ISOM 673 Network Analytics

Final Project: Twitch Network Analysis

Yifei Wang

May 3, 2024

Introduction

With the exponential growth of online gaming and live streaming platforms, such as Twitch, understanding the dynamics of user interactions within these digital ecosystems has become increasingly pertinent. In this context, social network analysis serves as a tool for us to uncover user connectivity and understand how users form ties. In this report, we will analyze the network of English-speaking Twitch streamers to understand the network structure, identify how top streamers interacts, and employ link prediction to suggest potential friend suggestions among users.

Background

The dataset being used is a Twitch user-user network of gamers who streams in English. Nodes are the users themselves and the links are mutual friendships between them. Vertex features are the streaming habits of each user, including:

- days: the number of days the streamer has been streaming.
- mature: whether the streamer uses explicit language.
- views: number of views the streamer has received.
- partner: streamers that participate in the Twitch Partner Program (those that are committed to streaming and could make earnings from streaming)

In addition, the streamers are uniquely identified by a user ID. We observe that there are 7,126 users in the dataset and the users make 35,324 directed ties. Roughly half of the users stream in explicit language, and majority of the users do not participate in the Twitch Partner Program. The

distribution of days is slightly right skewed, meaning that a mix of both established and newer streamers are within the network. On the other hand, the distribution of views is highly right skewed, suggesting that while most users have received a relatively low number of views, there are a few users who have a disproportionately high number of views, potentially indicating highly popular or viral streamers. The distributions are visualized in the below tables.

