.2 Mature Content 4](#_wvt1yub3cwv3)

[3.3 Lifetime 5](#_tklygp6h0qkw)

[3.4 Dead Accounts 5](#_lrfak7igp0o4)

[3.5 Language 6](#_riycxeuv225p)

[3.6 Affiliate Status 7](#_750pegq2wv4y)

[**4. Network Analysis 7**](#_omqyuebb3ieu)

[4.1 Homophily Analysis 8](#_dh4cmf9007d1)

[4.2 Propensity Score Matching 11](#_whcre028ztl)

[4.3 Poisson Regression 12](#_34g9kfkedotg)

[**5. Interpretations & Recommendations 13**](#_8svgyg762okb)

[5.1 Interpretations of Results 14](#_d9ihm5jq4npg)

[5.2 Business Insights and Recommendations 14](#_kynhsgqbtpzf)

[**6. Final Comments & Conclusion 16**](#_3gswr3gelhnf)

[**7. References & Code 17**](#_4ozdv4nb7h1s)

# Introduction

In the realm of digital entertainment live streaming has ascended to become a fundamental medium over the past decade. Among the various segments contributing to this growth, gaming has been a prominent driver. Twitch, an Amazon subsidiary, stands at the forefront of this evolution, not merely as a streaming platform but as a comprehensive social network. It provides a platform where streamers and viewers engage in a symbiotic relationship, influencing one another through direct and multifaceted interactions.

Twitch exemplifies the integration of same-side and cross-side network effects within its ecosystem, where participants, streamers, and viewers alike are interconnected nodes. These nodes are capable of establishing unidirectional or bidirectional connections, fostering a rich tapestry of social interactions. Such a model has been instrumental in Twitch's exponential growth, nurturing a dedicated user base that regards the platform as an unparalleled hub for content. This project proposes to conduct a thorough analysis of the dynamics and user interactions within Twitch's ecosystem. Our objective is to uncover the underlying patterns that contribute to the platform's success and to understand the characteristics that define its engaged and vibrant community.

# 

# Data Explanation

Our objective is to further understand this trend through the analysis of a SNAP dataset containing information about Twitch streamers and users. This dataset is available on Stanford's SNAP website and was originally collected via Twitch's public API in 2018. It was designed to address six key goals: detecting streamers who show explicit content, forecasting the language of broadcasters, calculating the lifespan of users, anticipating user departure, recognizing affiliate status, and predicting the number of views. Our aim is not to duplicate these studies but to utilize their findings as a comparative framework for our own research. We intend to use this dataset to identify network patterns and analyze how users cluster around certain broadcasts. The dataset includes 168,114 nodes and 6,797,557 edges, representing Twitch users and the mutual followership links between them, respectively. The characteristics of users that were analyzed are as follows:

| **Numeric\_id** | User identifier. Split into two categories for Edges dataset |
| --- | --- |
| **Views** | A count of the number of views a user’s streams have received |
| **Mature** | Whether the user streams explicit content (1 or 0) |
| **Life\_time** | The age of the user’s account in days |
| **Created\_at** | The date the account was created |
| **Updated\_at** | The date of the last update of the account |
| **Dead\_account** | Whether an account is classified as dead or not (1 or 0) |
| **Language** | The language of the user’s streams |
| **Affiliate** | Whether a user is a Twitch affiliate or not (1 or 0) |

# 

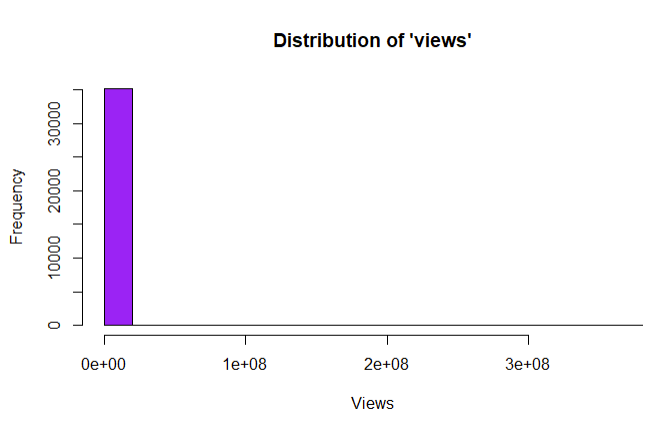
It should be noted that in the Edges dataset there are only two variables: numeric\_id\_1 & numeric\_id\_2 where each ID refers to an individual user that comes from the Features dataset. These edges are then added, recorded, and used to establish a relationship between two users. The way something such as views are tracked is that each ID has a count for how many views their stream has gotten (from the Features dataset) and then any viewers that have watched someone else’s stream would be listed in the opposing Numeric\_id column in the Edges dataset to show how they are connected. We can then use this to track to see which other additional features are related amongst these different networks that are being created.

# Exploratory Data Analysis

We began our exploratory data analysis by first testing how large of a network we could analyze with the computational power we had at our disposal. As the dataset itself had 168,114 features and 6,797,557 edges, it was simply taking too long for our devices to perform the calculations and analysis we wanted. Because of this we made the decision to perform multiple analyses on a subset of the data, this included a subset of 1000 nodes connected to the node with the highest degree which in our case was the node holding the ID 61863. From there we began our analysis by looking at the distribution of our variables. We observed the distributions of views, mature content, account lifetime, dead status, language, and affiliate status. It should be noted that the decision was made to ignore the “creation” and “last update dates” variables as account lifetime proved to be more insightful than those attributes which are derived from it. We found the following through our exploratory analysis:

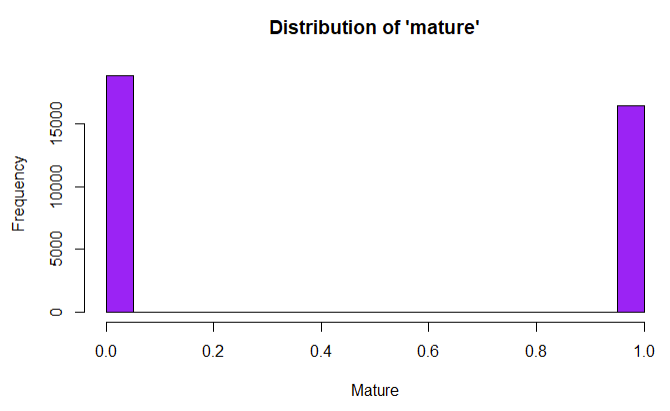
## 3.1 Views

An overwhelming majority of users in this network had little to no views on their accounts. Views as a category had an overall median of 4644.5 but a mean of 265513.4. This suggests that while there are large streamers within this network, most users either did not create streams themselves or rather had streams that did not generate much traffic. Our histogram confirmed this, with the volume of users at the 0e+00 mark being so high that other users are not showing up, though we know they are there due to the scale of the x-axis.



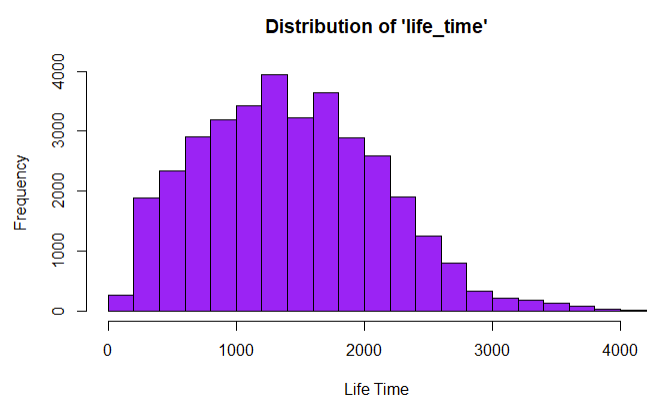
## 3.2 Mature Content

The users in our subsetted network were close to evenly split between those who created mature content and those who did not, with 53% and 47% of the data respectively. This suggests at first glance that there may not be much homophily in regards to the maturity of content on Twitch, and we made note of this. Mature is one of the features we hoped to do further analysis on and will do as such in our Network Analysis section.

