Advanced Analytics in a Big Data World

Assignment 4

Ana Sofia Mendes - r0925549
Chris Butcher - r0918809
Federico Soldati - r0924528
Guilherme Consul Soares de Bem - r0917829
Ishika Jain - r0915387
Sounak Ghosh - r0914328

Group_17

MSc Statistics and Data Science

May 28th, 2023

Contents

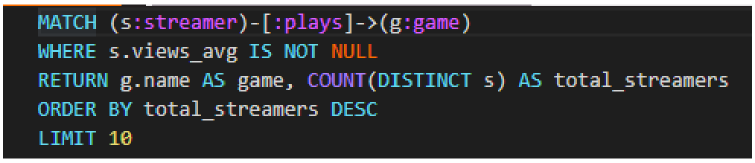
1 Introduction

Twitch is a popular live streaming platform that primarily focuses on video game streaming. It allows users, known as streamers, to broadcast their gameplay and other forms of creative content in real-time to a global audience. Viewers can watch these live streams and interact with streamers through chat and other features. It has diverse content and has become a thriving community for streamers and viewers. It offers customization options for streamers, monetization opportunities, and features to enhance the viewing experience. Twitch has grown into a significant platform in the streaming industry, providing a space for streamers to showcase their talents and connect with a worldwide audience of viewers who share similar interests.

The dataset that is used for this assignment is a graph which is extracted from Twitch. The aim of this project is to analyse the dataset and try to find interesting information, using Cypher and Gephi.

2 Most popular games

We start the analysis by finding the 10 most popular games. In order to find that, we will base their popularity on the number of streamers who play the game. We can obtain these results with the following Cypher query:



```
MATCH (s:streamer)-[:plays]->(g:game)
WHERE s.views_avg IS NOT NULL
RETURN g.name AS game, COUNT(DISTINCT s) AS total_streamers
ORDER BY total_streamers DESC
LIMIT 10
```

The top 10 games are:

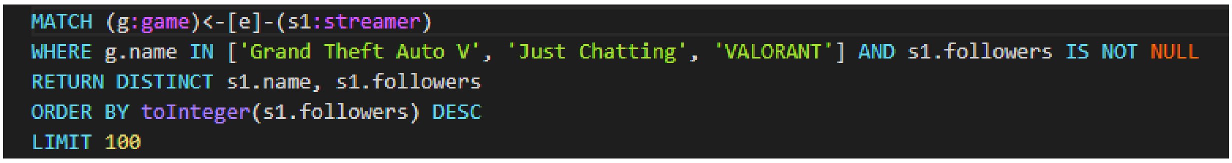
Game	No. streamers playing
Just Chatting	4648
VALORANT	1544
Grand Theft Auto V	1538
League of Legends	1387
Fortnite	962
Minecraft	944
Escape from Tarkov	695
Call of Duty: Warzone	694
Counter-Strike: Global Offensive	686
Apex Legends	686

Table 1: Top 10 games by no. followers

We see that 'Just Chatting' is by far the most used by streamers on Twitch. Note that this does not refer to the name of a game. 'Just Chatting' allows many streamers to talk to their viewers without having to game or do a specific activity while streaming. Most 'Just Chatting' streamers have special guests or host exciting events to further boost their stream.

Considering the top 3 games (*Just Chatting*, *VALORANT*, *Grand Theft Auto V*), we next see who are the top 100 streamers playing these games. The popularity of a streamer is based on the number of followers.

We achieve this with the following query:



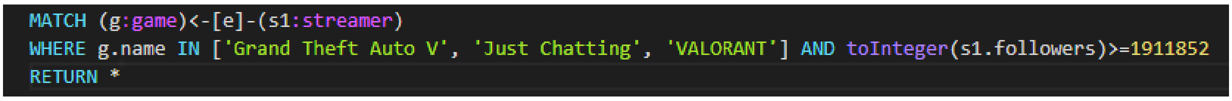
```
MATCH (g:game)<-[e]-(s1:streamer)
WHERE g.name IN ['Grand Theft Auto V', 'Just Chatting', 'VALORANT'] AND s1.followers IS NOT NULL
RETURN DISTINCT s1.name, s1.followers
ORDER BY toInteger(s1.followers) DESC
LIMIT 100
```

Where we get the following streamers:

Ranking	Streamer	No. followers
1	auronplay	15153321
2	rubius	14017799
3	ibai	13035147
...	...	...
99	pow3r	**1928373**
100	thedanirep	**1911852**

Table 2: Top 10 games by no. streamers playing

Using the number of followers of the 100th streamer as a lower boundary, we can obtain the graph of the top 100 streamers playing the top 3 games. That is given by the following query:



```
MATCH (g:game)<-[e]-(s1:streamer)
WHERE g.name IN ['Grand Theft Auto V', 'Just Chatting', 'VALORANT'] AND toInteger(s1.followers)>=1911852
RETURN *
```

With the .json file resulting from the query, we can display the graph using Gephi. The final result can be visualized in Figure 1:

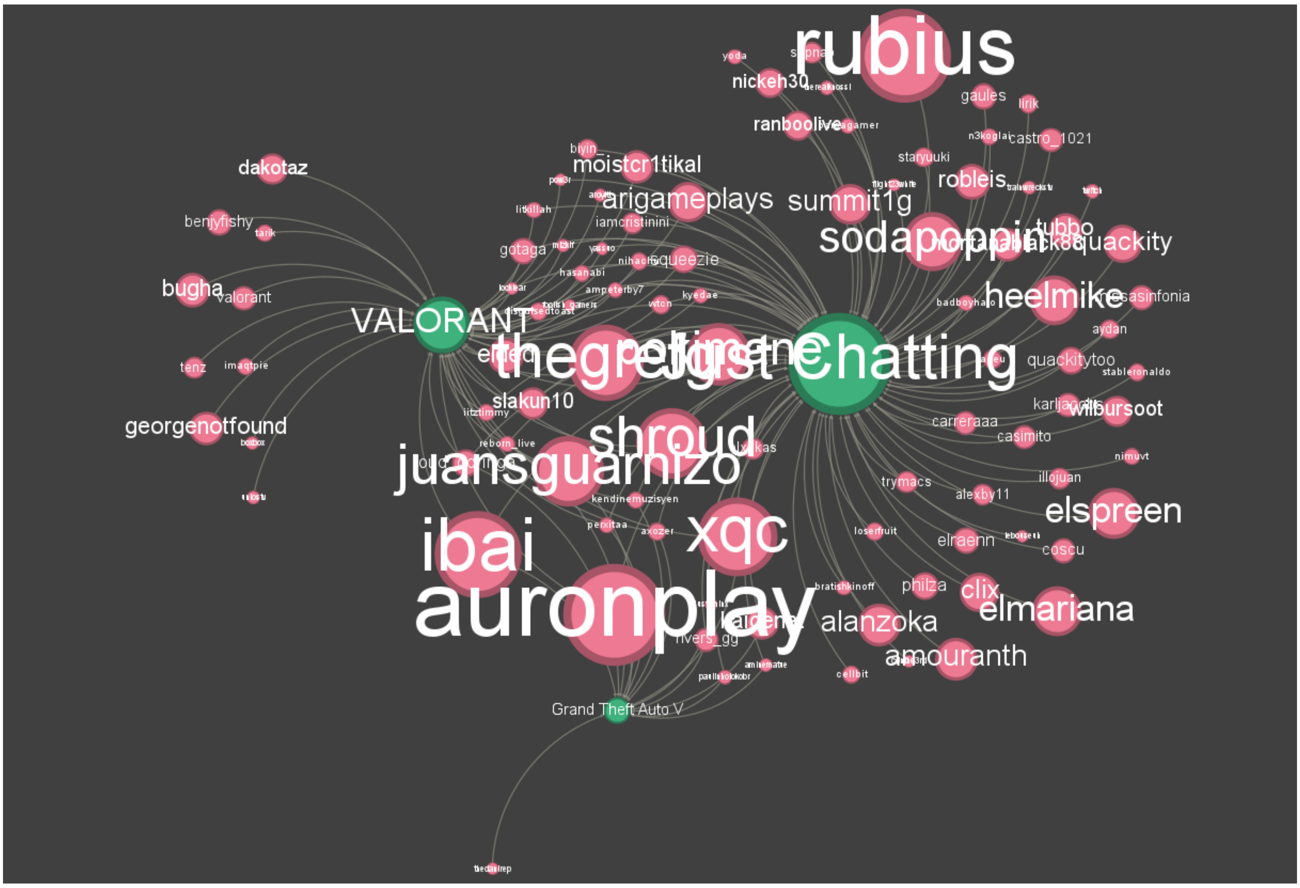


Figure 1: Top 3 games played by top 100 streamers

The size of the nodes related to streamers is proportional to the number of followers they have and the size of the nodes related to the games is proportional to the number of streamers playing them.

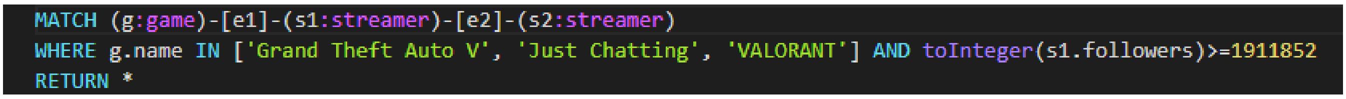
The graph highlights the connections between streamers who play the same games. Streamers playing the same game tend to form clusters or communities, indicating shared interests or collaborations. These clusters may represent different streaming communities or networks within the Twitch platform.

We observe that also the most popular streamers use Twitch for “Just Chatting”, and even a considerable number of them just for that, without playing any game.

2.1 Community mining based on recommendations

We tried to look for commonalities between the streamers recommended by other streamers. In order to have an implementable dataset, we chose streamers recommended by the top 100 followers who play one of the top three games found in the previous section.

The graph can be obtained with the following Cypher query:



```
MATCH (g:game)-[e1]-(s1:streamer)-[e2]-(s2:streamer)
WHERE g.name IN ['Grand Theft Auto V', 'Just Chatting', 'VALORANT'] AND toInteger(s1.followers)>=1911852
RETURN *
```

We classified the data of streamers into communities on the basis of modularity. Modularity is a measure of the structure of a graph, measuring the density of connections within a module or community. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. We obtained a modularity of 0.786. Gephi uses the Louvain method for calculating modularity and one thing to note is that high modularity doesn’t denote a good partition. The size of the nodes is proportional to the number of average viewers the streamer has.

We found that the algorithm classified the streamers into 14 communities. On looking at the streamers, and their description we found that most of the streamers in each community speak the same language as mentioned below.

- 0 - Portuguese
- 1 - German
- 2 - Russian
- 3 - Italian
- 4, 7, 8, 9, 10, 11 - English
- 5 - French
- 6 - Turkish
- 12, 13 - Spanish

By looking at Fig. 2, we can see that the Spanish-speaking communities (12, 13) are very close to each other. Similarly, the English-speaking communities (4, 7, 10) are also intermingling with each other as they consist of variety-gamers. Whereas, the 8th community is very close to the Portuguese-speaking community(0). On examining the 8th community we found that most of the streamers in the 8th community play FIFA, which is one of the most popular games in Portugal, Brazil, and other Portuguese-speaking regions. Thus it is common for few Portuguese-speaking streamers to recommend English-speaking streamers playing FIFA, and vice-versa.

Therefore, we can conclude that most of the streamers recommended by top streamers often share the same language, and games played.