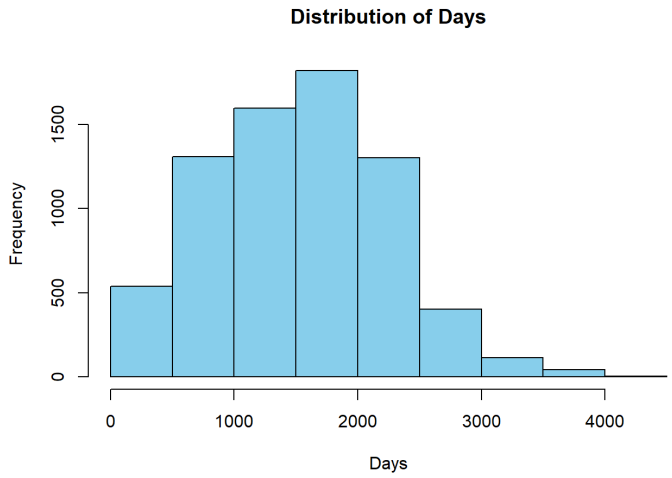

Table 1. Twitch streamer distribution based on days they have been streaming

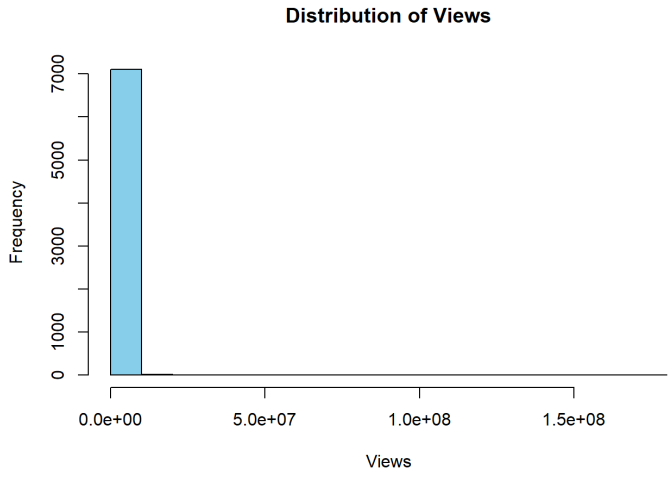

Table 2. Twitch streamer distribution based on views they have received

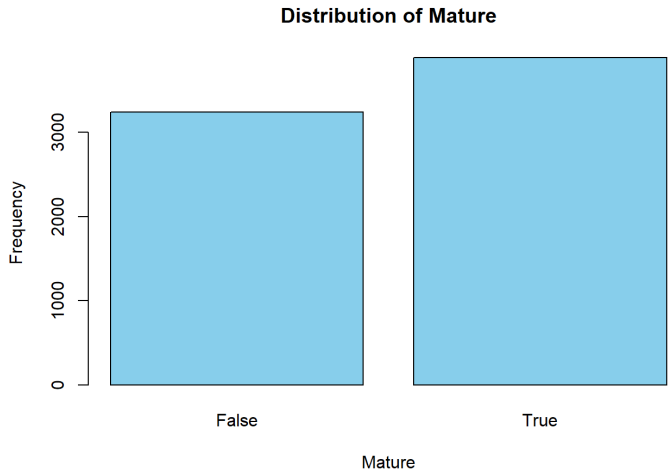

Table 3. Twitch streamer distribution based whether they stream in explicit language

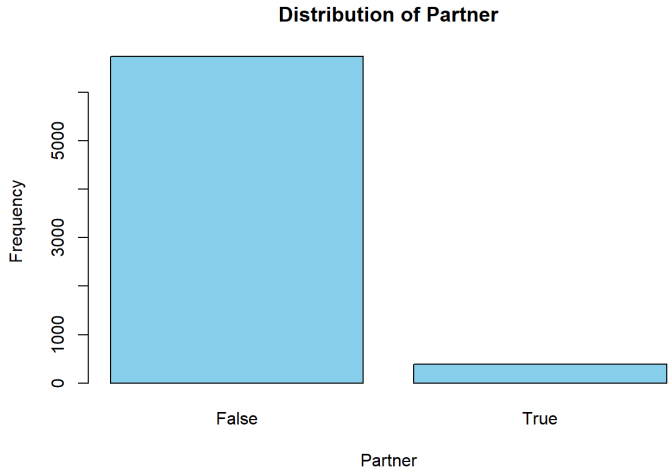

Table 4. Twitch streamer distribution based whether they are in the Twitch Partner Program

The analysis of the dataset will be two-fold. In the first part of the analysis, the focus will be on the node links to identify the network structure and the most central streamers based on the varying centrality measures. In the second part of the analysis, the focus will be on the user features and link prediction through the ergm package. By analyzing the network structure and central streamers, we would gain insights into how information spreads within the English

Twitch streamer network. Employing link prediction allows us to forecast the likelihood of connections forming between streamers who are not currently connected in the network, helping streamers forge new connections.

Analysis

The first step to data analysis was data preparation. The dataset was stored in two files, one containing the edges information from ID to ID, and the other containing the user features mentioned prior, uniquely identified by ID. After joining the two data files together based on ID, it was then converted to an igraph object. With all nodes and edges plotted, the network visualization becomes difficult to interpret. 10% of the total vertices was randomly sampled to alleviate this issue, and we observe that nodes are sparsely connected besides a group of densely connected nodes.

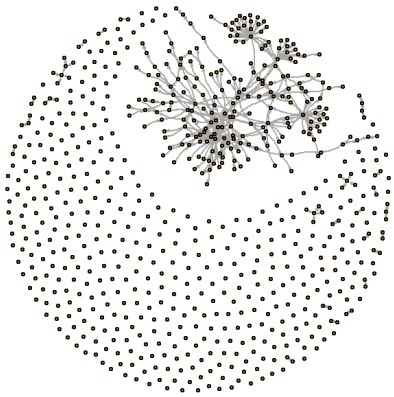

Table 5. Graph visualization of 10% randomly sampled nodes

Clustering through mclust on all data does not converge and clustering on sampled 10% data does not provide meaningful results (8 out of 9 clusters would consist of only 1 node). What this implies is that most Twitch streamers have sparse connections, but there exists a distinct subset of highly interconnected streamers who frequently collaborate with one another.

