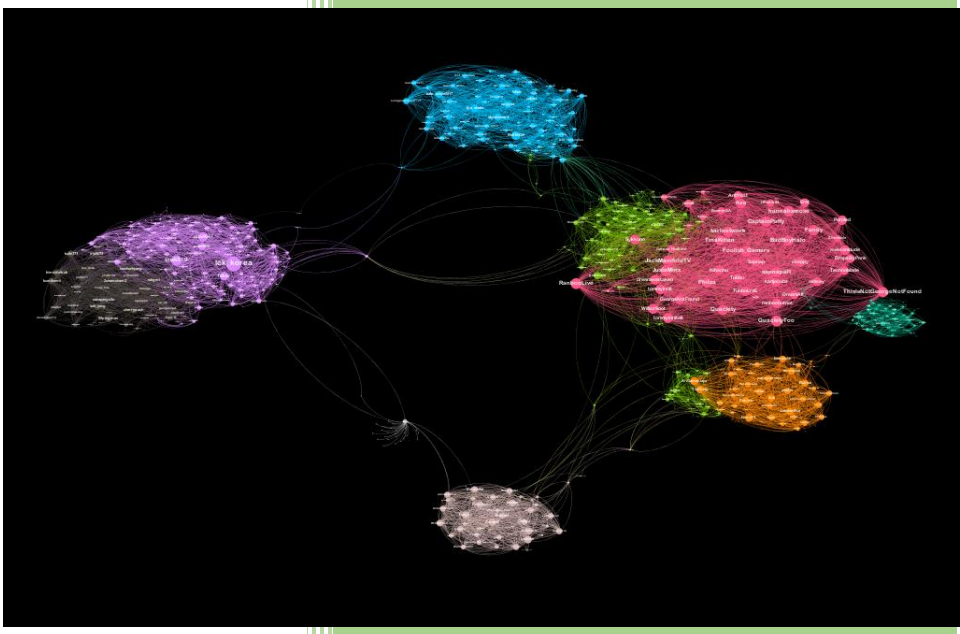
2021

Twitch Streamer Network Analysis



Bryan Huynh 30041522

CPSC 572

12/7/2021

Abstract:

Twitch being one of the largest streaming platforms with millions of viewers and thousands of content creators. Twitch has a variety of different content streamers that are split up into different categories. These categories include Games, IRL(In Real Life), Music, etc. We have collected data about users and streamer follow data to identify how followers of one twitch streamer are connected to another twitch streamer. With this data, our goal is to analyze the probability of which genre is most likely to share users. We used a variety of tools to visualize the twitch streamers and the users that connect to each of them.

Introduction:

Twitch is a live streaming platform that is managed by Twitch Interactive, a subsidiary company of Amazon. Originally created as Justin.tv where users could broadcast video streams of any subject. In 2011, Twitch was created to handle the gaming category of Justin.tv. Twitch's focus is to stream video games, which includes eSports, music broadcast and other forms of creative entertainment with the most recent edition being IRL(In Real Life) content.

Twitch platform has grown significantly over the years from its creation in 2011 and is now considered the biggest streaming platform. With more than 2 million concurring views monthly who watch their favourite streamer play games and interact with them via twitch's chat function.

Through the means of the chat function streams have direct interactions with their viewers. Users of the platform can follow the streamers to show interest in their content, subscribing by paying a monthly fee or users can donate money to the streamers directly to get a more immediate interaction. Although viewing the stream is completely free, users tend to donate or subscribe to show their support.

The focus of this project is to understand the underlying structure and interactions that exist between streamers. Specifically what links streamers together to create these communities. Other publications on the topic of twitch mainly focused on the users, and what the users' interactions looked like; for example, what type of combination of categories do people most often watch. I could not find any other publications that focused on streamers to streamer network analysis

Dataset Description:

The data was collected during the month of October using twitch's API and a web scraping tracker website to gather information on streamers. The gathering of data would start at a single twitch streamer, it would then find all other streamers who had mutual followers where the connection was more than N mutual followers and had a large standing on twitch [300,000 + followers]. A recursive process was applied to the new set of streamers and the process would continue for N steps. The result of this extraction would occur multiple times on different starting streamers to gather clusters of different communities. These included distinct types of games/categories, languages, and regions. We gathered the information for different communities from North American, South American, European, and Eastern streamers.

