Twitch Social Network Analysis

Carolina Garma Escoffié
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida – Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán, México
Email: st1809073@upy.edu.mx

Didier O. Gamboa Angulo
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida - Tetiz
Tablaje Catastral 4448. CP 97357
Ucú, Yucatán, México
Email: Didier.gamboa@upy.edu.mx

***Abstract*—In this document the results of performing a network analysis on the Twitch Social Network from SNAP are presented. For computational purposes, the smallest subnetwork, the Portuguese Twitch Social Network, was chosen. The PT Network contains 1,912 nodes that are the users themselves and the links are mutual friendships between them. Vertex features are extracted based on the games played and liked, location and streaming habits. These social networks were collected in May 2018. A typical analysis containing distance and centrality measures as well as network visualization was performed. An undirected simple connected component with over 31,299 links and an average degree of 32 was obtained. The network follows a Barabasi-Albert model with a scale-free property that fails in the left-side of the distribution which decreased its gamma coefficient to 1.36 (which would model an anomalous regime). Finally, a community detection analysis was performed over a connected component sample of approximately 200 nodes to detect the targeted binary communities of the network. The Kernighan bisection algorithm resulted to be a good clustering method to detect the partner communities but failed when detecting explicit content members as its distribution is really sparse among this network, a greedy approach was performed over the whole network and 20 main communities were obtained, from which the 2 largest represent the great majority of the network, however, this binary partitioning did not match with the distribution of the target communities.**

Keywords—Twitch Social Network, Erdos-Renyi, Watts-Strogatz , Barabasi-Albert, Centrality Measures, Network Characterization, Network Models, Community Detection, Graph Visualization.

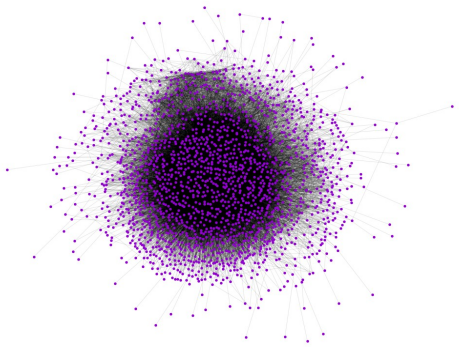


Figure 1. Twitch Social Network

I. INTRODUCTION

Twitch (also known as Twitch.TV) is a live-streaming platform owned by Amazon, Inc. whose main function is to stream live video games. In July 2011, Twitch launched its Partner Program, which reached 4,000 members by June 2013. Similar to the "Partner Program" of other video sites such as YouTube, it allows popular content producers to share in the revenue generated by advertising on their streams.

The network presented belongs to the Portuguese Speaking portion of the SNAP Twitch Social Networks, in which nodes are the users themselves and the links (undirected) are mutual friendships between them. Datasets share the same set of node features; this makes transfer learning across networks possible. Vertex features are extracted based on the games played and liked, location and streaming habits (days, views, partnership, mature content). These social networks were collected in May 2018.

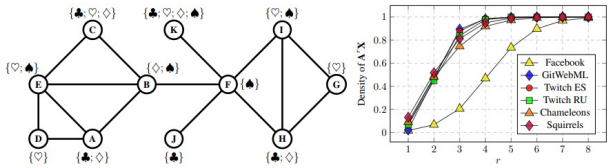
For this project, the related tasks involved exploratory analysis, visualization, and binary community detection to better understand the behavior and main characteristics of the network as well as to identify the principal communities such as the partners and the explicit content members. Other possible tasks would be link prediction, and transfer learning.

The possible applications of these tasks could be related to friendship recommendations or to identify potential “hub” or “influencer” members among the network to offer them the partnership program to ensure profits for both parties, since, as will be seen in the following sections, partner members are the most influential and far-reaching members of the entire network. Similarly, to detect possible violations of the platform's content and censorship policies, identifying members with explicit content is a good way to strategically monitor that the company is not violating legal streaming content regulations that could result in millionaire fines.