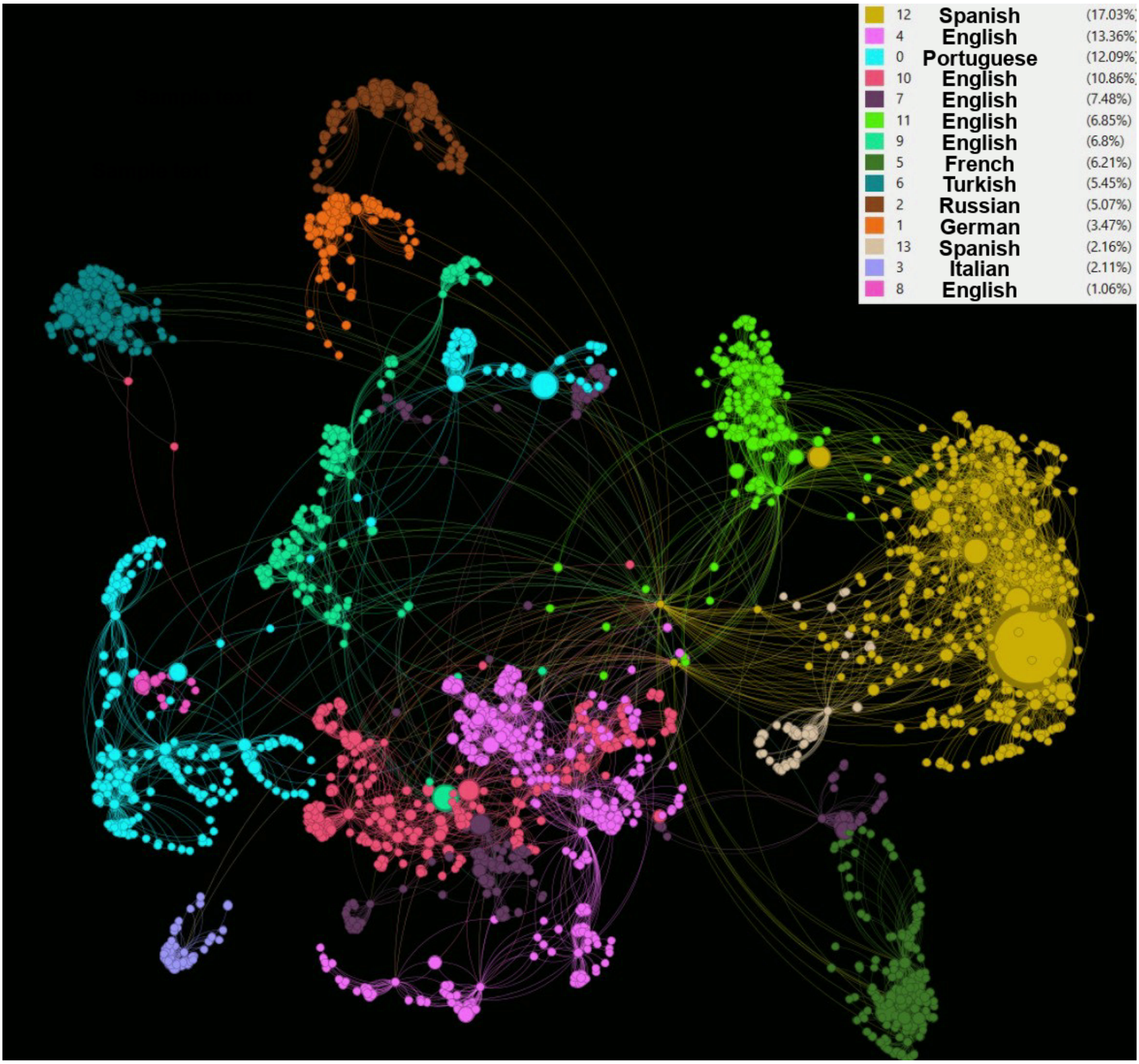


Figure 2: 14 Communities

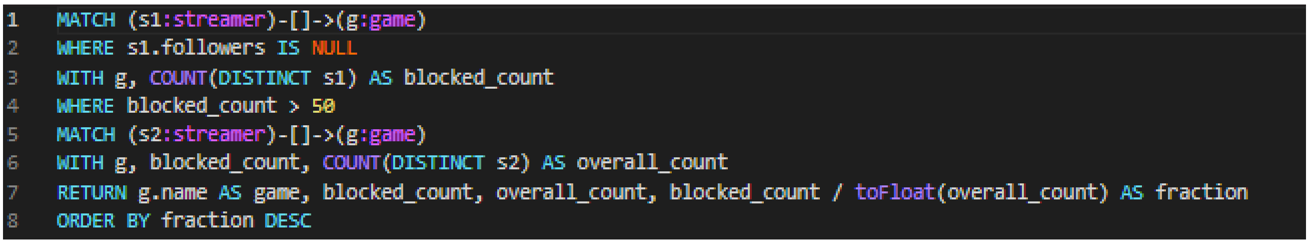
3 Blocked streamers

3.1 General analysis

Similar to the previous analysis, in this section we will focus on Twitch streamers who have been banned. We will try to identify patterns or reasons underlying Twitch's policies regarding streamers bans.

Firstly, we started by filtering games played by at least 50 streamers who are currently banned. Then we identified the games with the highest proportion of banned streamers. We will utilize this information to gain insight into which games, or category of games, are more likely to results in a streamer's ban.

We did it applying this query:



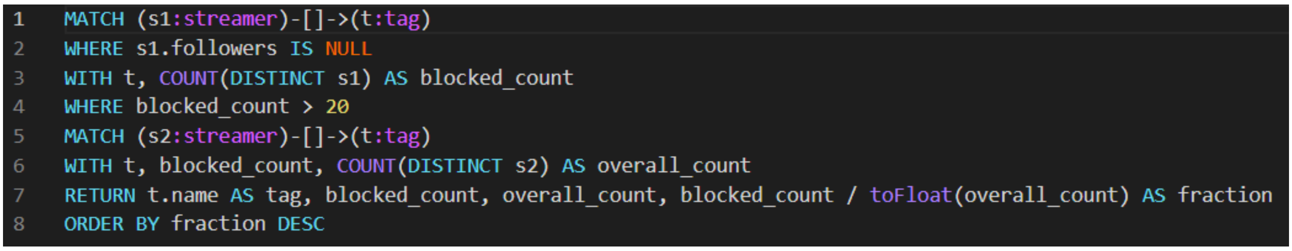
```
MATCH (s1:streamer)-[]->(g:game)
WHERE s1.followers IS NULL
WITH g, COUNT(DISTINCT s1) AS blocked_count
WHERE blocked_count > 50
MATCH (s2:streamer)-[]->(g:game)
WITH g, blocked_count, COUNT(DISTINCT s2) AS overall_count
RETURN g.name AS game, blocked_count, overall_count, blocked_count / toFloat(overall_count) AS fraction
ORDER BY fraction DESC
```

Here we can see a table presenting the results, including the game name, the number of banned streamer who used to play the game, the total number of streamer who played the game ((including both banned and non banned ones) and the fraction between these two numbers:

Rank	Name of the Game	N. of Blocked Streamers	N. of Streamers played the Game	Proportion
1	Slots	71	243	0.292
2	Virtual Casino	100	366	0.273
3	Fortnite	55	962	0.057
4	Grand Theft Auto V	79	1538	0.051
5	Just Chatting	234	4649	0.050