These data sets were then combined to form a single data set that would contain, Source, Target, and Weight. The Source would be some streamer, Target would be a streamer who had mutual followings with the source, and Weight would be the count of how many followers they found in common. This type of network was a directed network. Following this, we took all the streamers and web scraped more information about them like their top 5 games/categories and spoken language of the streamer.

The project can be found at https://github.com/BryanHuynh/Twitch-Streamer-Network

Basic Statistics:

	Twitch Network	ER Network	DP Network
Nodes	831	830	831
Edges	7850	6245	7850
Connected Components	1.0	1.0	1.0
Clustering Coefficient	0.33	0.03	0.01
Average shortest path	2.69	0.88	1.93

Degree distribution:

The null model I used was a degree preservation null model, and the random network is a Erdos-Renyi model

From the degree distribution in blue, we can see that it does not resemble the random network in green.

From the degree distribution we can see it generate a stretched exponential sublinear graph and not a linear power law graph

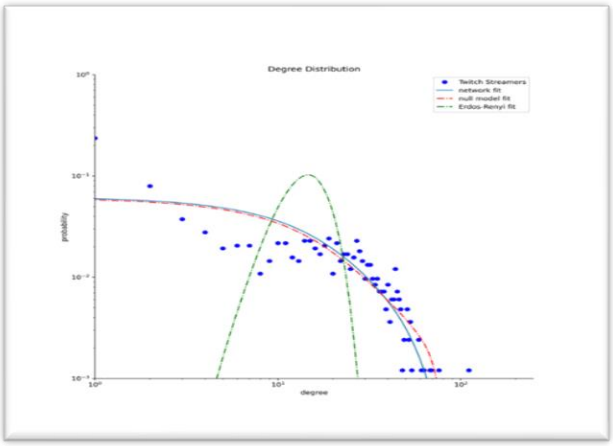


Figure 1 Degree Distribution

Clustering coefficient:

From the graph we can see that there is a high clustering coefficient

This is not replicated in the null model or random network so we can see that there is something unique in the graph.

We can see later that this immerges from the tightly knitted community structure present in this network

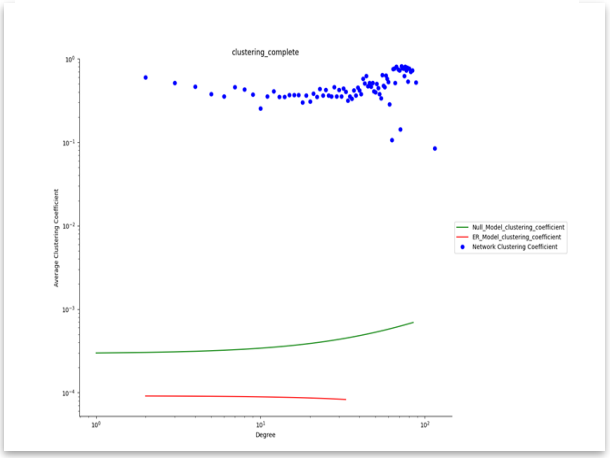


Figure 2 Clustering Coefficient

Modularity:

The question I want to ask is “what elements drives streamers to be in communities on Twitch?” Often the average twitch viewer watches streamers who all together have some similarity between all of them.