After understanding the network structure, we want to analyze the trends of top streamers, as identified by varying centrality measures (indegree, outdegree, closeness, and betweenness). We observe the top 10 streamers differ depending on the metric that is used and that the correlation between the 4 measures vary.

```
##              Indegree Outdegree Closeness Betweenness
## Indegree    1.0000000 0.4942637 0.4201114   0.7710909
## Outdegree   0.4942637 1.0000000 0.3724490   0.7298713
## Closeness   0.4201114 0.3724490 1.0000000   0.2638392
## Betweenness 0.7710909 0.7298713 0.2638392   1.0000000
```

Table 6. Correlation matrix for identified centrality measures

More specifically, indegree and betweenness are the most correlated while closeness and betweenness are the least correlated. This suggests that streamers who receive a high number of incoming connections (indegree) are also likely to act as intermediaries or bridges between other streamers in the network (betweenness). The low correlation between closeness and betweenness suggests that streamers that are well-connected to others (high closeness) are not necessarily streamers that act as intermediaries (high betweenness), and this makes sense given the general sparse nature of the network.

When looking closely at the network subgraph visualizations formed by the top 10 streamers for each centrality measure, we note that strong triadic closure is pronounced in the closeness subgraph, but not as much in the others.

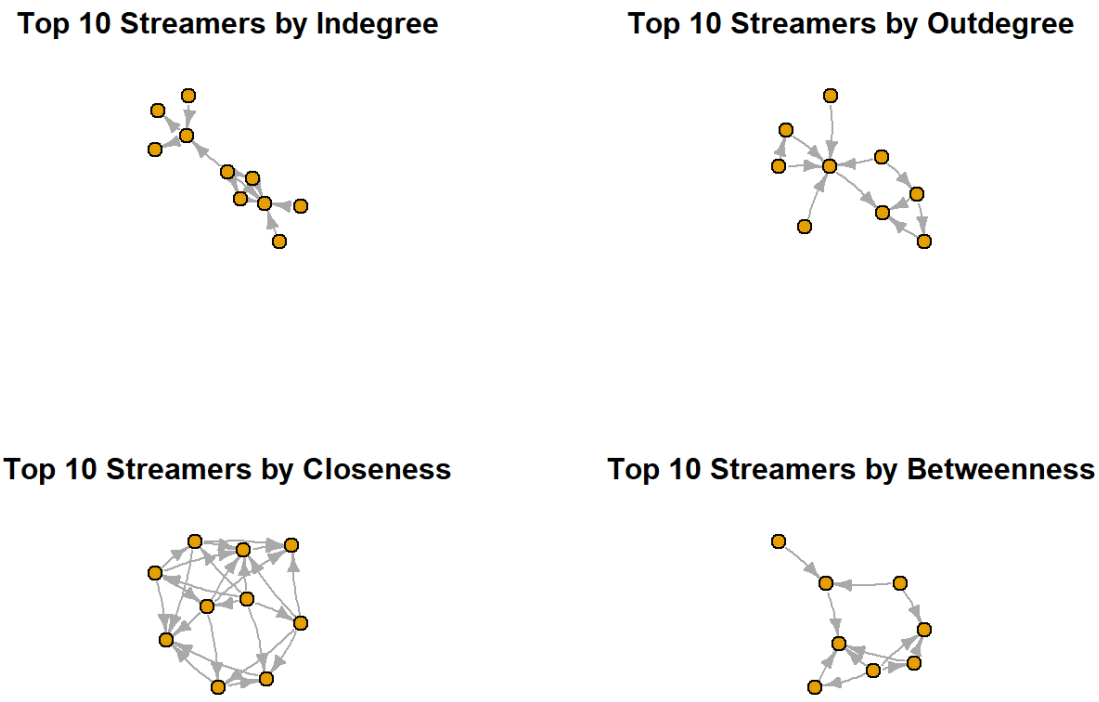

Table 7. Graph visualization of top 10 streamers for each centrality measure

This makes sense as the top streamers with high closeness centrality are users that are structurally close to many other users in the network, making it more likely for them to have frequent interactions with a core group, leading to the formation of strong triadic closures. On the other hand, top streamers as determined by betweenness are less likely to have pronounced strong triadic closure because these streamers act as intermediaries between other streamers in the network and may not necessarily be closely connected to each other. Similarly, streamers

with high indegree or outdegree centrality may be popular or active in reaching out to others, but they may not necessarily form ties with each other.