Table 3: Top Games which have the highest banning rates

Then we did the same for tags, now filtering for the ones used by at least 20 streamers who got banned. Using this query:



```
MATCH (s1:streamer)-[]->(t:tag)
WHERE s1.followers IS NULL
WITH t, COUNT(DISTINCT s1) AS blocked_count
WHERE blocked_count > 20
MATCH (s2:streamer)-[]->(t:tag)
WITH t, blocked_count, COUNT(DISTINCT s2) AS overall_count
RETURN t.name AS tag, blocked_count, overall_count, blocked_count / toFloat(overall_count) AS fraction
ORDER BY fraction DESC
```

The tags are listed below:

Rank	Tag	N. of Blocked Streamers	N. of Streamers used the Tag	Proportion
1	Bonus	28	30	0.933
2	Gamble	26	30	0.867
3	Gambling	27	32	0.844
4	казиноонлайн (casino online)	21	28	0.750
5	Casino	37	51	0.725
...				
9	Русский	163	1632	0.100
...				
11	Español	119	2172	0.055
12	Deutsch	74	1373	0.054
13	Português	60	1478	0.041
...				
16	English	389	10595	0.037

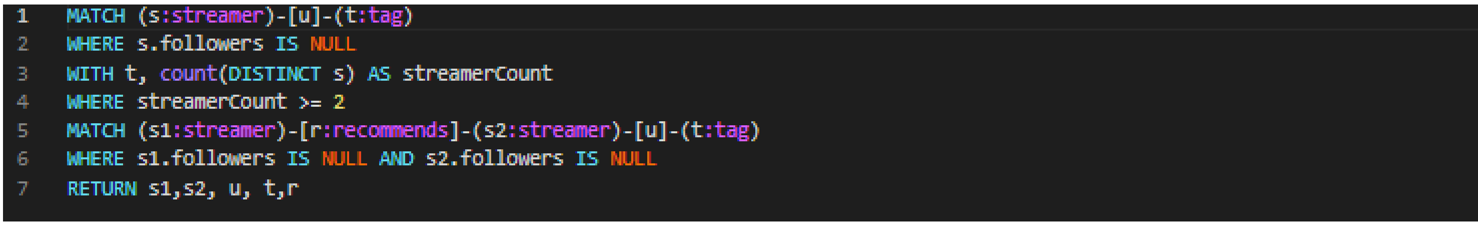
Table 4: Top tags used by the blocked streamers

From the table 3 we can observe that the top 2 games are - Slots and Virtual Casino, both of which are online gambling games. So, we can see that most of the streamers who plays these games have faced bans from Twitch. From table 4, it is noticed that the top 5 tags are again related to gambling games: Bonus, Gamble, Gambling, казиноонлайн (casino online in Russian) and Casino. This aligns with Twitch's recent policy update, which states, 'starting October 18th, we are further tightening our rules to also prohibit any streaming of listed sites that contain slots, roulette, and dice games and are unlicensed in the U.S. or other jurisdictions that offer consumer protections...'. This stricter policy reinforces Twitch's stance on betting-related content, and help explaining the pattern we noticed in our dataset.

In the table we can also see that the tags linked to the highest number of banned streamer are all linguistics tags. Of these tags English appear to be the most popular, while German, Spanish and especially Russian, presents higher proportions of banned streamers. The proportion of Russian-speaking streamers banned by Twitch are around 10%, while the same figure is 3.7% for the English-speaking counterpart. If we look into the total number of streamers who got banned, "English" is the most common tag used by them and it is 389 streamers.

3.2 Community mining

Next, we created a graph using the blocked streamers and their tags as nodes, with their links and the recommendations between streamers as hedges. We filtered the dataset to keep only the tags connected to at least two blocked streamers. The following query was used:



```
MATCH (s:streamer)-[u]-(t:tag)
WHERE s.followers IS NULL
WITH t, count(DISTINCT s) AS streamerCount
WHERE streamerCount >= 2
MATCH (s1:streamer)-[r:recommends]-(s2:streamer)-[u]-(t:tag)
WHERE s1.followers IS NULL AND s2.followers IS NULL
RETURN s1,s2, u, t,r
```

Using this graph, the json file which is created is to be used to display the graph in Gephi. We classified the data of streamers into communities on the basis of modularity, obtaining a modularity of 0.708. We used Resolution equals to 3, to avoid having too many communities. This led to the creation of 8 distinct groups. The size of the nodes is based on their Degree, and because of this, the majority of the larger nodes are either linguistic tags or tags related to casino (gambling, bonus, Casino, or the Russian translations of the same words).