****

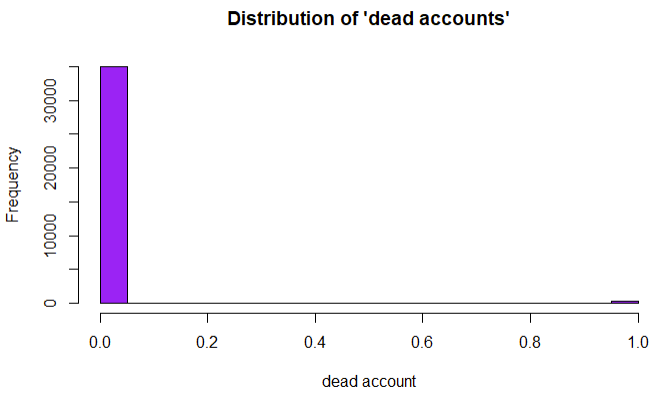
## 3.3 Lifetime

We found the lifetime age of accounts is very close to being normally distributed, with a majority of users having accounts between 2.5 and 5.5 years of age. Most notably there were users older than Twitch itself, which was officially launched in 2011. While this dataset contains data up until 2018 it is more likely that the users with accounts older than 7 years old were users who had been with the platform while it was still under the domain “Justin.tv.” With Twitch (or twitch.tv as it is accessed via URL) tracing its origins to founder Justin Kan’s livestreaming website launched in 2007.



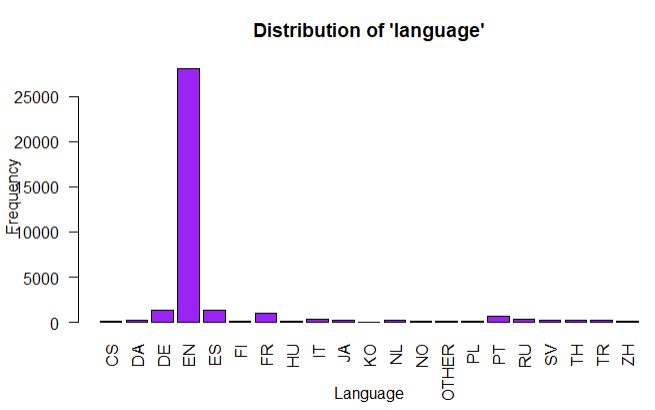
## **3.4** **Dead Accounts**

We found that over 99% of accounts were not considered dead within our dataset. This means that very few accounts had been inactive for long enough to be considered a dead account, and thus our analysis should be relevant towards active users, who are the targets we would like to look at.



## 3.5 Language

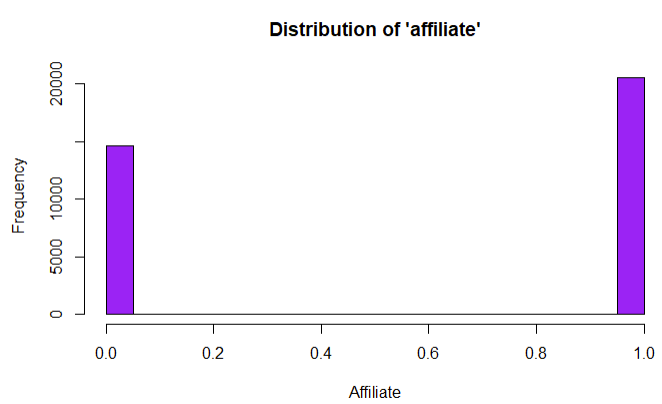
English was by far the most common language for streams to be in. While Twitch is very much a global sensation, especially after the pandemic, it seems that its popularity was mostly limited to the United States or other English speaking countries back in 2018. If we were to look at data for 2024, it is likely the distribution would be quite different, though even now the United States is still the largest market for Twitch.



## 

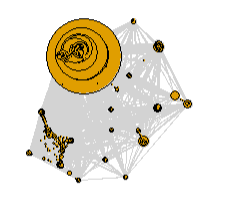
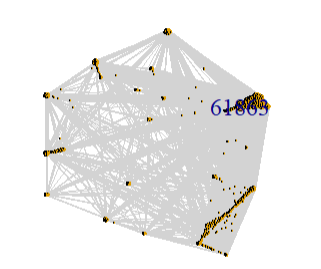
## 3.6 Affiliate Status

Our network is split with 58% of users being affiliates and 42% listed as non-affiliates. Looking at the overall dataset, it seems that affiliates were oversampled, which proved to be helpful for our analysis, allowing for us to understand the relationship between affiliate status and mutual follower connections as well as overall channel views.



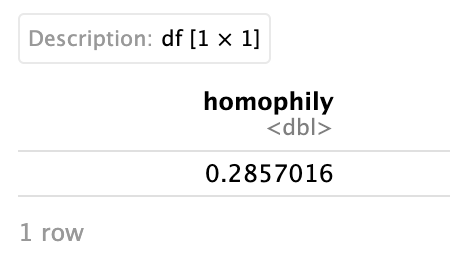
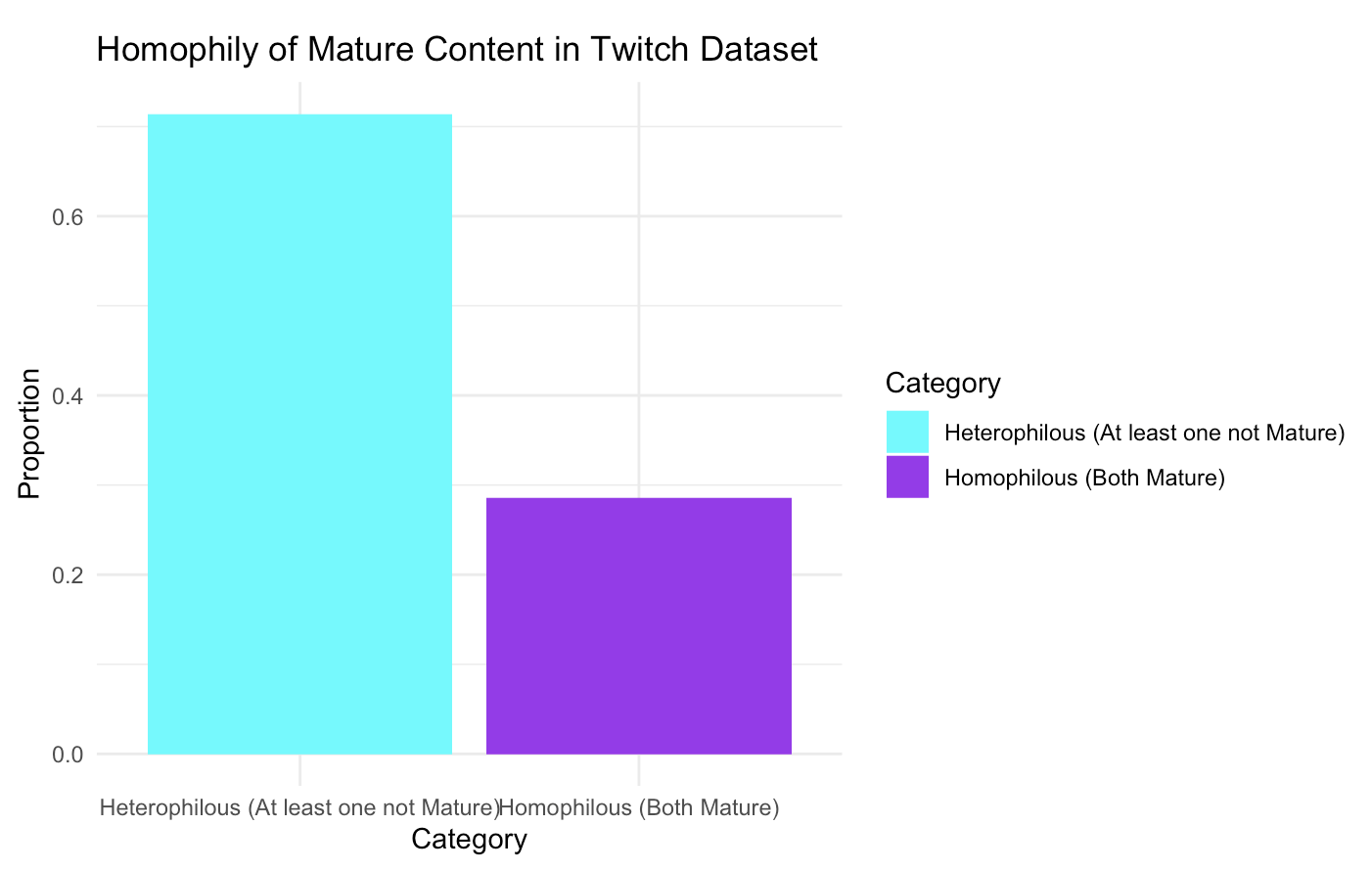
# Network Analysis

We visualized our network in various ways using iGraph in R, and came to understand various properties of this network. We found extremely clear clusterings of nodes, signifying the existence of communities on Twitch. Users within a community would have large groups of followers that consisted of the same viewers. Additionally, we found that most of the streamers with the highest number of mutual followers were a part of the same community. This hinted at the importance of networking on Twitch. As the largest content creators found themselves creating content for the same community, the success of a single streamer within a community seemed to correlate with the success of other streamers in that community. We also saw that although the ties are strongest within a community, there were also mutual follower connections between communities, with smaller communities being linked together through their connection with the largest streamers.