Category Type	Result
Greed community Detection algorithm	0.784
Based on Top Category	0.362
Based on Language	0.737
Based on Language and Top Category	0.409
Greedy community Detection on a Null Model	0.220

Using networkx’s greedy modularity function, the modularity value come in at 0.78. This is a very high score and expected considering that the clustering co-efficient for this network was also very high. Using Gephi to generate a graph [figure 1.1] and colourizing the communities, we can see a common theme of an overall subgraph community in the shape of an oval, and these community being split into smaller subgraphs of different communities. So, we can see that there is something that links a community together, but there’s also this additional element that separates them

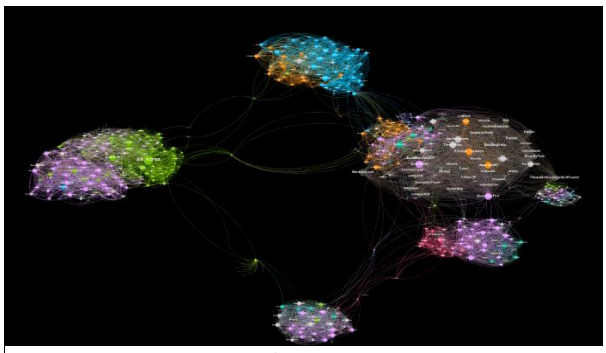


Figure 2.1. Greedy Community Detection

When we infer communities based on the streaming category the modularity value comes in at 0.41. So, we can say that the category that the streamer streams play a role in the community structure in the network. Looking at the figure 1.2, we can see that there are additional details. Take for example the blue, orange community at the very top. In the greedy community detection, it detected that this was a single solid community, but we can now say that this like many other communities has many parts. In addition we can also see many of the community divides in figure 1.1. replicated on 1.2, So the divides we saw must have been the categories. The value for M can be maximized if a greedy approach was applied where for each node, it would cycle through all their top games, instead of their top game, and found the one with the maximum M value, but I found that the M value here was sufficient for this finding.

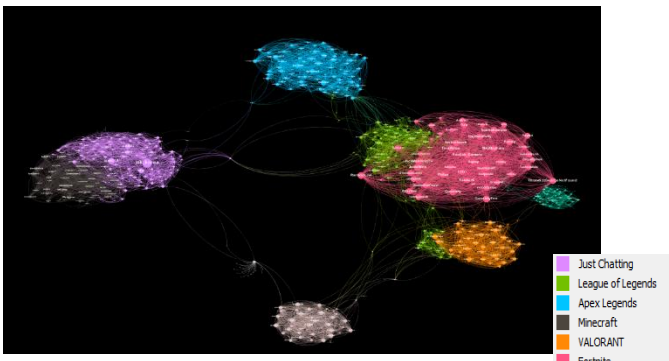


Figure 3.2. Top streamed category communities

When we infer communities based on the streamers spoken language the modularity value come in at a massive 0.73. So, we can see that a large set of communities can be setup around languages and that it plays a huge role in the structure of the community. Looking at figure 1.3 we can see that while it easily represents communities, we can also see that this method is also like casting a wide net. It loses many of the details, we saw in figures 1.1 and 1.2. But because of that we can also say that this must have been the element that sticks communities together that we in figure 1.1

Figure 4.3. Spoken Language category communities

Spread:

How well does information spread on twitch and what is the best method is spread information?

I ran a simulation of a spread network on the twitch network like the one discussed in class. In this model, there is a list of starting nodes that will be activated. Every node on this network has a chance of spreading once and being activated once. A node can influence another node if there is an out-going connection to the node. The probability of a node being influenced depends on the amount of mutual followers between them and the average mutual followers on an edge for the overall network. This method of probability mimics what I would expect information to cascade on twitch. A streamer says something memorable, to which the chat replicates it there and then goes onto other streamers to replicate the same meme there too. The chances of the meme catching on or spreading depends on the amount of people who hop over to the other stream and do it too.