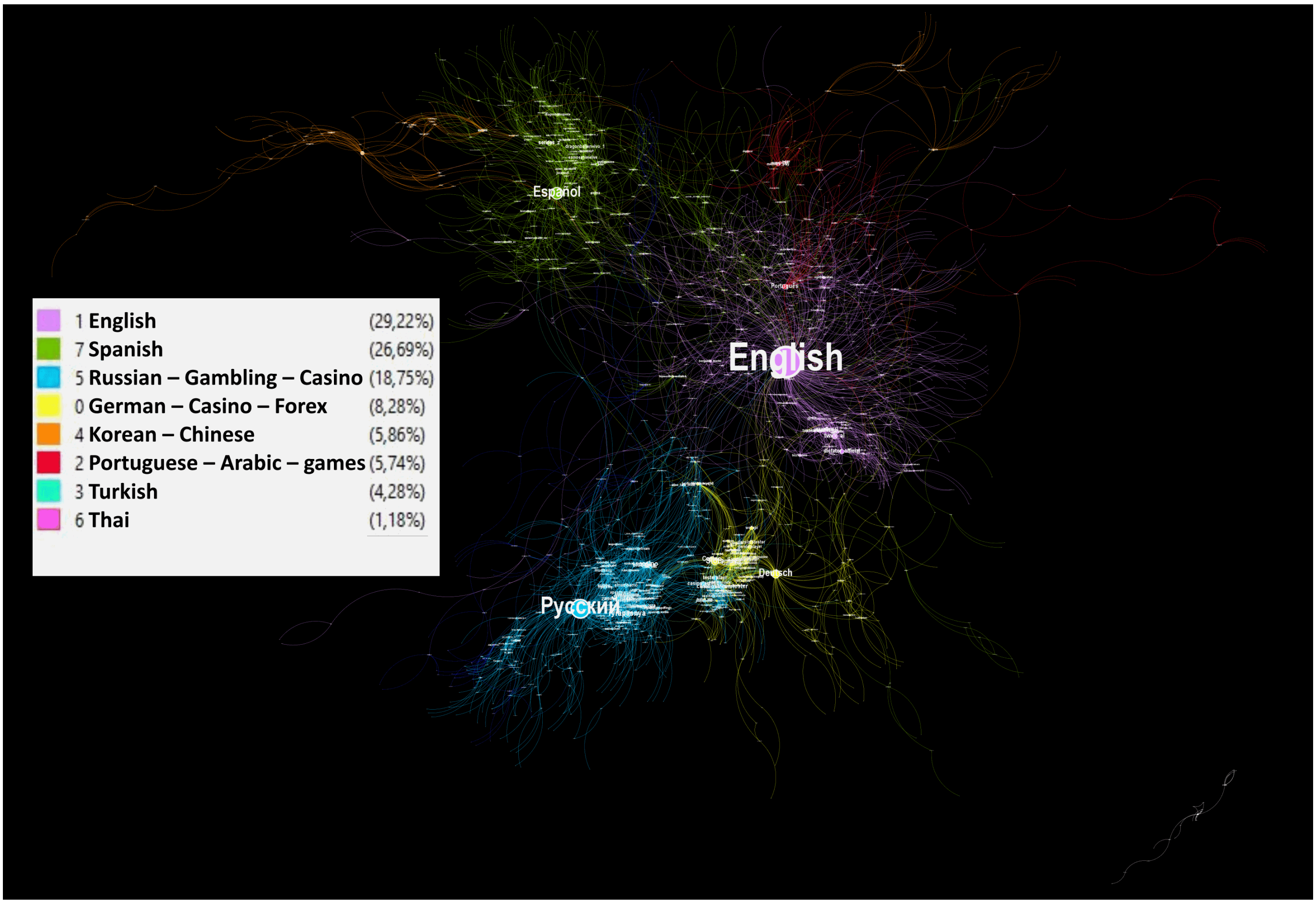

Español
English
Русский
Deutsch
1 English (29,22%)
7 Spanish (26,69%)
5 Russian – Gambling – Casino (18,75%)
0 German – Casino – Forex (8,28%)
4 Korean – Chinese (5,86%)
2 Portuguese – Arabic – games (5,74%)
3 Turkish (4,28%)
6 Thai (1,18%)

Figure 3: Community mining for Banned Streamers

In Figure 3 we can see the obtained Graph. All the obtained communities have been mainly classified because of the language spoken by the streamers composing them. Differently from the others communities, we saw an important role of tags related to Casino, gambling and trading and forex in the Russian and German speaking communities. The only community completely separated by the others is the Thai one, which can be seen in the right-bottom corner of the plot.

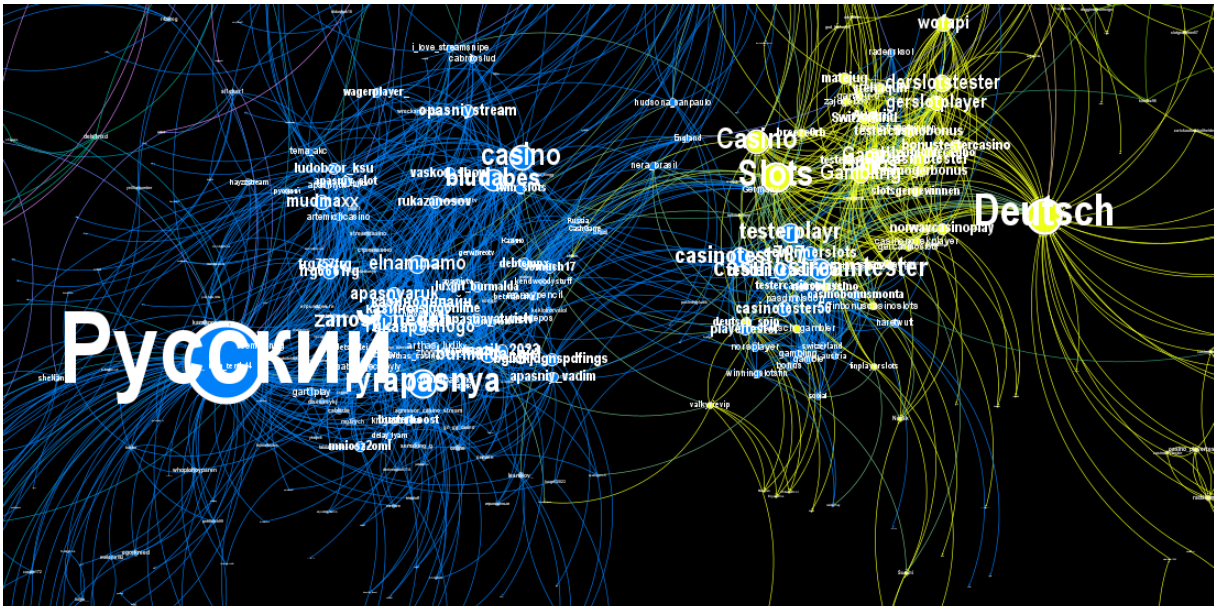


Figure 4: Details of the German and Russian Communities for Banned Streamers

In Figure 4 we have a more detailed picture of the Russian and German communities. It's possible to see how tags like Casino Slots, casino, gambling and also streamers with related names, are connecting the two communities.

Two distinct theories can be derived from these findings. Firstly, it is possible that both German and Russian authorities have implemented stricter regulations or enforcement measures regarding online gambling or betting content. However, is also true that this could be because of similar streaming practices and content creation patterns among streamers in these regions. There could be a larger presence of streamers playing casino games in Germany and in the Russian Federation compared to other countries, contributing to the higher proportion of banned streamers being associated with those tags.

4 Twitch Recommendation System reverse engineer analysis

This analysis was made as an attempt to reverse engineer how the Twitch recommendation system works, by taking a closer look at the graph structure around the recommendation relationships. For additional information and code, refer to recommendation_analysis.ipynb.

To do that, we started by creating a table with information regarding all the recommendation relationships present in the graph. Every row in the table is a recommendation relationship between the recommendee and recommended streamers, along with several information regarding the two streamers, detailed below.

