

Twitch Gamers Dataset from Stanford Network Analysis Project

By: Cella Wardrop, Mateo Jolivert, Florencia
Murphy and Rümeysa Türkbey





INDEX

- Dataset summary
- Research question
- Why did we choose this?
- Findings
- Possible implications



Dataset summary

Description: “A social network of Twitch users which was collected from the public API in Spring 2018. Nodes are Twitch users and edges are mutual follower relationships between them. The graph forms a single strongly connected component without missing attributes. The machine learning tasks related to the graph are count data regression and node classification.”

- Includes
 -
 - Explicit content streamer identification.
 - Broadcaster language prediction.
 - User lifetime estimation.
 - Churn prediction.
 - Affiliate status identification.
 - View count estimation.



Background research:

- **Title:** Twitch Gamers: a Dataset for Evaluating Proximity Preserving and Structural Role-based Node Embeddings, The University of Edinburgh
- **Hypothesis:** Analysis of the social network and node classification experiments illustrate that Twitch Gamers is suitable for assessing the predictive performance of novel proximity preserving and structural role-based node embedding algorithms
- **Focuses:**
 - Explicit Content
 - Broadcaster Language
 - Dead Account
 - Affiliate Status variables
- **Sampling type:** snowball effect
- **Findings:**
 - Users who broadcast in more commonly spoken language (English, German, French) are more likely to have connections with users who broadcast in the same language
 - Users who broadcast in more commonly spoken languages are well connected.
 - Broadcasters who use explicit language are less popular
 - Twitch users who turned away from the platform are well incorporated in the social network and do not form new communities
 - Affiliate Status attributes had no particular insights



Our research questions

1. What is the most popular language among streamers?
2. What is the relationship between explicit context and language and views
?(who consumes more explicit material)?

We focus on:

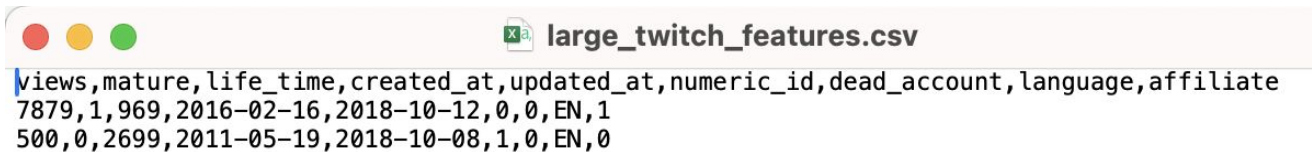
1. Explicit content streamer identification.
2. Broadcaster language prediction.
3. View count estimation.
4. Following relationships



Background research:

Overview of data we used:

- 1) large twitch features = Columns below
- 2) large twitch edges = the id of users



```
large_twitch_features.csv
views,mature,life_time,created_at,updated_at,numeric_id,dead_account,language,affiliate
7879,1,969,2016-02-16,2018-10-12,0,0,EN,1
500,0,2699,2011-05-19,2018-10-08,1,0,EN,0
```

Overview of data...



```
▶ import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt

#read the edges CSV file
df=pd.read_csv('/content/large_twitch_edges.csv')

# Create a graph from the DataFrame
G = nx.from_pandas_edgelist(df, source='numeric_id_1', target='numer

# Number of nodes
num_nodes = G.number_of_nodes()

# Number of edges
num_edges = G.number_of_edges()

print("Number of nodes in the dataset:", num_nodes)
print("Number of edges in the dataset:", num_edges)
```

➞ Number of nodes in the dataset: 168114
Number of edges in the dataset: 6797557

Overview of top 1% creators

⇒ Total creators: 168114

Top 1% count: 1681

Top creators:

	views	mature	life_time	created_at	updated_at	numeric_id	\
96473	384396587	0	2416	2012-03-01	2018-10-12	96473	
129896	368912220	0	2379	2012-04-07	2018-10-12	129896	
61862	340602050	0	2826	2011-01-16	2018-10-12	61862	
128864	297117159	0	2314	2012-06-11	2018-10-12	128864	
144643	294116493	0	2577	2011-09-22	2018-10-12	144643	
...	
158740	2573843	0	2532	2011-11-06	2018-10-12	158740	
95576	2565777	0	1544	2014-07-21	2018-10-12	95576	
125753	2564289	0	3294	2009-10-05	2018-10-12	125753	
104474	2563424	1	1546	2014-07-19	2018-10-12	104474	
138204	2558656	0	3152	2010-02-24	2018-10-12	138204	

	dead_account	language	affiliate
96473	0	RU	0
129896	0	EN	0
61862	0	EN	0
128864	0	EN	0
144643	0	EN	0
...
158740	0	EN	0
95576	0	PL	0
125753	0	EN	0
104474	0	EN	0
138204	0	EN	0