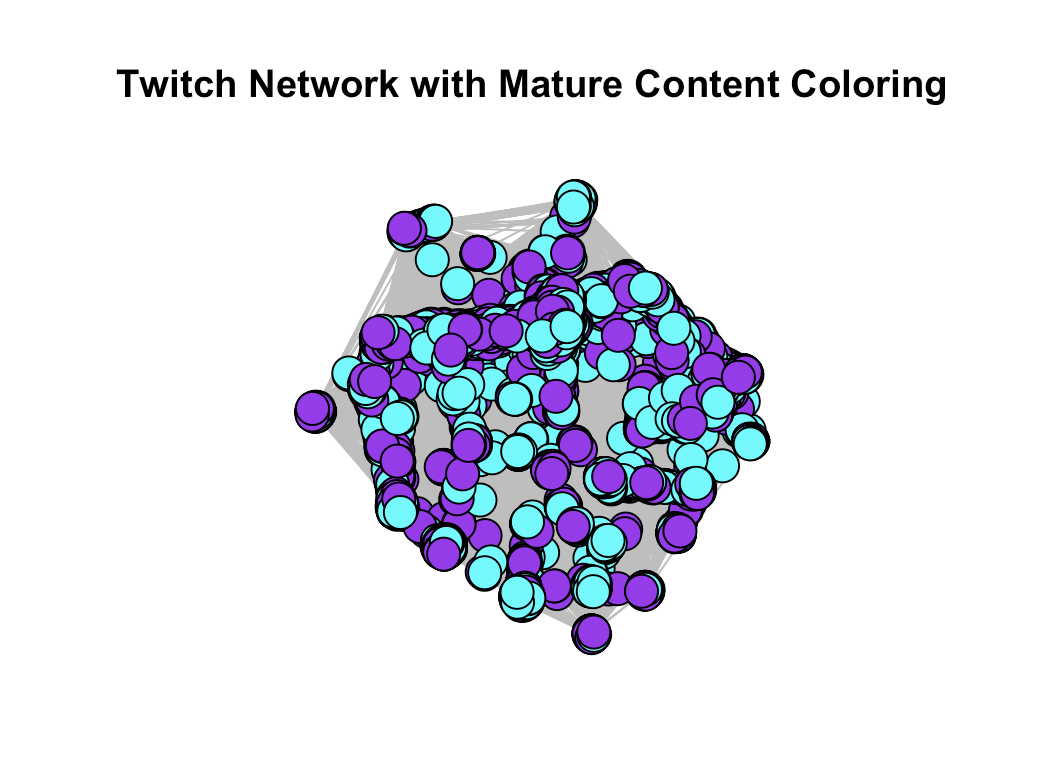


## 4.1 Homophily Analysis

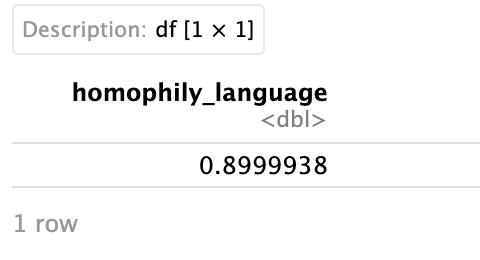
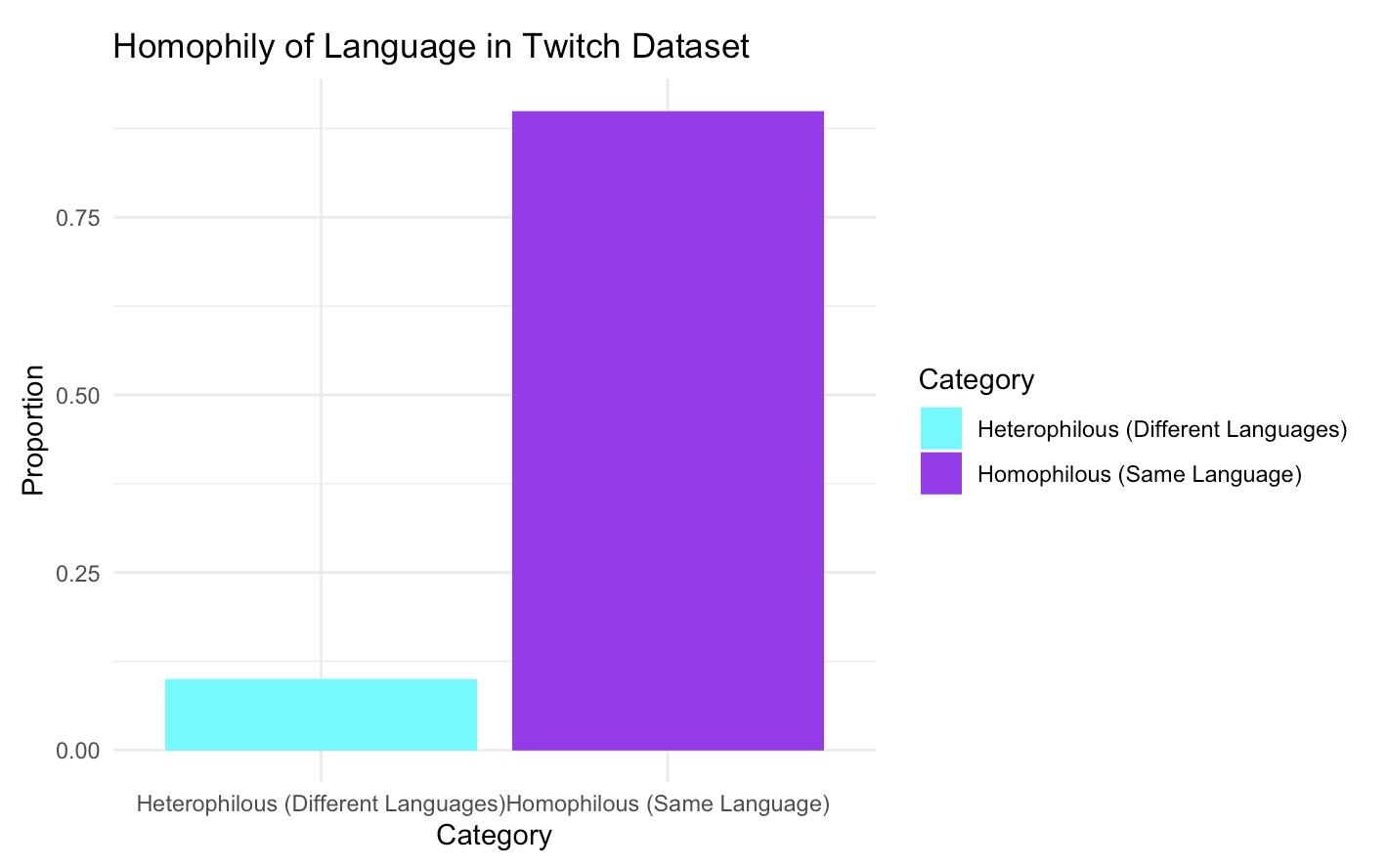
By utilizing both iGraph & ggplot2 in R we were able to perform homophily analysis on the mature content and language features of the Twitch dataset. It should be noted that this analysis was not conducted on a subset of the data similar to what we did during our EDA, but rather on the dataset as a whole. The analysis on the mature content feature resulted in showing moderate homophily (homophily rating of 0.2857) as users do show a tendency of connecting with others that produce similar types of mature/non-mature content, but this was far outpaced by the number of cross connections that existed between the two. In short, while there is a number of viewers that prefer to watch only mature content, or users that prefer to only watch non-mature content the vast majority within this dataset enjoy a mix of the two showing an appreciation for a wide variety of streams.



Below is the node network of the mature content feature, and as can be seen while there is some grouping the vast majority of nodes are scattered next to one another showing more cross connections between them.