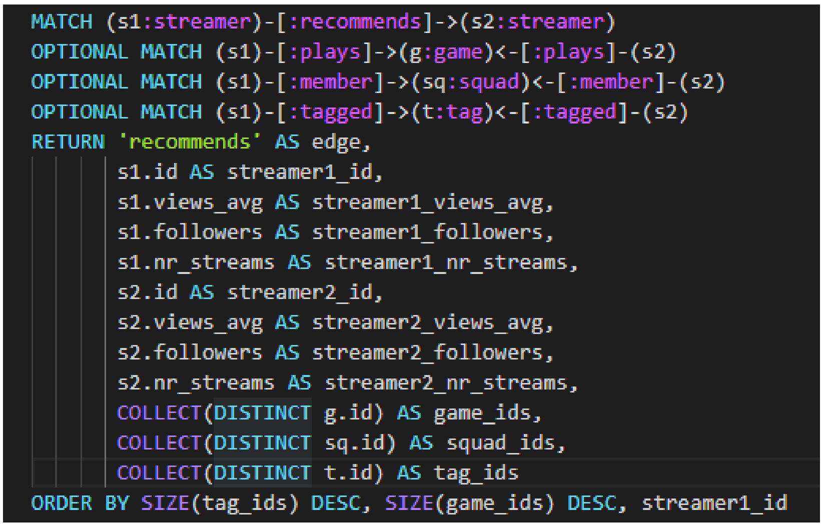
Individual Recommendee and Recommended streamers information:

- Average views
- Number of followers
- Number of performed streams
- Number of games played
- Number of tags in the profile
- Number of squads it is a member of
- Number of times the streamer recommends another
- Number of times the streamer gets recommended by another

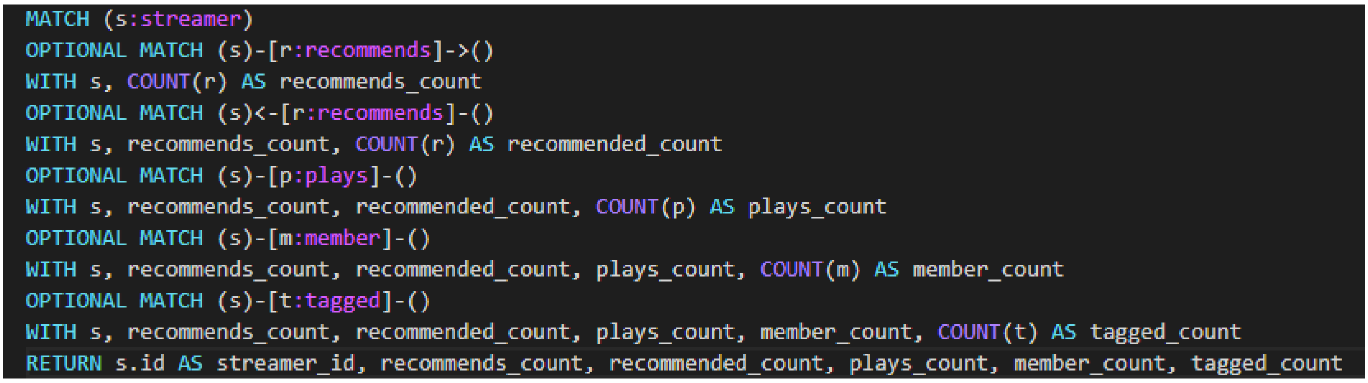
Recommendation Relationship information:

- List with shared games (empty if none)
- Number of shared games
- List with shared tags (empty if none)
- Number of shared tags
- List with shared squads (empty if none)
- Number of shared squads

Some of the columns were created directly from Cypher queries on *Memgraph*, while others were inserted using Python afterwards. The Cypher queries that created the original tables (which were joined later on Python to create the full table) can be found below.



```
MATCH (s1:streamer)-[:recommends]->(s2:streamer)
OPTIONAL MATCH (s1)-[:plays]->(g:game)<-[:plays]-(s2)
OPTIONAL MATCH (s1)-[:member]->(sq:squad)<-[:member]-(s2)
OPTIONAL MATCH (s1)-[:tagged]->(t:tag)<-[:tagged]-(s2)
RETURN 'recommends' AS edge,
       s1.id AS streamer1_id,
       s1.views_avg AS streamer1_views_avg,
       s1.followers AS streamer1_followers,
       s1.nr_streams AS streamer1_nr_streams,
       s2.id AS streamer2_id,
       s2.views_avg AS streamer2_views_avg,
       s2.followers AS streamer2_followers,
       s2.nr_streams AS streamer2_nr_streams,
       COLLECT(DISTINCT g.id) AS game_ids,
       COLLECT(DISTINCT sq.id) AS squad_ids,
       COLLECT(DISTINCT t.id) AS tag_ids
ORDER BY SIZE(tag_ids) DESC, SIZE(game_ids) DESC, streamer1_id
```



```
MATCH (s:streamer)
OPTIONAL MATCH (s)-[r:recommends]->()
WITH s, COUNT(r) AS recommends_count
OPTIONAL MATCH (s)<-[r:recommends]-()
WITH s, recommends_count, COUNT(r) AS recommended_count
OPTIONAL MATCH (s)-[p:plays]-()
WITH s, recommends_count, recommended_count, COUNT(p) AS plays_count
OPTIONAL MATCH (s)-[m:member]-()
WITH s, recommends_count, recommended_count, plays_count, COUNT(m) AS member_count
OPTIONAL MATCH (s)-[t:tagged]-()
WITH s, recommends_count, recommended_count, plays_count, member_count, COUNT(t) AS tagged_count
RETURN s.id AS streamer_id, recommends_count, recommended_count, plays_count, member_count, tagged_count
```

After exporting those tables as .csv files and some manipulation on Python, some key insights regarding the Twitch Recommendation system could be found, and are listed below.