[1681 rows x 9 columns]

Number of nodes in the graph: 1681

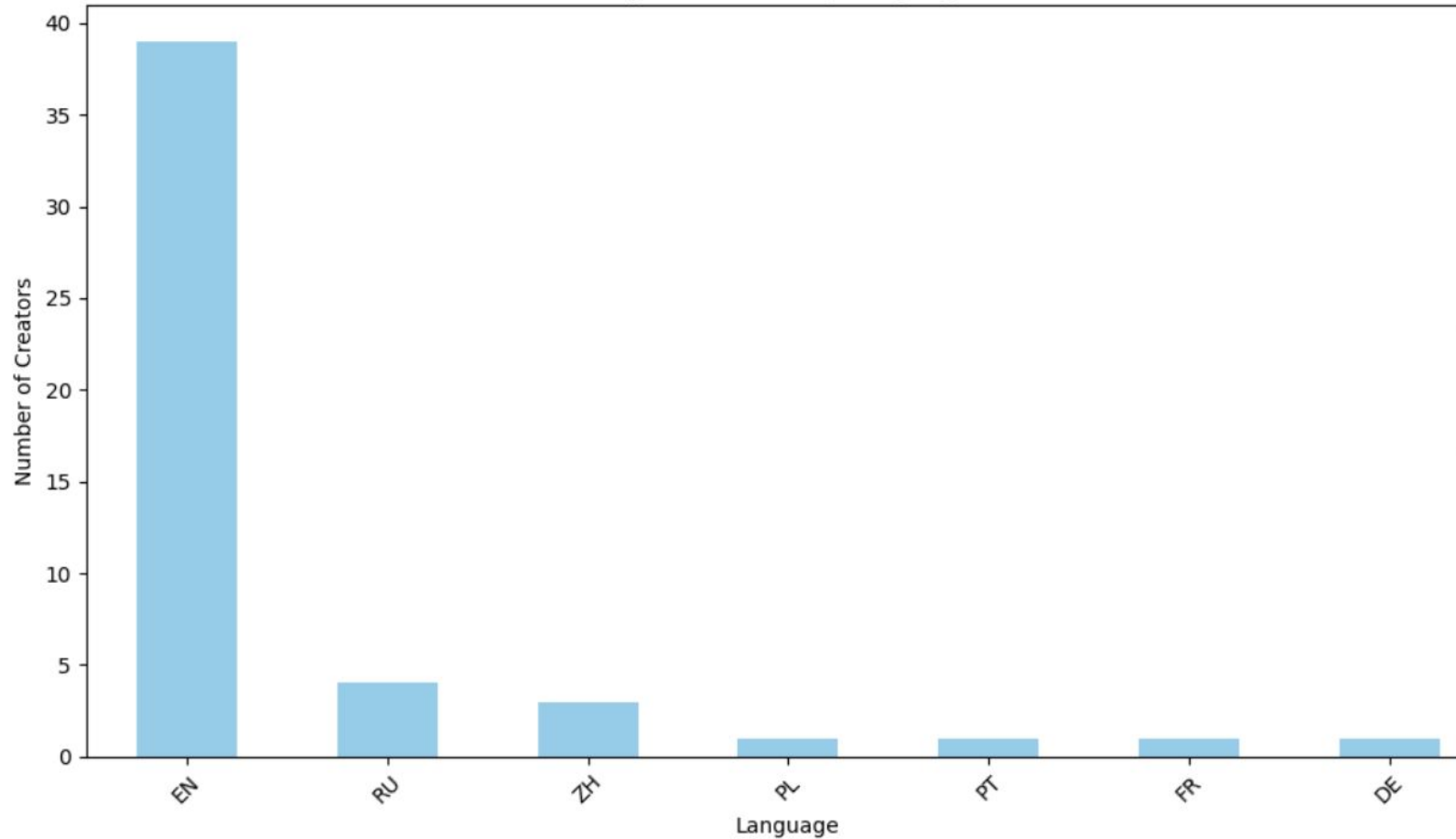


Overview of the Largest Creators (Based on Views):

	numeric_id	views	language	affiliate
96473	96473	384396587	RU	0
129896	129896	368912220	EN	0
61862	61862	340602050	EN	0
128864	128864	297117159	EN	0
144643	144643	294116493	EN	0
110345	110345	264643346	EN	0
125642	125642	243451177	EN	0
64605	64605	240718261	EN	0
71050	71050	218559516	EN	0
161362	161362	213931993	EN	0