The second part of the analysis focuses on linked prediction. To do so, we utilize the ergm package and for each variable (days, mature, views, and partner), we assess its influence on tie formation. The function nodematch() is used for numerical variables (days and views) and nodefactor() is used for categorical variables (mature and partner). The output is in the below table:

```
## Call:
## ergm(formula = net ~ edges + nodematch("mature.x") + nodematch("mature.y") +
##     nodematch("partner.x") + nodematch("partner.y") + nodematch("views.x") +
##     nodematch("views.y") + nodematch("days.x") + nodematch("days.y"))
##
## Maximum Likelihood Results:
##
##                      Estimate Std. Error MCMC %  z value Pr(>|z|)
## edges               -7.299582   0.013071      0 -558.437   <1e-04 ***
## nodematch.mature.x   0.069102   0.010693      0    6.463   <1e-04 ***
## nodematch.mature.y   0.007788   0.010659      0    0.731   0.4650
## nodematch.partner.x -0.005490   0.010743      0   -0.511   0.6093
## nodematch.partner.y -0.003369   0.011071      0   -0.304   0.7609
## nodematch.views.x   -0.622738   0.256635      0   -2.427   0.0152 *
## nodematch.views.y    0.494777   0.327988      0    1.509   0.1314
## nodematch.days.x     0.115370   0.248859      0    0.464   0.6429
## nodematch.days.y    -0.268601   0.296318      0   -0.906   0.3647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 70385977  on 50772750  degrees of freedom
##  Residual Deviance:   584160  on 50772741  degrees of freedom
##
## AIC: 584178  BIC: 584319  (Smaller is better. MC Std. Err. = 0)
```

Table 8. Output of the ergm model

We observe that nodes with a high number of existing ties are less likely to form additional ties, and this makes sense given our earlier observation that network ties are sparse aside from a closely-knit group. It is also interesting to note that ties are more likely to be formed when the ‘from’ statchreamer uses explicit language (when mature is factor level 2, or true).

Beyond these two points, the other significant variable that influences tie formation is when the ‘from’ streamer is in the Twitch Partner Program. When the ‘from’ streamer is in the Twitch Partner Program, ties are more likely to be formed. However, there are no significant relationships observed when the "to" streamer is in the Twitch Partner Program. In other words, streamers who are part of the Twitch Partner Program are more likely to initiate connections with other streamers in the network, but whether the recipient of these connections is also a Twitch Partner Program member does not significantly influence tie formation.

Conclusion

In conclusion, the analysis of Twitch streamer networks has revealed several key insights. First, the network is characterized by sparse connections, with a distinct subset of highly interconnected streamers who collaborate with one another. Centrality measures such as indegree, outdegree, closeness, and betweenness highlight different sets of top streamers, indicating varying roles within the network. The observation of strong triadic closure among top streamers with high closeness centrality underscores the formation of tight-knit communities within the network. Moving beyond network structure, the analysis of tie formation suggest that existing ties and streamer characteristics from the streamer initiating the connection influence new tie formation.

Further exploration through the incorporation of tie strength analysis and longitudinal data using tools like RSiena could deepen our understanding of network evolution and the underlying processes driving tie formation. More specifically, we may represent tie strength through the number of times the two users have collaborated. Incorporating this information can help us better assess the node centrality and network structure, as well as improving the ergm models through weighted edges.