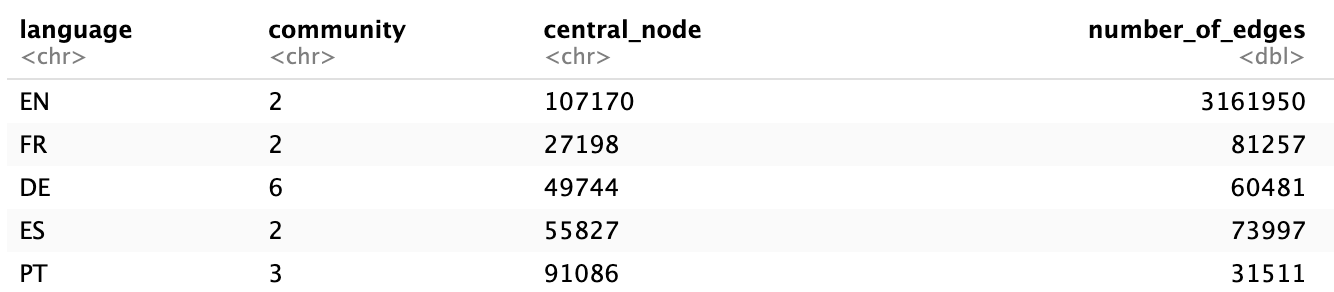


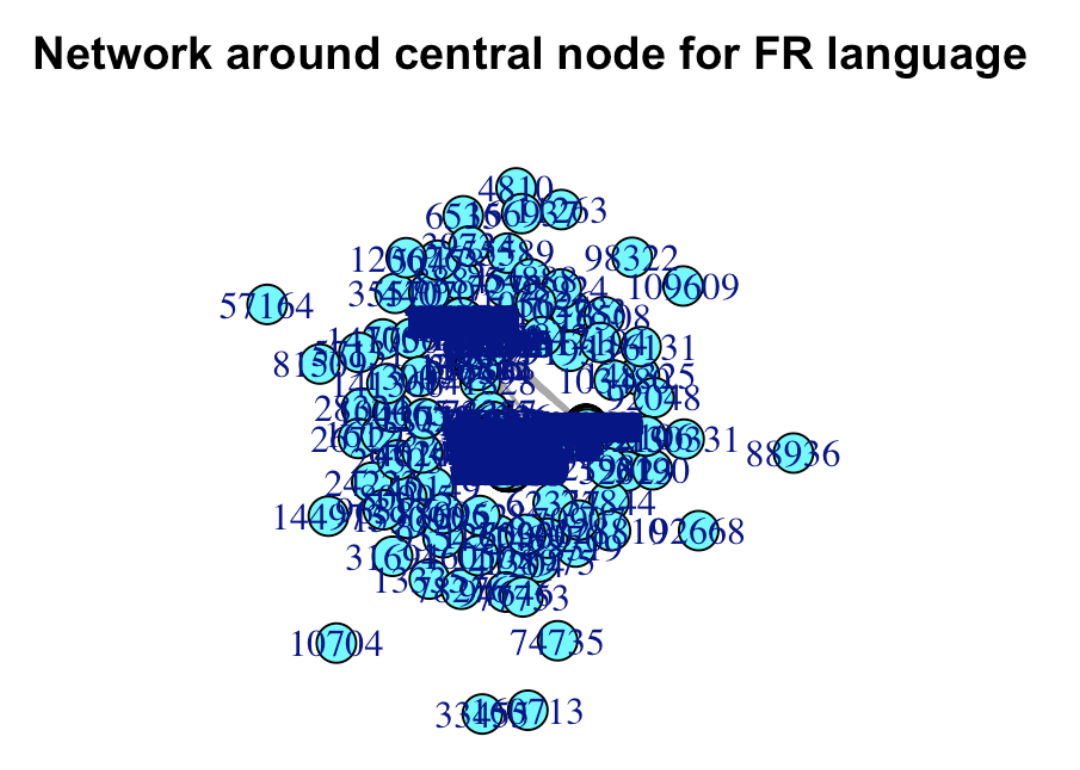
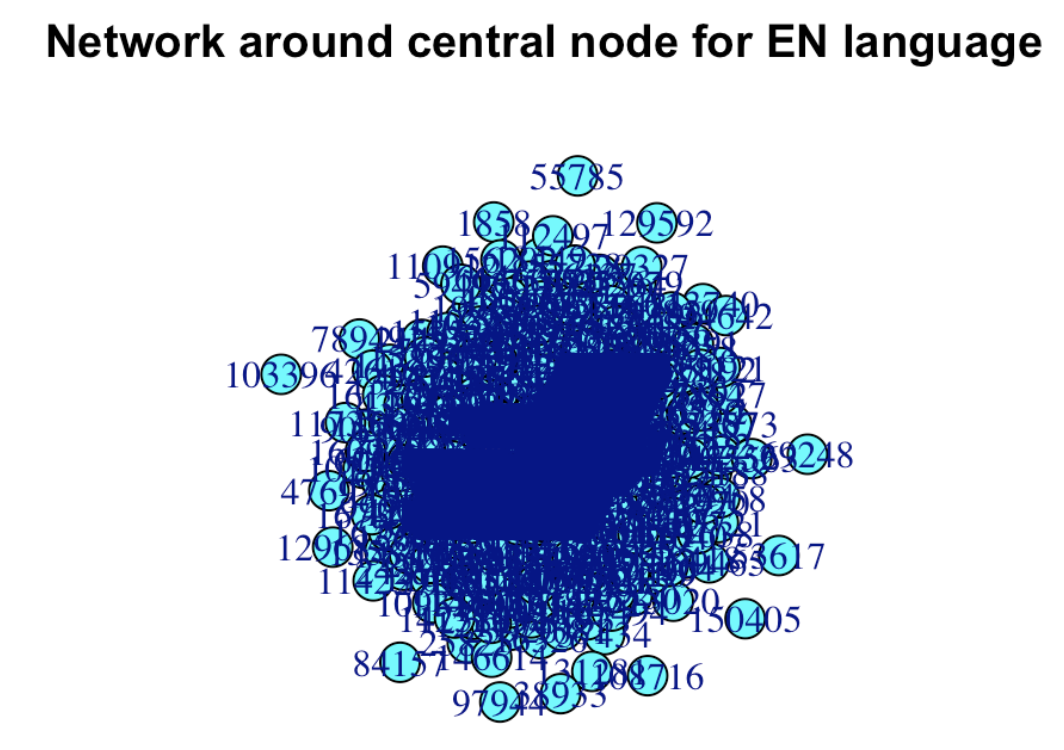
In contrast there is very strong homophily based on language (homophily rating of 0.8999). With a rating this high it appears that most users are overwhelmingly connected to others streaming in the same language. This likely reflects both viewer preferences and the real-world human tendencies to cluster based on shared language and culture.