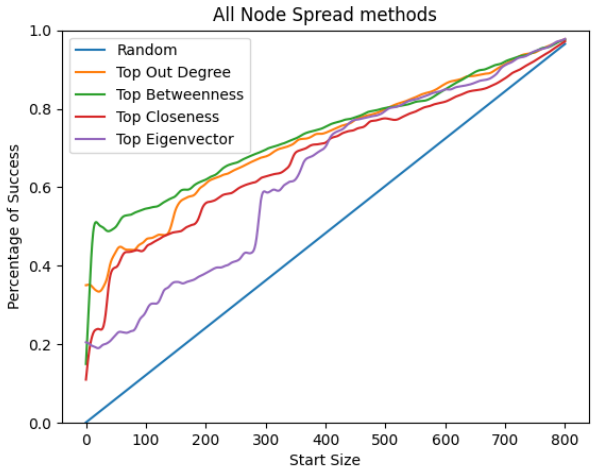


Figure 3.1. Spread. Starting node method comparison

I would have done a system, where a node has multiple chances of being influenced and every time a node tries to activate it, if it fails then the probability of it goes up for the next attempt but found that this easily made starting sizes of 1 have +90% spread.

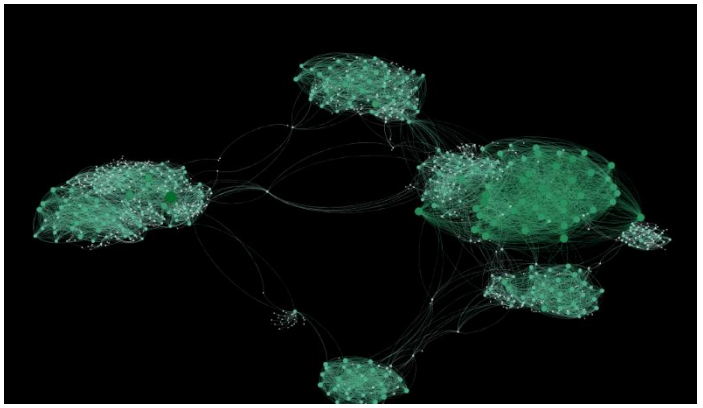


Figure 3.2 Out degree colourized

I wanted to understand what the best method was to maximize spread in this network. I took 4 node centrality measures and a method where the nodes added were random. For each analysis, I ran each starting size 5 times and averaged the overall spread.

What is interesting is that random node selection results in a linear line. I have no idea what is occurring here.

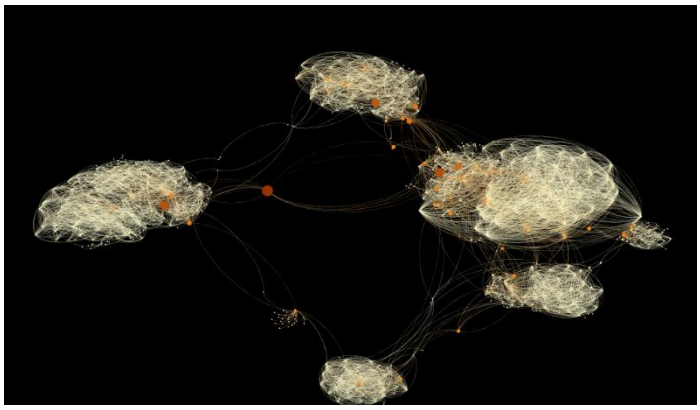


Figure 3.3 Betweenness colourized

There are 2 notable centrality measures that I want to point out. Looking purely at a spreading from a single node, we can see that top out degree [orange] has the highest at ~35%. What is interesting is the betweenness measure [green] where while its initial starting size has low spread, it quickly gains massive coverage when more nodes are added.

If we look at who these nodes are; looking at figure 3.2 we can see that there is a high concentration on high out degree nodes in the English and a few in the Korean community. I guess at what is going on is that because there is a high concentration in a few communities, these communities are almost guaranteed to have been reached. This would be like buses in a community. The buses would travel around the local community very effectively but its ability to spread to other communities isn't as strong.

Looking at figure 3.3, we can see that the saturated nodes all exist around the edges of communities, and that there seems to be at least one node in every community. This allows for it to not only spread to one community but also start affecting others that are in proximity. This would be akin to airports. This allows for spreading across different communities.