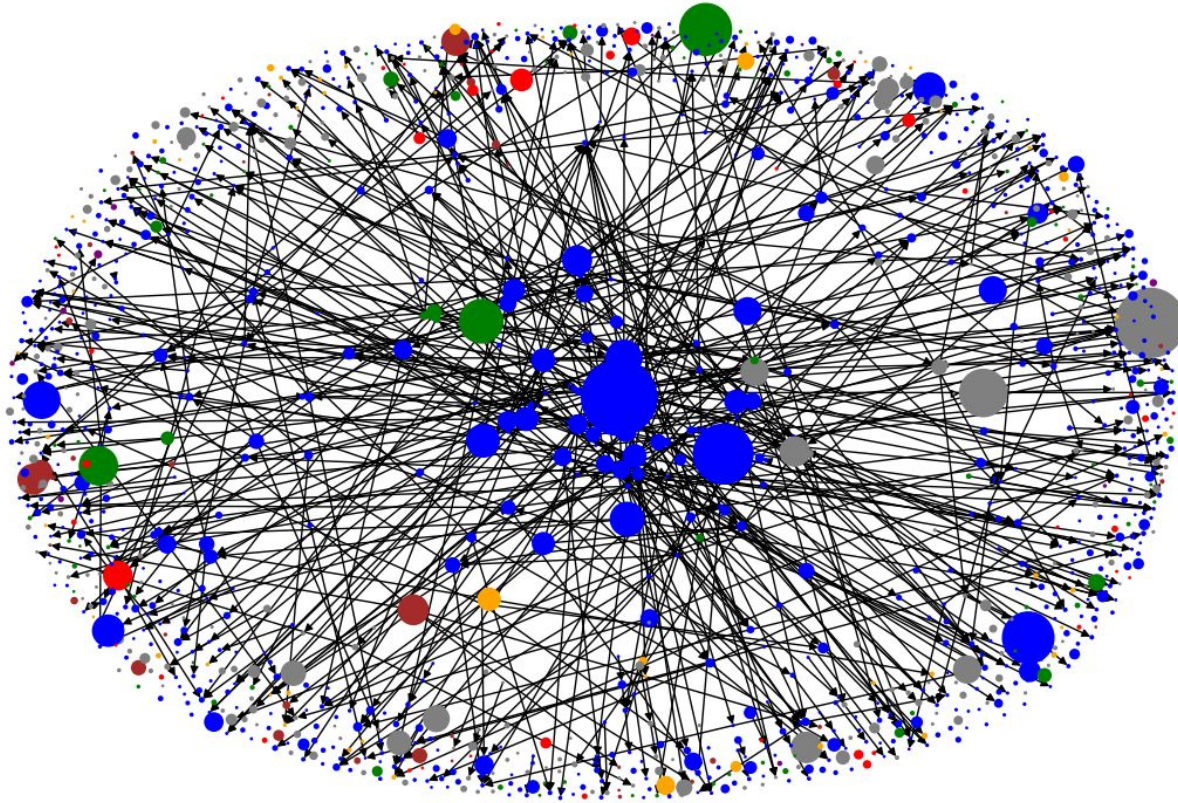
Top 50 Creators by Language



Findings...



Network Graph of Top 1% Creators and Their Neighbors



Each line in the network graph represents a mutual follower relationship between two Twitch users, with the direction of the line indicating the direction of the relationship (i.e., who follows whom).