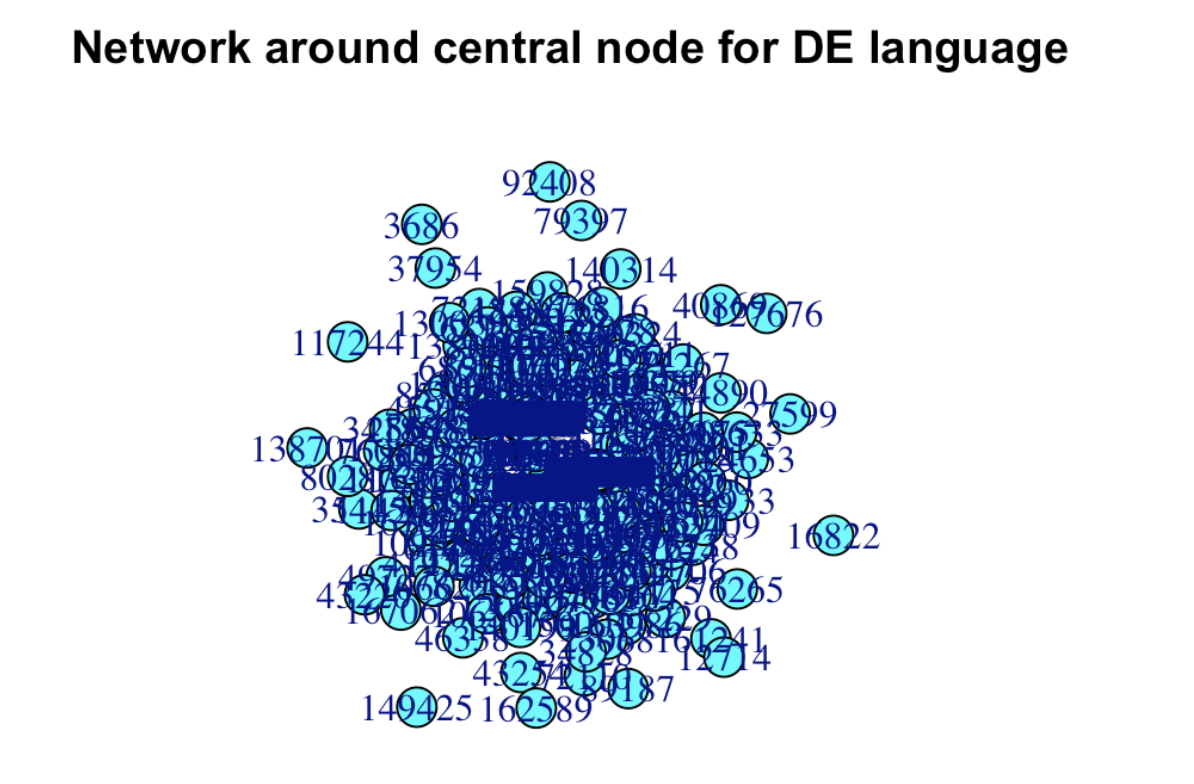
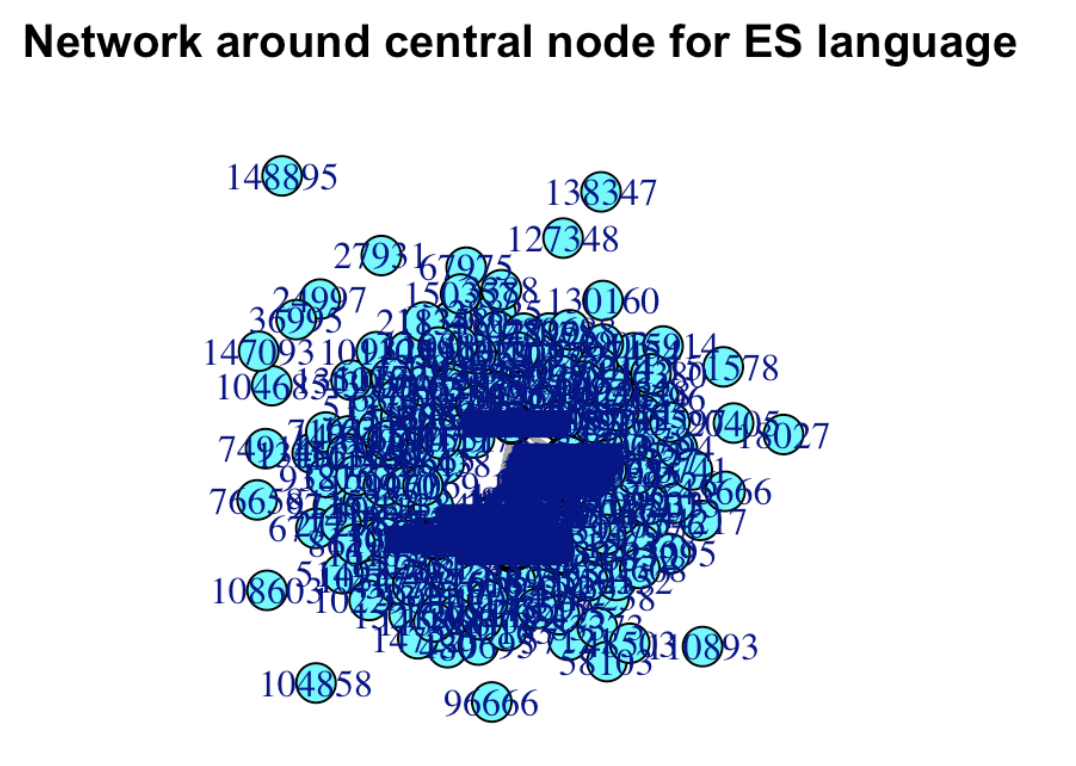


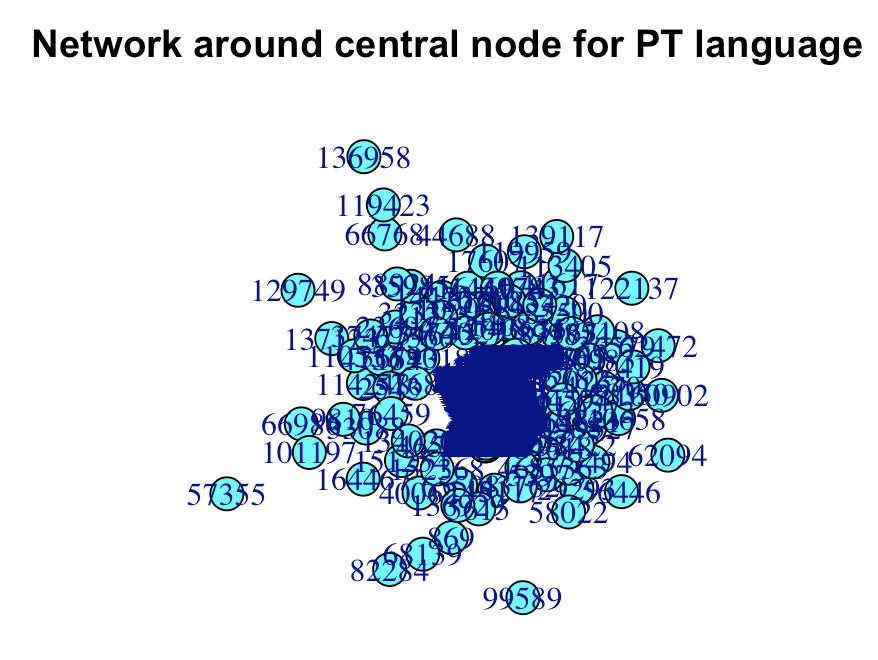
However, another factor we needed to consider because of this was that the data may be skewed because there are so many English speakers on the platform. Maybe other languages didn’t show the same signs of homophily? To test this we once again subsampled the data but this time with central nodes of each of the top 5 languages. This included English, French, German, Spanish, and Portuguese. We then calculated the number of edges each language in these subsets have and finally which sub-community they belong to. It should be noted the sub-community calculations are not assigned by the dataset or even infer that they are very close to one another just that the way the data was split up and separated the central nodes were closest to these particular communities.

What we noticed is that even without calculating the exact homophily of each language there were clear signs and indications that the languages of English, German, and Spanish are heavily homophilous whereas French and Portuguese have concentrated central nodes but other nodes that seem to be branching out of their community. This shows that while English is playing a significant role in influencing how homophilous language is in the dataset, other top languages echo similar sentiments.







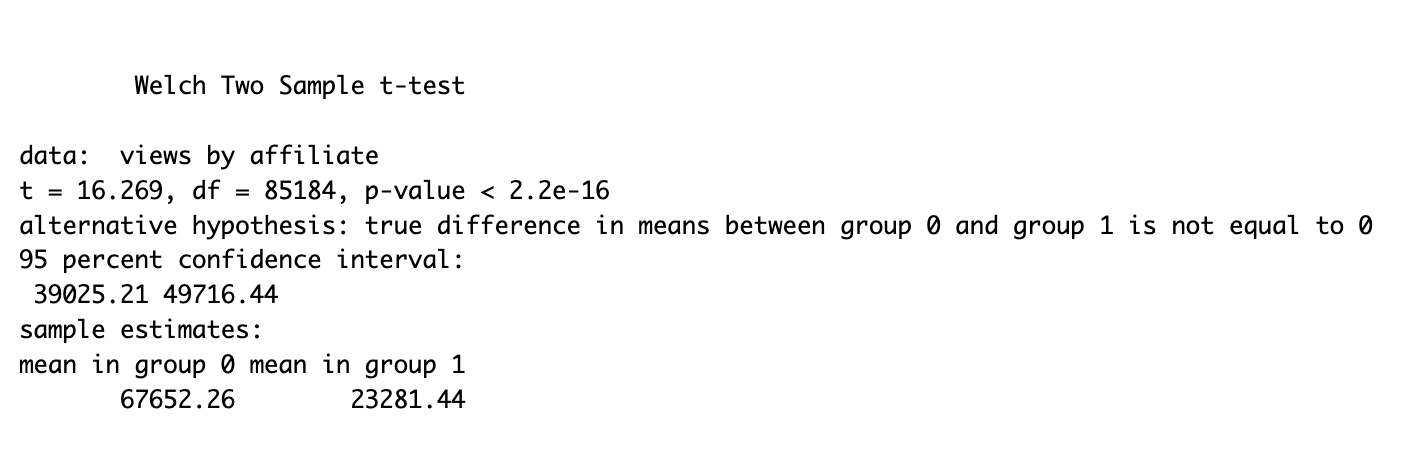


## 4.2 Propensity Score Matching

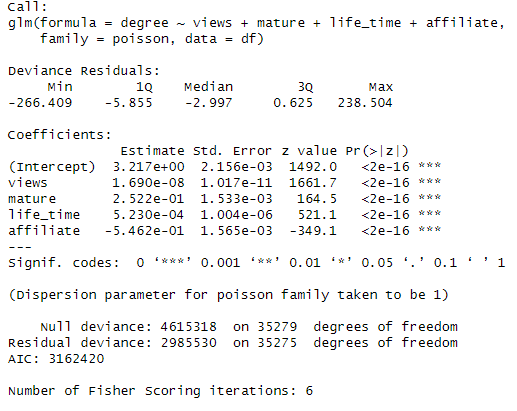
The propensity score matching analysis quantifies how user attributes relate to network position (degree) and viewership. In our initial correlation models we noticed that affiliate status was negatively associated with degree centrality, suggesting affiliates may have more dedicated individual followings rather than being as highly connected across the network thus leading to less views.