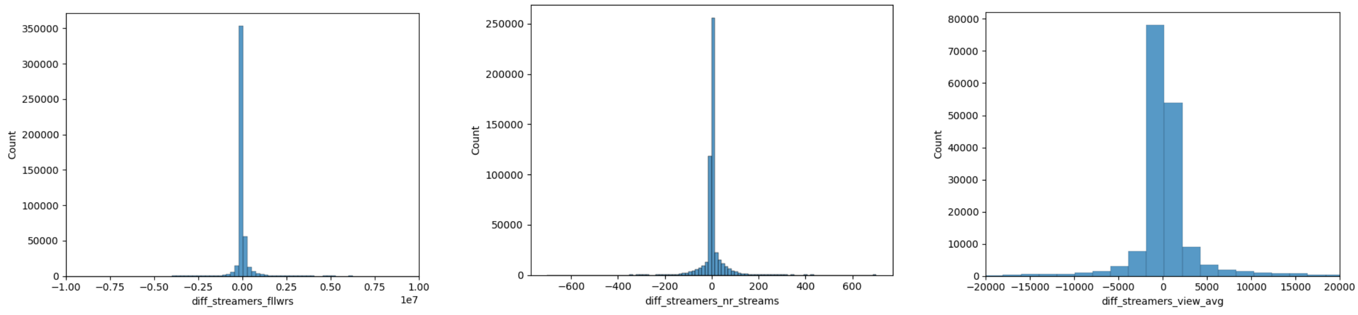
- Around 74% of the recommendations where both recommendee and recommended are linked to at least one game, share at least one or more of these games in common (both linked to the same game).
- Around 87% of the recommendations where both recommendee and recommended are tagged in at least one tag, share at least one or more of these tags in common (both tagged to the same tag).
- Around 19% of the recommendations where both recommendee and recommended belong to a squad, share the same squad (intra-squad recommendations).

From the information above, we can see that, for active streamers (those who play games, have tags and belong to squads), the tags are highly influential in determining recommendations, closely followed by which games are played. We see that squads are not so influential in the recommendations, as only 19% of the recommendations from people that have squads are between streamers in the same one.

- 48% of the recommendations are from streamers that have no tags, play no games and belong to no squads.
- 45% of the recommendations are for streamers that have no tags, play no games and belong to no squads.
- 27% of the recommendations involve both inactive recommendee and recommended streamers.

These three statistics above tell us that there is a significant amount of streamers that are relatively inactive and still give and get recommendations, which means that for those, other factors than games, tags, and squads are used to define recommendations.

When comparing the differences in number of average views, followers and streams between recommendee and recommended streamers, we see that most recommendations are between streamers that are not so different regarding these three statistics. As can be seen in the plots below, the median difference falls very close to 0 in all three comparisons, with few large deviations. From this analysis we can say that recommendations tend to happen more between streamers that have a similar number of followers, streams and average views.



(a) No. followers

(b) No. streams

(c) Average views

Figure 5: Statistics (histogram)

We can now also create a correlation plot with the variables of the final table, and look for interesting insights we can extract from it.

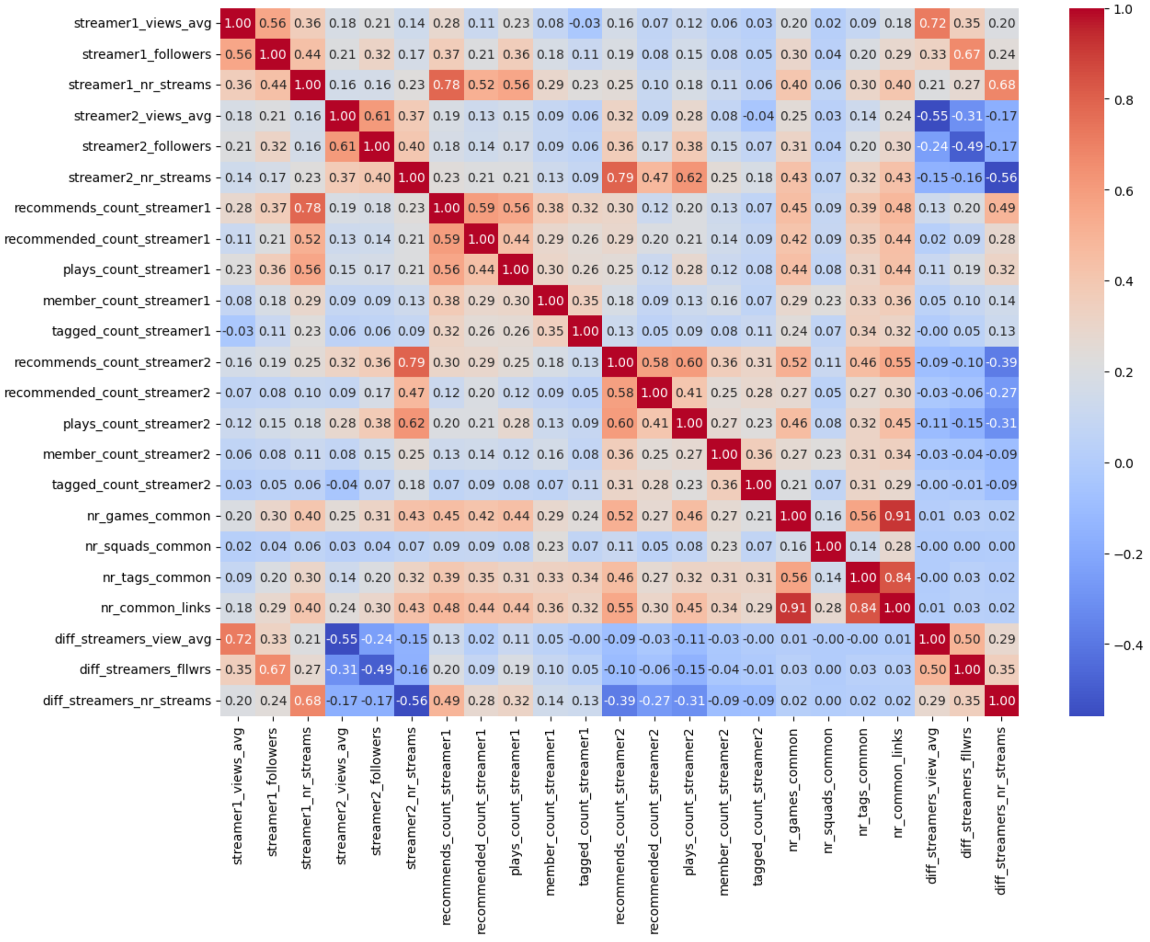


Figure 6: Correlation plot

An interesting observation is regarding the number of streams and the recommendations. There is a strong correlation between the number of streams that a streamer performs and the amount of times it recommends someone (0.78 and 0.79). The same goes for the amount of times it gets recommended, but less strongly (0.52 and 0.47). This can be due to the fact that more active streamers tend to play more games (0.56 and 0.62), and as was seen before, common games is a factor that influences recommendations between active users. To conclude, the more a streamer streams, the higher the chance it has to have a diversified content, which spreads its recommendation span, hence increasing the number of recommendations involving that streamer.