Top Betweenness Streamers

Streamer	Language	Top Category
loltyler1	English	League of Legends
Ludwig	English	Just Chatting
Faker	Korean	League of Legends
Euriece	English	Apex Legends
Valorant	English	Valorant
Aceu	English	Apex Legends
SolaryFortnite	French	Fortnite
Gaules	Portuguese	CS GO
Gumayusi	Korean	League of Legends

Top out degree streamers

Streamer	Language	Top Category
Lck_Korea	Korean	League of Legends
JustaMinx	English	Just Chatting
JackManifoldTV	English	Minecraft
Quackity	English	Minecraft
RanbooLive	English	Minecraft
Foolish_Gamers	English	Valorant
Hannahxxrose	English	Minecraft
CaptainPuffy	English	Minecraft
BadBoyHako	English	Minecraft

Taking an even closer look at who these streamers are, we can see that the nodes who have high betweenness [left table above] have a wide variety of languages and streaming categories, which we have concluded early are the essential blocks for Twitch's community structure. This is opposed to top out degree streamers [right table above] where we see a more limited set of languages and categories therefore limited number of communities.

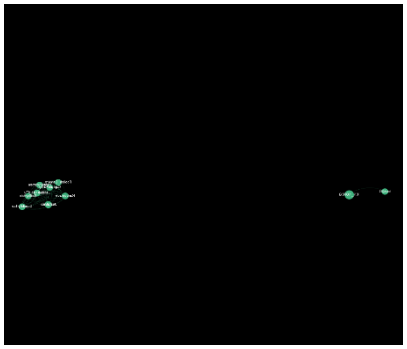


Figure 6.1 Top Degree Spread Timeline Initial

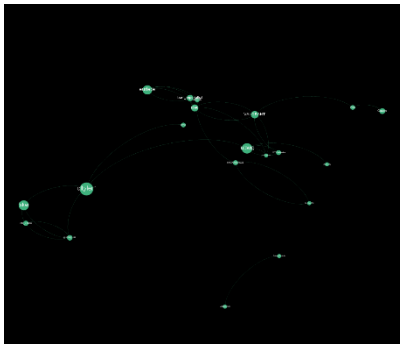


Figure 5.2 Top Betweenness Spread Timeline Initial

Looking at the end results of top degree (figure 4.1) spread we can see that they are concentrated in 2 areas, these are the English and Korean communities since they are where nodes have the highest concentration of high degrees.

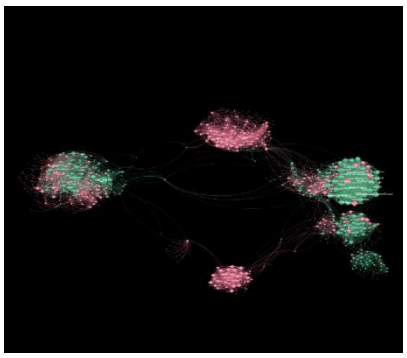


Figure 8.3 Top Degree Spread Timeline Final

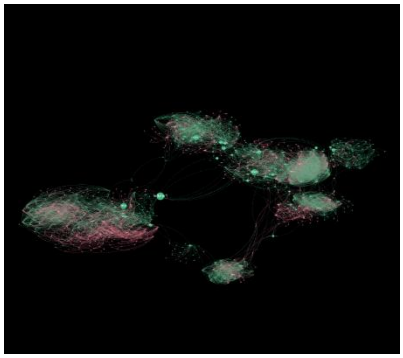


Figure 7.4 Top Betweenness Spread Timeline Final

As a result of this, the network (figure 4.3) we can see that both of those communities have been reached and have had a chance of being influenced. The top and bottom (Japanese and French communities respectively) were not reached. This follows my hypothesis originally that with sampling nodes with high degrees results in their respective communities being reached more easily. While nodes outside of those communities are less likely to even hear about it.