II. PREVIOUS WORK

Previous work on this dataset is presented in the paper "Multi-Scale Attribute Node Embedding" by B. Rozemberczki, C. Allen and R. Sarkar in 2019. Efficient unsupervised learning of node embeddings for large networks has presented a large growing over recent years. The main goal of the authors was to learn similar latent representations for nodes with similar sets of features in their neighborhoods, both on a pooled and multi-scale basis (Twitch network).

In Figure 2a attributed nodes D and G have the same feature set and their nearest neighbors also exhibit equivalent sets of features, whereas features at higher order neighborhoods differ. Figure 2b shows that as the order of neighborhoods considered (r) increases, the product of the adjacency matrix power and the feature matrix becomes less sparse. This suggests that an implicit decomposition method would be computationally beneficial.



(a) Attributed example graph. (b) Densification of the target matrix.

Figure 2. Phenomena affecting and inspiring the design of the multi-scale attributed network embedding procedure (Rozemberczki, 2019)

In overall, the authors investigated attributed node embedding and proposed efficient pooled (AE) and multi-scale (MUSAE) attributed node embedding algorithms with linear runtime. They proved that these algorithms implicitly factorize probability matrices of features appearing in the neighborhood of nodes. On several datasets (Wikipedia, Facebook, Github, and citation networks) they found that representations learned by their methods, in particular MUSAE (Twitch), outperform neighborhood based node embedding methods (Perozzi et al. (2014); Grover & Leskovec (2016)), multi-scale algorithms (Tang et al. (2015); Perozzi et al. (2017)) and recently proposed attributed node embedding procedures (Yang et al. (2015); Liao et al. (2018); Huang et al. (2017); Yang et al. (2018); Yang & Yang (2018)).

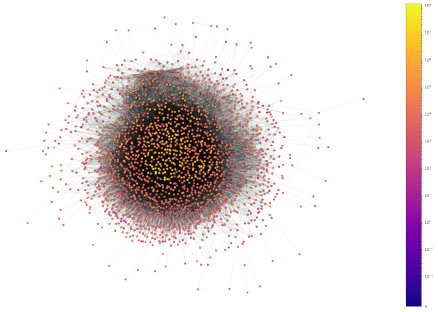


Figure 3. Twitch Social Network with node colors weighted by number of views.

III. NETWORK CHARACTERISTICS

The network is a simple giant connected component of 1,912 nodes with undirected links that represent friendship among members. The nodes have attached four inner attributes of the member they represent which are its partnership and mature content as booleans and the number of days and views the user have reached on the network as integers. In Table 1 the results of the principal metrics are attached.

TABLE I. NETWORK MEASURES

Measure	***Result***
Nodes	1,912
Links	31,299
Density	0.017
Average Clustering Coefficient	0.32
Transitivity	0.13

As seen in Table I, the network has a strong average clustering coefficient which also is supported by a good transitivity which translates into strong connections among members, it is probable that the neighborhood of a node would know each other. The density is low, something that is expected for real networks as most pairs of nodes are not directly connected.

TABLE II. DISTANCE MEASURES

Measure	***Result***
Average path length	2.53
Diameter	7
Radius	4
Peripheri Nodes	0.1 %
Center Nodes	14.91 %

Again, few are the nodes with a maximum eccentricity of 7 and almost 15% of the total have the minimum eccentricity of 4. Supported by the average shortest path length, this table reinforces the statement of the Small-World Phenomenon for real social networks.

TABLE III. CENTRALITY MEASURES

Measure	***Top Score***	***Node***
Degree Centrality	0.40	127
Closeness	0.60	127
Betweeness	0.10	127

IV. CENTRALITY MEASURES

As seen in Table III, we have that the top scores for each measure all belong to one single node, the 127. As the data collected was anonymous, we cannot be sure of who was this streamer, but as the data is from 3 years ago, maybe this node could belong to someone in the top 10 of current Portuguese Streamers.