To verify this we conducted a propensity score matching of the affiliate feature and did so to truly verify if there was a negative association with affiliate status. Based on the output below it can be seen with statistical significance that affiliate status does in fact have a negative correlation with views, suggesting that affiliates prefer to cater towards their own niche communities rather than grow their stream.

Group 0 represents non-affiliates while Group 1 represents affiliates and as can be seen the views for Group 0 far outnumber that of Group 1.

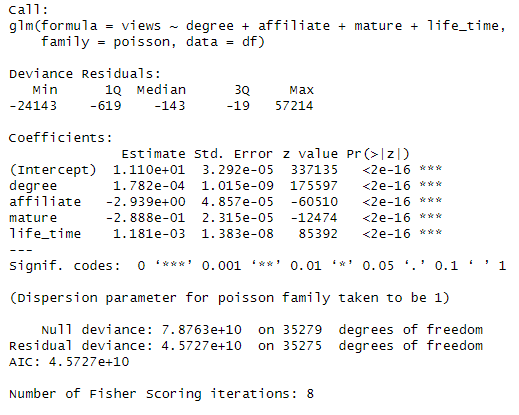


## 4.3 Poisson Regression



We had initially run our variables under a linear regression, measuring the relationship between variables and mutual follower connections (degree), and the relationship between variables and overall views, but as the scale of views was so large in comparison to every other variable, we opted to use a Poisson regression model to understand these relationships instead. In relation to degree, we found that views, being an account that creates mature content, and the lifetime of the account all had positive coefficients, while affiliate status had a negative coefficient. This suggested that within this network, Those with more views, longer account lifetimes, and mature content streamers were more likely to have more mutual follower connections. Interestingly, affiliates had less mutual follower relations.

In relation to views, we found that degree and lifetime of the account had positive coefficients while mature content and affiliate status had negative coefficients. This suggested that users who had larger communities (more mutual follower connections) and whose accounts were older tended to have more views, while interestingly mature content creators and affiliates had less views.



# Interpretations & Recommendations

In the intricate landscape of Twitch, a live streaming platform that has burgeoned into a vibrant community of creators and viewers, a detailed analysis offers a plethora of interpretations, insights, and actionable recommendations. This section delves into these aspects, drawing from a comprehensive study conducted on Twitch's network dynamics, user engagement patterns, and the overarching influence of content and language preferences. Through a meticulous exploration of views distribution, content maturity, account lifetime, language dominance, and affiliate status, we unveiled the multifaceted nature of Twitch's ecosystem.

## 5.1 Interpretations of Results

Our analysis illuminated a stark disparity in views distribution, revealing a concentrated viewership among a select cadre of streamers. This phenomenon underscores the challenges and opportunities within Twitch's competitive landscape, where visibility and engagement are pivotal. The exploration of mature content revealed a balanced dichotomy, suggesting that Twitch harbors a diverse audience with eclectic preferences, thus negating any significant polarization between mature and non-mature content. Intriguingly, the account lifetime analysis underscored Twitch's prowess in sustaining long-term user engagement, a testament to the platform's enduring appeal and its strategic initiatives to retain users.

The investigation into account activity unearthed a vibrant community, with over 99% of accounts being active. This highlights the platform's success in fostering an engaged user base, a critical component of its social fabric. Furthermore, the linguistic analysis brought to light the predominance of English, hinting at Twitch's stronghold in English-speaking territories. However, this also spotlighted potential growth avenues in diversifying language support to tap into global markets.

A pivotal revelation was the nuanced relationship between affiliate status and network centrality. Affiliates, despite having dedicated followers, exhibited a lesser degree of interconnectedness within the Twitch network compared to non-affiliates. This insight opens up discussions about the structural dynamics of Twitch's affiliate program and its impact on community integration and content dissemination.

## 5.2 Business Insights and Recommendations

The study's findings underscore the critical role of community and network dynamics in shaping user experience and engagement on Twitch. The pronounced cluster formations and niche communities signify the platform's organic growth through shared interests and interactions. To leverage this, Twitch could enhance its community building tools, fostering deeper connections and facilitating discovery across these micro-communities. This approach not only strengthens the social bonds within Twitch but also amplifies content reach and engagement.

The strong homophily based on language presents Twitch with a golden opportunity to diversify its content and expand its global footprint. By promoting localized content and bolstering support for non-English streams, Twitch can unlock new audience segments and cultivate a more inclusive and diverse community ecosystem. This strategy aligns with the global trend towards localization and could significantly enhance Twitch's market penetration and viewer engagement across different regions.

The analysis also calls for a critical reevaluation of the affiliate program. The negative correlation between affiliate status and network centrality suggests a potential misalignment with Twitch's broader community engagement goals. Revamping the affiliate program to encourage wider network participation and cross-community interactions could redefine affiliate pathways, aligning them more closely with Twitch's community-centric ethos.

Moreover, Twitch could capitalize on the influence of "hub" streamers as catalysts for community bridging and content discovery. Encouraging collaborations and cross-promotions among streamers from different communities could foster a more interconnected network, enhancing content diversity and viewer experiences.

Lastly, an investment in advanced data analytics is paramount. By harnessing real-time analytics and deeper insights into user behavior, content trends, and network dynamics, Twitch can equip its creators with the tools to adapt and thrive in a constantly evolving digital ecosystem. This forward-looking approach not only benefits the streamers but also enriches the viewer experience, ensuring Twitch's sustained relevance and growth in the live streaming domain.

In conclusion, Twitch stands at a pivotal juncture, with the potential to redefine its platform through strategic community engagement, content diversification, and an enriched understanding of its vast network. By embracing these recommendations, Twitch can navigate the complexities of the digital age, fostering an environment where creators and viewers alike thrive in a dynamic, inclusive, and engaging community.

# 

# Final Comments & Conclusion

Through analyzing this Twitch dataset, we uncovered novel insights about the platform's community structures, homophily patterns, and growth strategies. The Twitch network exhibits distinct clustering, with small niche communities densely interconnected around shared interests. However, these niches are bridged by highly influential "hub" streamers who connect disparate groups.

Profound homophily exists based on language, with users overwhelmingly connecting to others streaming in the same language. Cross-language ties are more rare, but do occur. And moderate homophily surrounds mature content, but it is significant that there are more instances of viewers who watch both mature and non-mature content.

Surprisingly, factors like more views, older accounts, and mature streams correlated with having more mutual connections across the network. In regards to degree centrality it was shown that having expansive mutual connections was the primary driver of higher viewership as well. But affiliate status was negatively associated with degree centrality, suggesting affiliates cultivate dedicated individual viewers rather than being influencers. This was backed up by running propensity score matching calculations and poisson regression, showing that this phenomenon does exist within this dataset.

In summary, developing niche communities paired with strategically expanding mutual ties through cross-promotion by language-sharing hubs appears more effective for growth than individual audience-building. The Twitch network rewards expansive reach over siloed strategies, highlighting community integration as more powerful than top-down incentive programs. These counterintuitive insights hold valuable implications for both live streaming platforms and individual creators aiming to combat viewer fragmentation

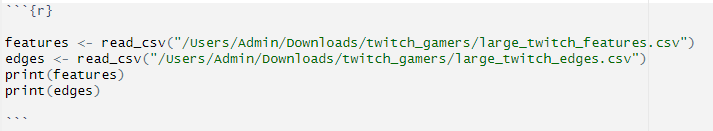
# References & Code

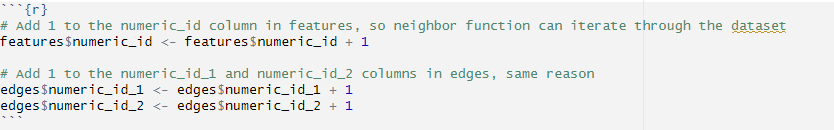
1. “What Happened to Justin.Tv & Why Did They Shut Down?” *RSS*, Failory, 10 Jan. 2024, www.failory.com/cemetery/justin-tv#:~:text=Founded%20in%202007%2C%20Justin.tv,sports%20to%20video%20game%20gameplays. Accessed Feb. 2024.