Looking at the initial state of top betweenness (figure 4.2), we can see that they are more spread out, and at least one node exists in all communities. As a result, in the final state (figure 4.4), we can see that there are nodes affected in every community. Surprisingly it would seem that the overall spread in each community results in almost the same amount covered. I believe that is a result that if at some point, a high degree node is spread to, it will result in the same spread as if it were starting on a high out degree node. In addition, Top betweenness completed in 11 iterations while top out degree completed in 9 iterations. Going further I would analyze, the number of iterations required to reach a minimum spread. This would allow us to understand what type of model would allow for quicker spread.

What makes these nodes have high betweenness?

Taking a closer look at the nodes with high betweenness, I wanted to understand why there can be connected to many communities.

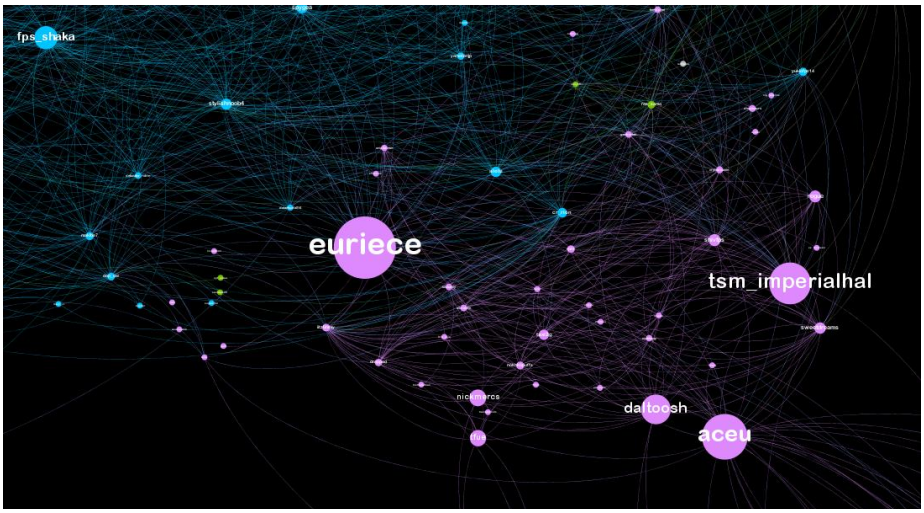


Figure 5.1

Looking at more closely at Euriece in figure 5.1, they are Filipino and found that they spoke English. So why are they able to be so close to the Japanese community? It turns out that they also play Apex legends, which is the most popular streamer category in the Japanese community, in addition they also can speak in Japanese as an addition language. So, being playing the communities most popular game and speaking their language allows you to connect to that community more easily. But there is also tsm_imperialhal, who is Caucasian and does not speak Japanese as a secondary language. In this case, they also play Apex Legends, but are also a professional player. So, it should suggest that speaking their language isn't a requirement, but if you are able to play the game professionally, they will watch you. The same is true for Aceu and Daltoosh, who do not speak the language, but are professional players. This seems to be the case for most of the streamers with high betweenness. Either they also speak that language, or they play the communities game professionally or very well.

The Null Model

The null model that I used in my project was a degree preserving null model. I took the twitch network and grabbed every node and added stubs to maintain the number of in and out degrees. I generate 100 models that were used in this project where I took the average of all of them. I choose to do a degree preserving null model over a configuration model because I wanted to maintain the number of in and out degrees in addition to my real network not having any self-loops and multiple edges.

The Random Model

The random model I used in this project was the Erdos-Renyi model. I chose to use this model because I wanted to have a model that would clearly indicate a lack of community structure to contrast with my network. If I were to do this again, I would choose to use a Barabasi-albert model, to compare my network more closely to one that follows preferential attachment and a scale free model.

Conclusion

When analyzing streamer to streamer relationships, we can see that there is huge emergence of multiple community structures. Amongst the factors that structure these communities, common language and streaming category are the main traits links streamers together. From our spread analysis we can also see that there are also nodes who have traits of multiple communities that allows them to be apart of both worlds. In addition, because of it, allows them to be great instruments for spreading information on the network.