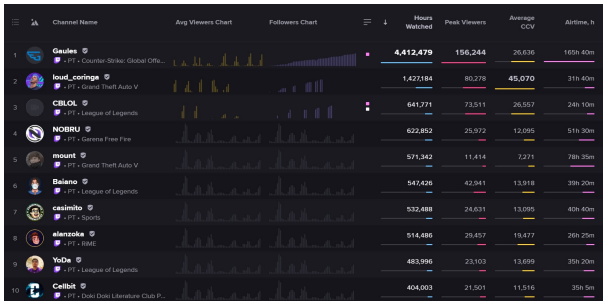


Figure 4. Top 10 Twitch Portuguese Streamers in 2021

Going back to the centrality measures, node 127 turned to be the most important node according to its degree centrality, it is directly connected to the 40% of the total Portuguese network. It is also the most important node according to the reachibility distance to all the other members (closeness centrality), and finally is also the most important given that is the one with the highest number of connections among other members that involve this node as intermediary (betweness centrality). In Figure 5, is easy to see that node 127 (the most in yellow) is located near the center given its higher centrality values and we can also see that belongs to the group of nodes with the highest number of views (Figure 3).

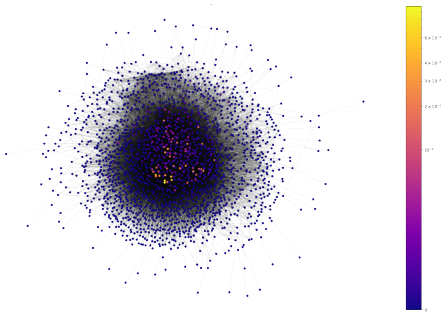


Figure 5. Betweenness Centrality on a sym-log scale

We can also see that the behavior of the remaining central nodes is similar, members with higher number of views (or top streamers) tend to be allocated in the center of the network as they are the ones with the highest centrality scores. See Figures 5, 6 and 7.

TABLE IV. DEGREE MEASURES

Measure	***Result***
Min degree	1
Max degree	767
Avg degree	32

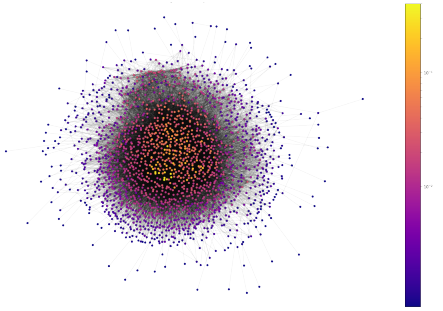


Figure 6. Degree Centrality on a sym-log scale

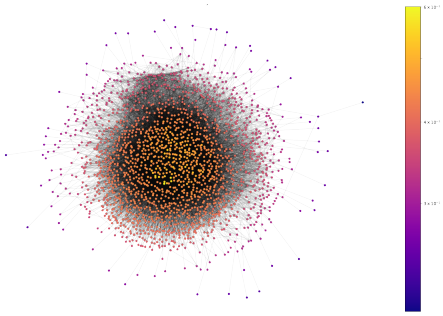


Figure 7. Closeness Centrality on a sym-log scale

V. DEGREE DISTRIBUTION AND NETWORK MODELS

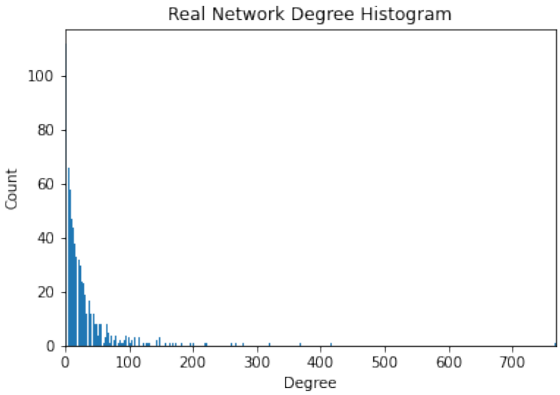


Figure 8. Degree distribution histogram

TABLE V. DEGREE PERCENTILES

Percentile	***Degree***
25	7
50	17
75	36
95	114

The degree distributions of the network cover from nodes of degree 1 to nodes of degree 767, an extensive range at first glance. However, as seen in Table V, 95% of the network includes degrees less than 114, therefore the distribution is right-tailed, most node degrees are relatively small compared to the range of degrees.