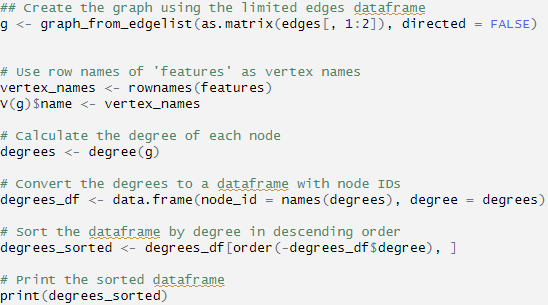
**Dataset:** [**https://snap.stanford.edu/data/twitch\_gamers.html**](https://snap.stanford.edu/data/twitch_gamers.html)

**Code**

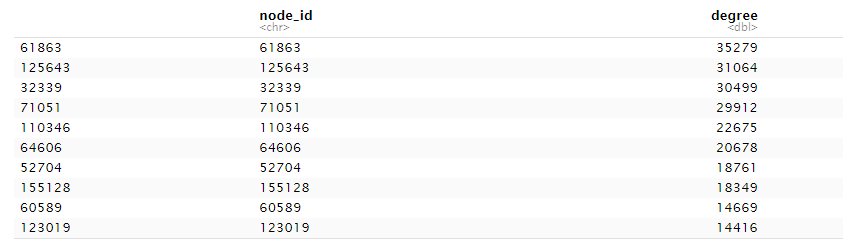
**Data Import & Cleaning**

****

****

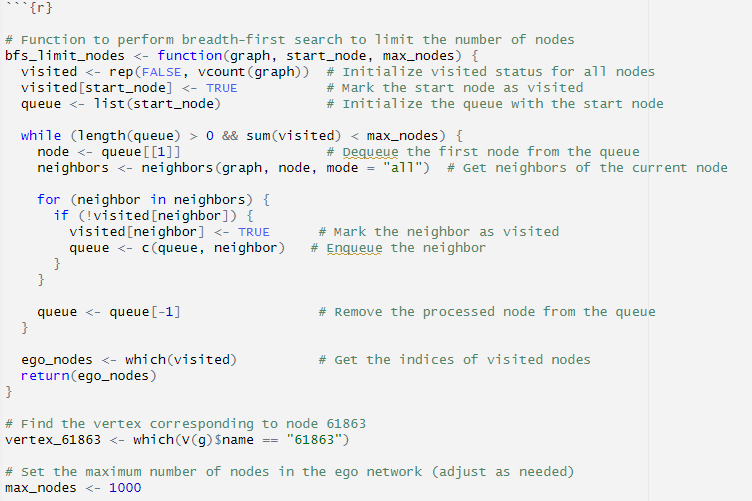
****

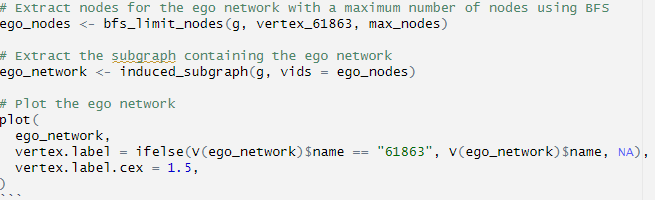
**^**This was used to find our node with the highest degree. Output:



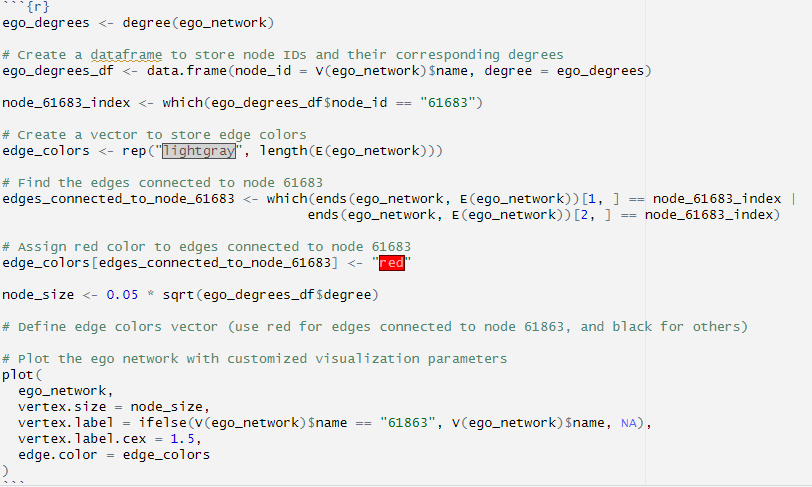
**Node Network Analysis**

**Creating the Network:**

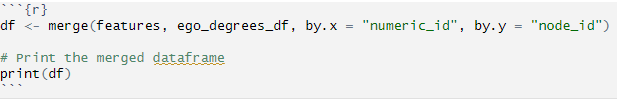


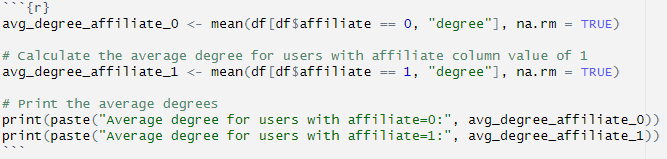


**Network with node size scaled with degree:**

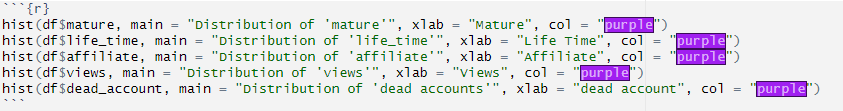
****

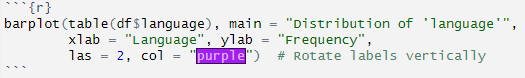
**Exploratory Analysis:**

****

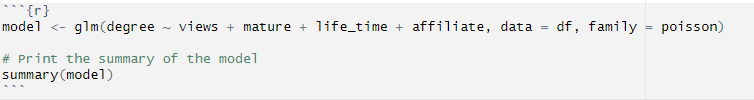
****

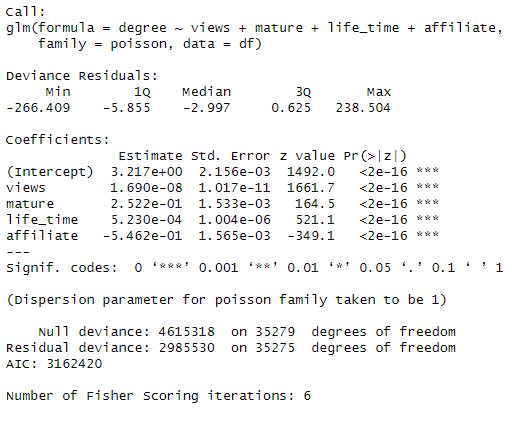


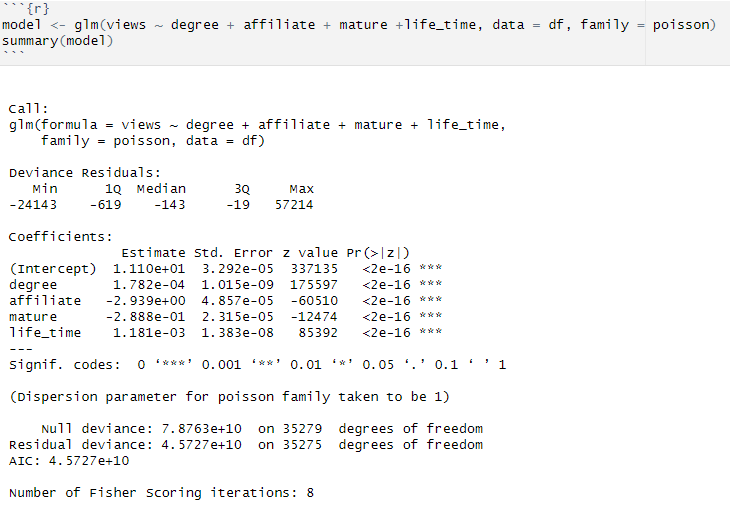


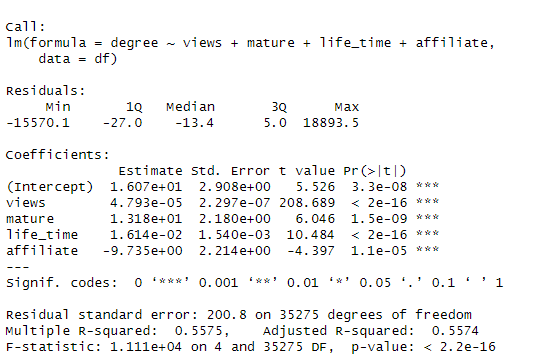


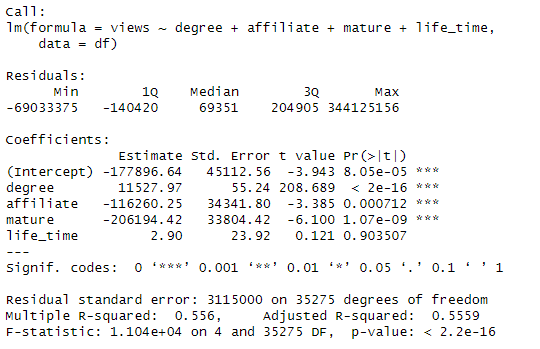
**Poisson Regression**

****

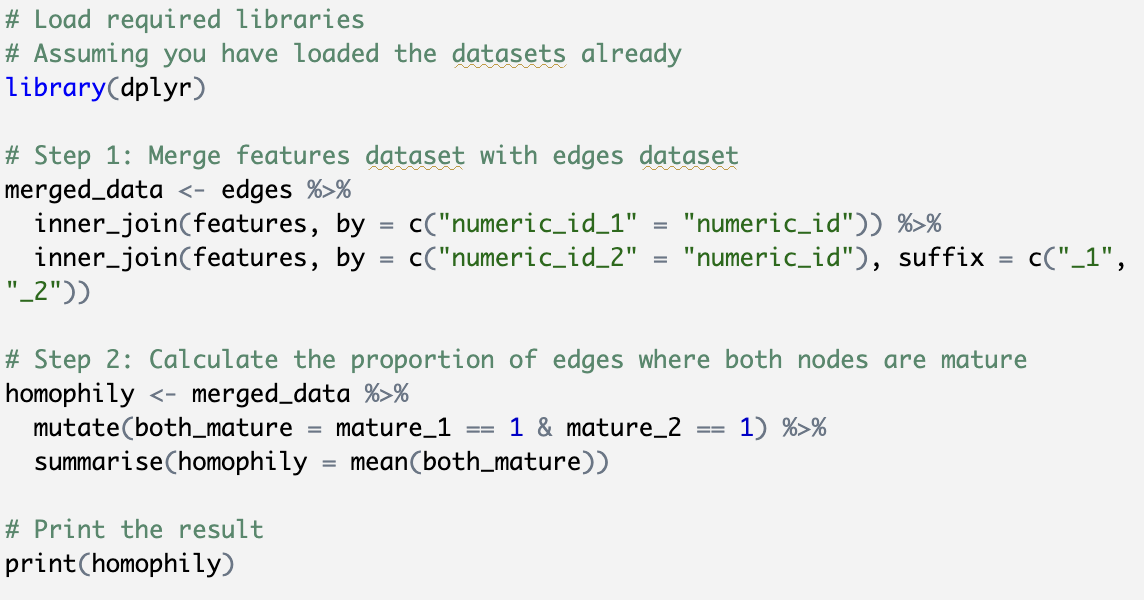
****

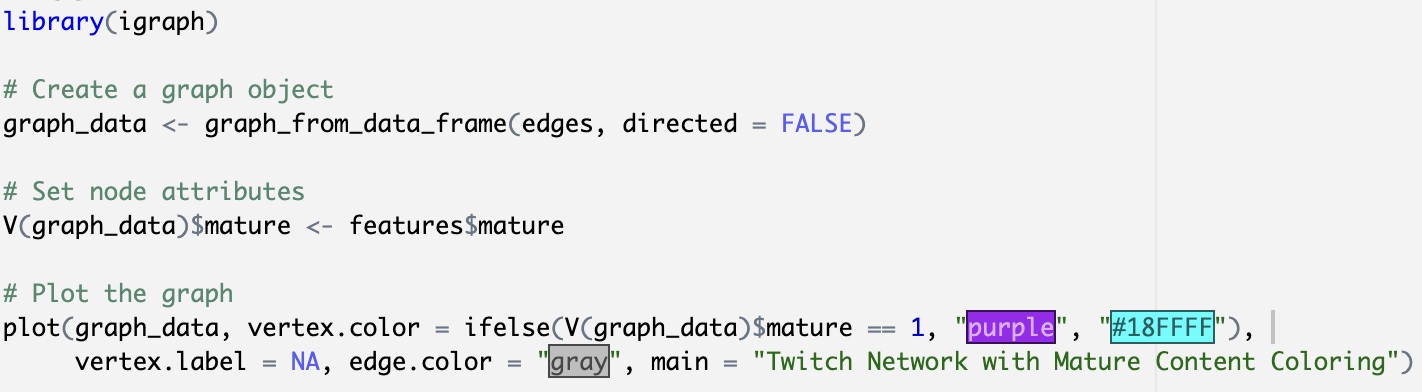
****

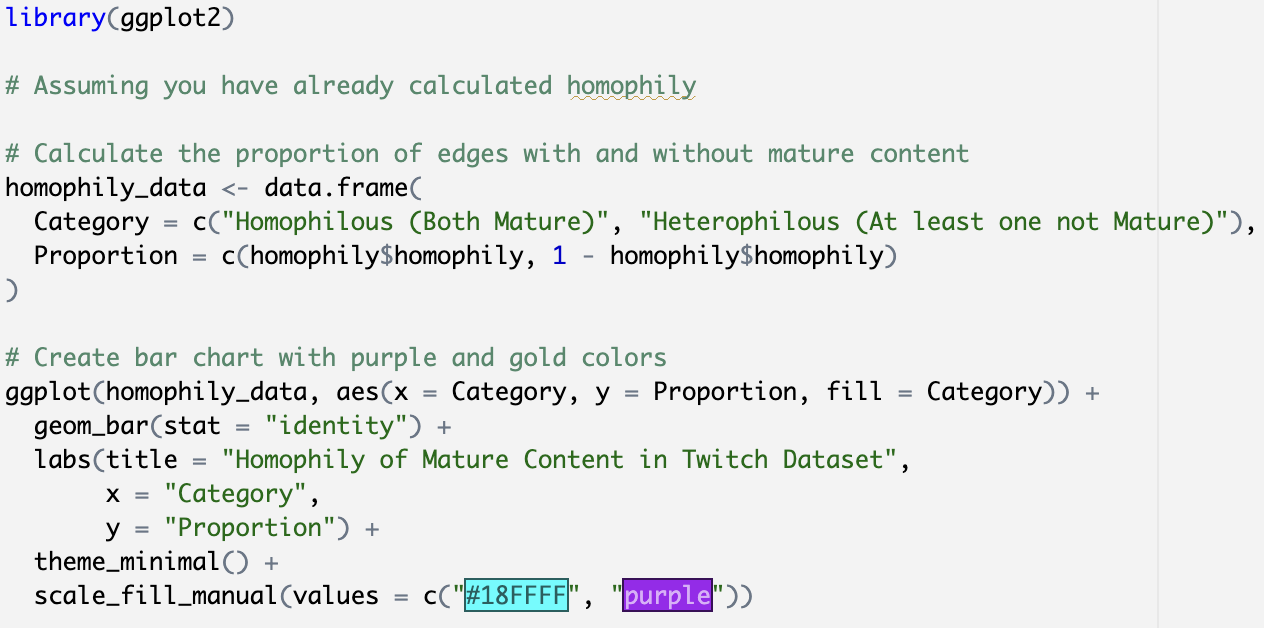




**Homophily of Mature & Language Features Code:**

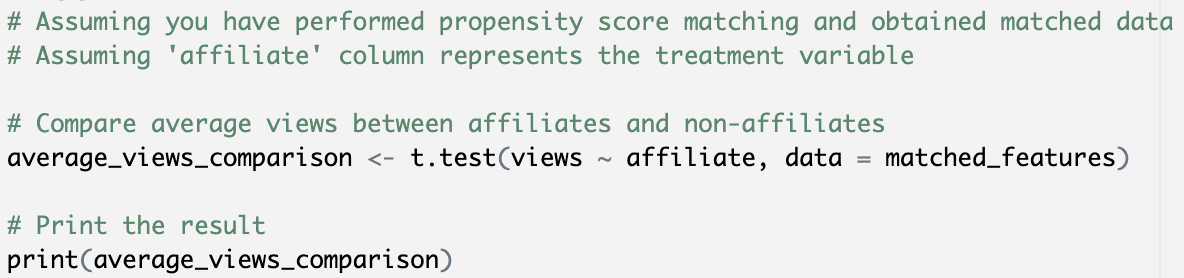






**Propensity Score Matching Code:**





**Homophily Top 5 Languages Code:**