As a final analysis, we can take a look at the games and tags that most appear as shared between recommendee and recommended streamers. The top 10 tags and games can be found on the tables below.

The tags are listed below:

Rank	Tag	no. shares
1	English	57644
2	Español	12514
3	Русский	10575
4	한국어(Korean)	9062
5	Français	8896
6	Português	7972
7	Deutsch	7223
8	日本(Japanese)	6253
9	中文(Chinese)	3940
10	Türkçe	2972

Table 5: Most shared tags between recomendee and recommended

The games:

Rank	Game	no. shares
1	Just Chatting	36832
2	League of Legends	12193
3	VALORANT	10503
4	Grand Theft Auto V	9855
5	Minecraft	6110
6	Escape from Tarkov	4733
7	Apex Legends	4002
8	World of Warcraft	3587
9	Counter-Strike: Global Offensive	3541
10	Fortnite	3536

Table 6: Most 'shared' games between recomendee and recommended

It is seen that the most shared tags in recommendations are all regarding language, which shows that the Twitch recommendation system is heavily influenced by the language of the stream. When it comes to games, the most popular is not a game, but a category for streamers chatting with the public. This is most likely the case because most streamers perform this activity (it is unlikely that a streamer just plays games and do not have breaks where it is just chatting with the public).

To visualize the connections between streamers, their recommendations, tags and games played, one can visualize a Graph containing as tags and games only those appearing in the top 10.

From the image, we notice that languages then to be arranged more towards the edges of the graph's body, not forming a defined cluster per language but a region where most streamers are situated. Regarding games, some games are very well distributed around the languages, like Just Chatting (not a game), League of Legends and Minecraft, appearing more towards the center of the Graph, while other like CS Go is closer to the Portuguese, Russian and German regions, World of Warcraft is more centered in the English region, and so on.

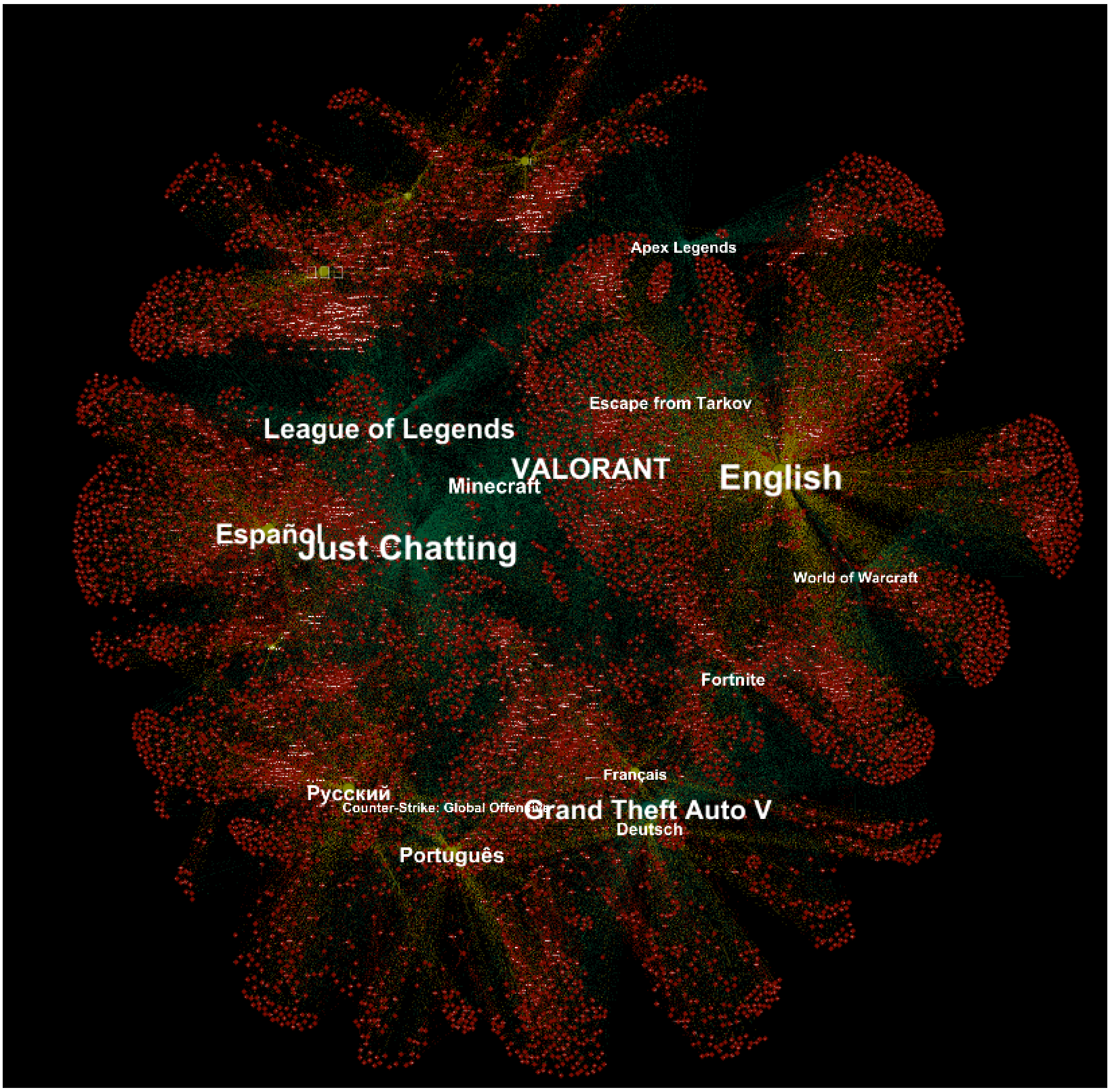


Figure 7: Graph of the top 10 games and tags

5 Conclusion

In summary, the analysis allowed us to derive some conclusions:

- The popular streamers use Twitch mostly to chat.
- Valorant and Grand Theft Auto V are the most played games by popular streamers.
- Most streamers recommended by popular streamers share the same language.
- The most popular games and tags among banned streamers are all related to gambling
- The streamers who use the Russian tag are more likely to be blocked than the streamers who use others linguistic tags
- There is a particularly high proportion of banned streamers from the German and Russian communities
- For the Twitch Recommendation System reverse engineer analysis, the key findings was that common tags significantly increase the chances of a recommendation, while common games also but less strongly. The number of followers, average views and streams are also influential, where recommendations tend to happen between streamers with similar numbers.