The hypothesis proposed here is then that the network follows a Barabasi-Albert model. To test this hypothesis, we will proceed to compare the network with the three main models seen in class: Erdos-Renyi, Watt-Strogatz and Barabasi-Albert. We will proceed to calculate the parameters p (random and small-world) and gamma (scale-free) to visualize how this network behaves in the different models with the same average degree $<k>=32$.

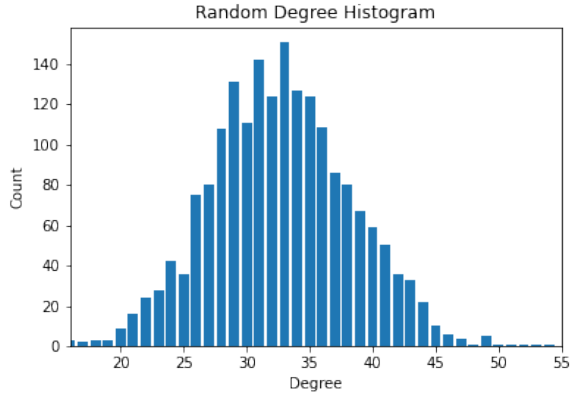


Figure 9. Random degree distribution histogram

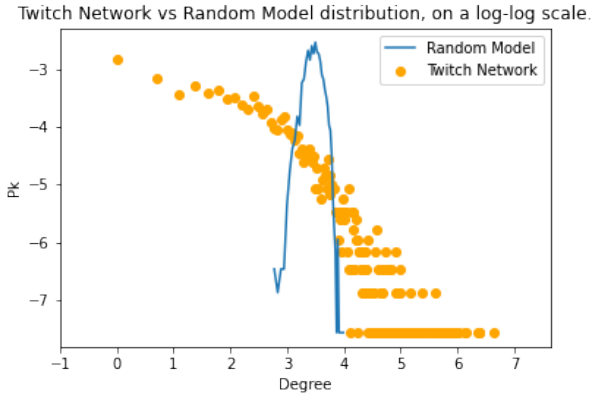


Figure 10. Random distribution compared to real distribution on a log-log scale

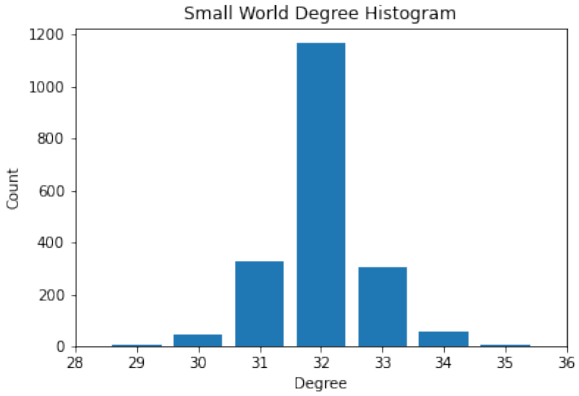


Figure 11. Small World degree distribution histogram

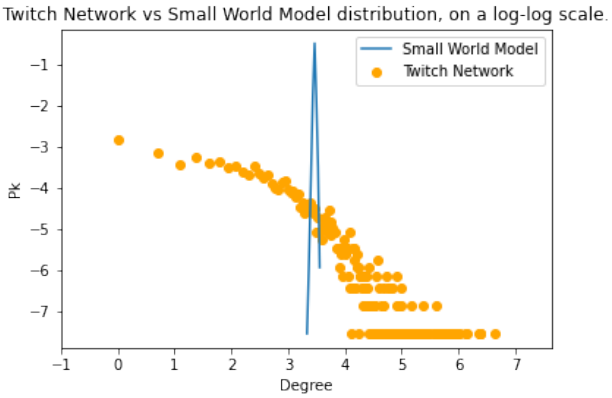


Figure 12. Small World distribution compared to real distribution on a log-log

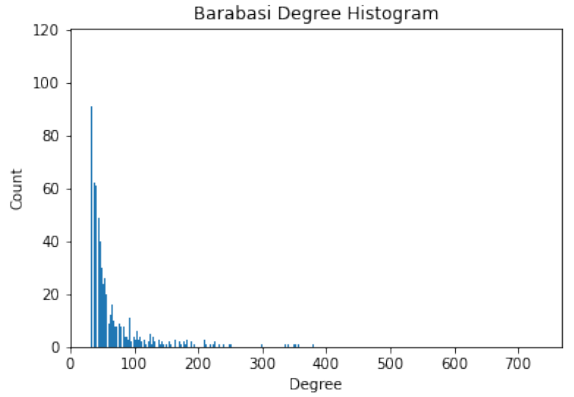


Figure 13. Barabasi-Albert degree distribution histogram

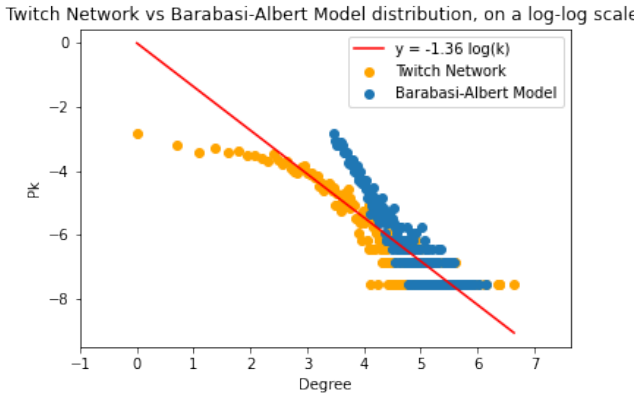


Figure 14. Barabasi-Albert distribution compared to real distribution on a log-log scale

A. Random (Erdos-Renyi) Model

We calculate the probability p of creating a link in the Erdos-Renyi model by $p = \langle k \rangle / (N - 1)$ obtaining $p = 0.017$. As seen in the Figure 9, the random model follows a normal distribution around the average degree, thus ignoring the high presence of smaller degrees in the original distribution as well as hubs.

In Figure 10, when plotting the DF of the random distribution over the original on a log-log scale we can see how the random model fails to adjust to the real behavior of the network.

B. Small World (Watt-Strogatz) Model

Using the probability p now as the wiring probability of the Small World model, we can see that it follows a Poisson distribution which is even narrower than the random model, almost 62% of the nodes are concentrated around the average degree meanwhile just 1% of the nodes have this same degree on the real network.

C. Scale Free (Barabasi-Albert) Model

As stated above, even though the average degree of the real network <k> is 32, it only represents 1% of the nodes in the original distribution, therefore using the average degree as input parameter is meaningless for the scale-free model, this means that when we randomly choose a node, we do not know what to expect: the selected node's degree could be tiny or arbitrarily large.

As Figure 12 indicates, the Poisson form offers a poor fit for the Twitch's degree distribution. Instead, on a log-log scale the data points follow a straight line, suggesting that the degree distribution of the real network is well approximating using a power-law distribution with degree exponent γ.

For calculating the γ that better fits our real distribution, we use the following formula:

$$\gamma = 1 + \frac{1}{\sum ln \frac{d}{min(d)}}$$

We obtained γ = 1.36 which based on the ultra-small world theory indicates that the graph belongs to an anomalous regime. For most real systems, the degree exponent is $2 < \gamma < 3$, for $\gamma < 2$, the exponent $1/(\gamma - 1)$ in the power-law equation is larger than one, hence the number of links connected to the largest hub grows faster than the size of the network. This means that for sufficiently large N the degree of the largest hub must exceed the total number of nodes in the network, hence it will run out of nodes to connect to. This, in words of the Twitch network, would mean that after some time, there will be nodes who will end up having friendship with the other N-1 members of the network, and even then, they would not stop growing, however, real networks are the result of a growth process that continuously increases N. Therefore, in reality, the network can be fitted by the Barabasi-Albert model in which growth and preferential attachment, play a particularly important role in shaping the network's degree distribution.

Finally, here are presented some distances measures of each distribution compared to the Twitch network.

TABLE VI. NETWORK MODELS METRICS

Model	*Avg Path Length*	*Clustering Coefficient*	*Avg. Degree*	*Gamma*
Twitch Network	2.53	0.32	32	1.36
Random	2.54	0.02	32	-
Small World	3.86	0.69	32	-
Barabasi-Albert	2.10	0.08	32	2.98

Random network asserted in the average path length but failed in the degree distribution and clustering coefficient, the WS model captured the small world characteristics, but not the degree distribution, the BA model captured the degree distribution, at least approximately, and the average path length, but not the clustering coefficient.

VI. COMMUNITY DETECTION

A binary node detection task was performed to identify two target communities among the Twitch network: partner members and mature content members. Two different algorithms were used, the Kernighan bisection over a connected component sample of 200 nodes, and a greedy approach over the whole dataset.

A. Twitch Partnership

As previously mentioned, Twitch partner members are those with a potential reachability over the whole network, therefore it is expected that these members would be in the center of the graph as observed in Figure 15. In addition, Figure 16 shows this same characterization but now weighting the node sizes according to the number of views and is easy to see that the number of views of this group, which only represents the 15% of the total members, represent 96% of the total views.

TABLE VII. PARTNER COMMUNITY

Measure	*Partner Members*
Percentage of total views	96 %
Percentage of total members	15 %
Percentage of mature content members among the partner community	36 %

TABLE VIII. MATURE CONTENT COMMUNITY

Measure	*Mature Members*
Percentage of total views	17 %
Percentage of total members	35 %
Percentage of partner members among the mature community	16 %

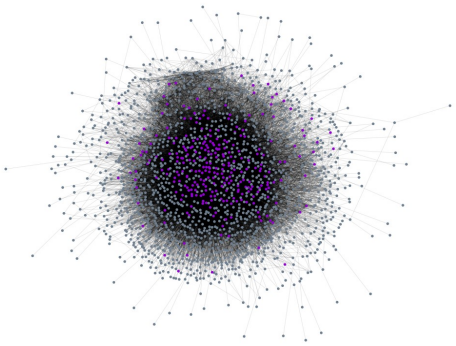


Figure 15. Twitch Partner members in purple and non-partners in grey.

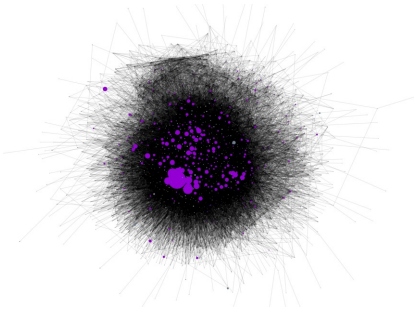


Figure 16. Twitch Partner members in purple and non-partners in grey weighted by number of views.

It is notorious now that the non-partner members are undistinguishable given its lower contribution to the number of views in the network.

For the following subsection, the Kernighan approach was performed over a connected component sample of approximately 200 nodes.

1) Kernighan Bisection

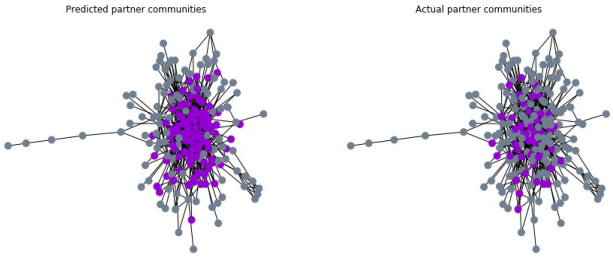


Figure 17. Predicted partner bisection (left) and actual partner bisection (right)

Figure 17 illustrates that the Kernighan bisection approach had a good outcome to identify the partner community, it identified 74% of the partner members correctly, even though the bisection was not perfect, a tendency for classifying the most central nodes as partner members is noticed.

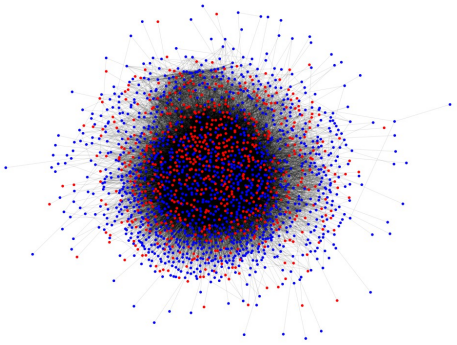


Figure 18. Mature content members in red and non-mature in blue.

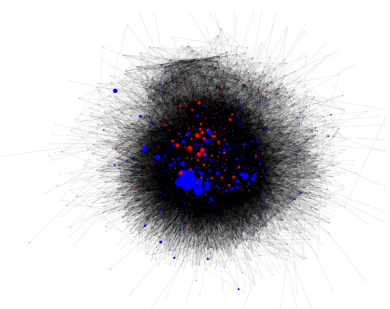


Figure 19. Mature members in red and non-mature in blue weighted by number of views.

B. Mature Content Members

Mature content members are those with a potential use of explicit language in their streaming, in Figure 18 we can see that the distribution of this community among the network is really sparse. Figure 19 shows a slightly predominance of non-mature content for the nodes with the highest number of views.

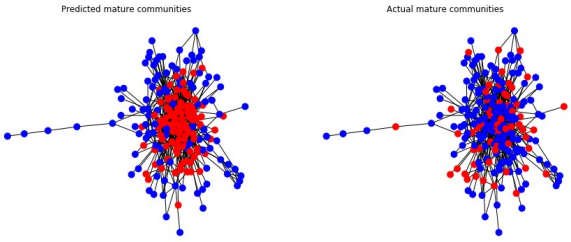


Figure 20. Predicted mature bisection (left) and actual mature bisection (right)

Figure 20 illustrates that the Kernighan bisection approach failed to identify the mature community, as it is really sparse among the network, even between partner members.

II) Greedy Algorithm

For this subsection, a greedy approach over the whole dataset was performed. The partitions resulted in 1,912 nodes divided among 20 communities from which the 2 largest represent almost the whole network, thus enhancing again that there is a natural binary partition of the network.

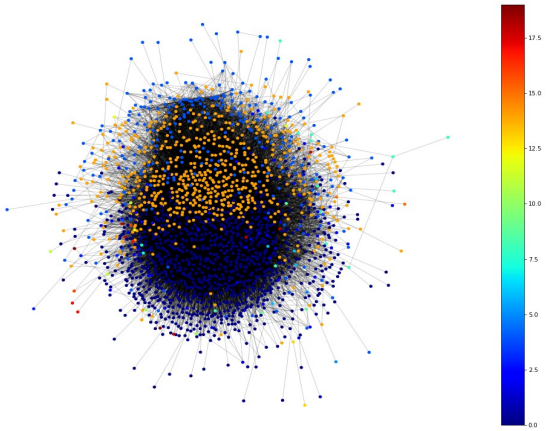


Figure 21. Greedy partitioning of Twitch's social network.

As seen in figure 21, there are two predominant groups among the network, however, neither the partner nor the mature communities follow the bisection obtained from the greedy algorithm. This new binary partition could reveal the two main types of games viewed, or they could even be telling us that we have two main locations in which Portuguese is spoken such as the Brazilian and Portuguese community. It is hard to tell what this partition means without more context about it.

VII. CONCLUSIONS

A successfully network analysis was performed from scratch. The initial hypothesis of the Twitch Network following a Barabasi-Albert Model got acceptable results but the scale-free property failed in the left-side of the distribution generating a gamma coefficient of 1.36, putting the distribution inside an anomalous regime, this part was really unexpected for me and I had to research a little bit more about what does it mean to be inside an anomalous regime, which actually turned to be that this network could not exist. For most real systems, the degree exponent is $2 < \gamma < 3$, for $\gamma < 2$, the exponent $1/(\gamma - 1)$ in the power-law equation is larger than one, hence the number of links connected to the largest hub grows faster than the size of the network. This means that for sufficiently large N the degree of the largest hub must exceed the total number of nodes in the network, hence it will run out of nodes to connect to. This, in words of the Twitch network, would mean that after some time, there will be nodes who will end up having friendship with the other N-1 members of the red and even then, they would not stop growing, however, real networks are the result of a growth process that continuously increases N. Therefore, I strongly believe that the Twitch network can be fitted by the BA model in which growth and preferential attachment, play a particularly important role in shaping the network's degree distribution that also explains the existence of hubs that turned to be the most influential streamers on this platform. The only measure that none of the models captured about the real network was the clustering coefficient of 0.31, the BA and the Random model underfitted the result and the WS overfitted it to more than the double (0.69), this result was similar to the clustering coefficient of the Facebook network (0.61). If we compare the average clustering of these two social networks, we could easily say that the probability of a moderate connected neighborhood in Twitch is just the half of the probability of seeing the same phenomena in Facebook, or in more simply words, Facebook users are twice as connected compared to Twitch users.

Regarding the community detection tasks, I had really low expectations about the results as the complexity time of two out of three algorithms seen in class would have troubles working on the whole dataset and even when working with small samples. However, we had a moderate success over the sample network when detecting partnership or potential influencers using the Kernighan bisection algorithm which successfully discerned among central and peripheral users in the network. Further work would be replicating a similar multi-scaled embedded node attribute analysis to detect complex communities as the mature content one. In the case of the greedy partitioning, neither the partner nor the mature communities follow the bisection obtained from the greedy algorithm. This new binary partition could reveal that perhaps we are lacking context to identify these communities that could be different gaming styles or even location-based communities as Portuguese is spoken in at least two different continents.

REFERENCES

[1] Barabási, Albert-László (2016) Network Science. USA. Chapters 4 and 5.

[2] Barabási-Albert model. (2021, May 1). Retrieved July 3, 2021, from https://eng.libretexts.org/@go/page/46602

[3] B. Rozemberczki, C. Allen and R. Sarkar. Multi-scale Attributed Node Embedding. (2019). Retrieved July 3, 2021, from https://www.researchgate.net/publication/336147076_Multi-scale_Attributed_Node_Embedding

[4] Heavy-tailed distributions. (2021, May 1). Retrieved July 3, 2021, from https://eng.libretexts.org/@go/page/46601

[5] Mark Newman (2018) Networks: The empirical study of Networks, UK: Oxford University Press

[6] Reza Zafarani, Mohammad Ali Abbasi, Huan Liu (2014) Social Media Mining: An Introduction., UK.

[7] Twitch Social Networks. (2021). SNAP. Retrieved July 3, 2021, from https://snap.stanford.edu/data/twitch-social-networks.html

[8] Twitch Top Streamers. (2021). TwitchTracker. Retrieved July 3, 2021, from https://twitchtracker.com/channels/ranking/portuguese