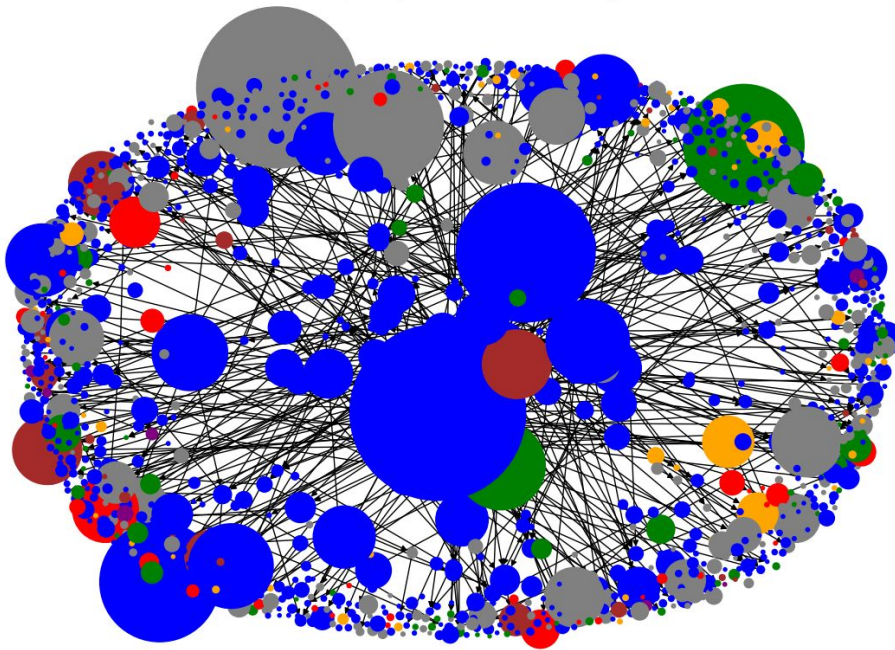
Language and colors relation

- 'EN' (english): 'blue'
- 'DE' (Germany): 'green'
- 'FR' (France): 'red'
- 'ES' (Spain): 'orange'
- 'IT' (Italy): 'purple'
- 'PT' (Portugal): 'brown'
- Default color for languages not explicitly mapped: "gray"

Overall Average Degree Centrality for English Users only: 0.022665337321369584

Graph with standardization of sizes

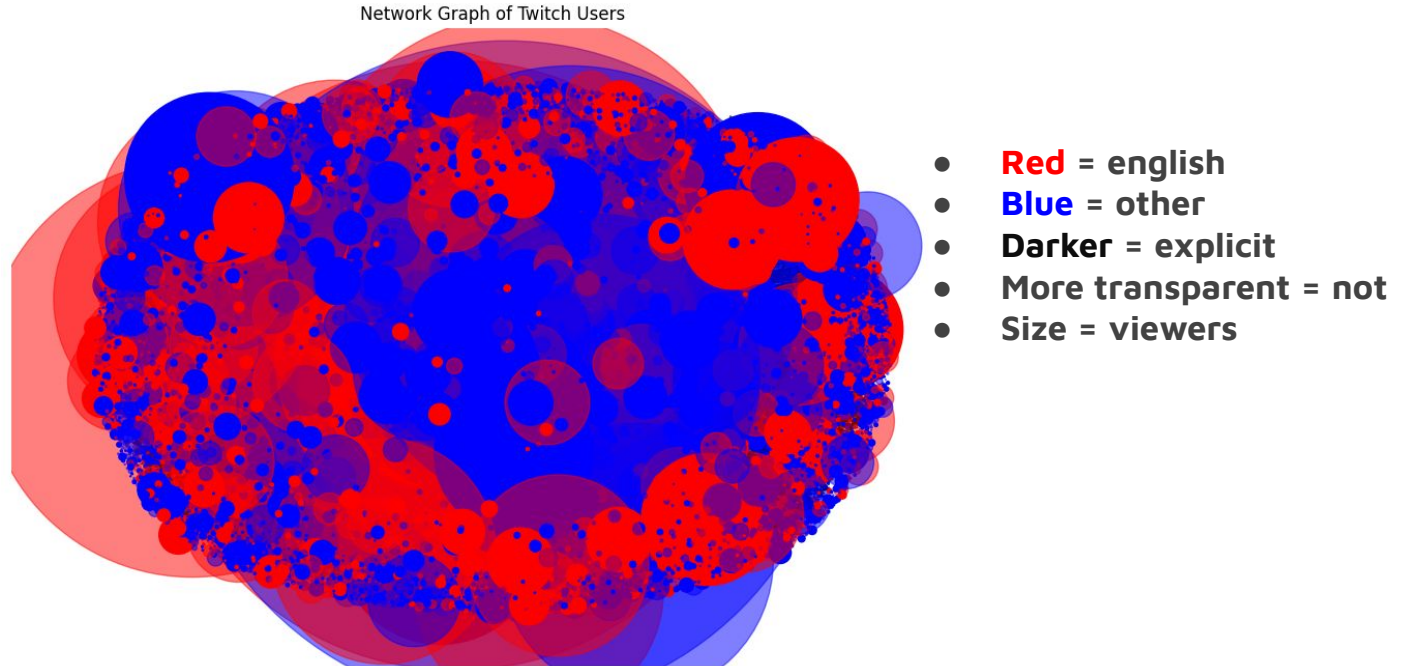
Network Graph of Top 1% Creators and Their Neighbors



Of the top languages, what are trends?

- We can see english is by far the most common language
- This is followed by Dutch in second position
- There is a big predominance of relationships between English and German speaking creators

Relationship between explicit context and language and views (who consumes more explicit material)

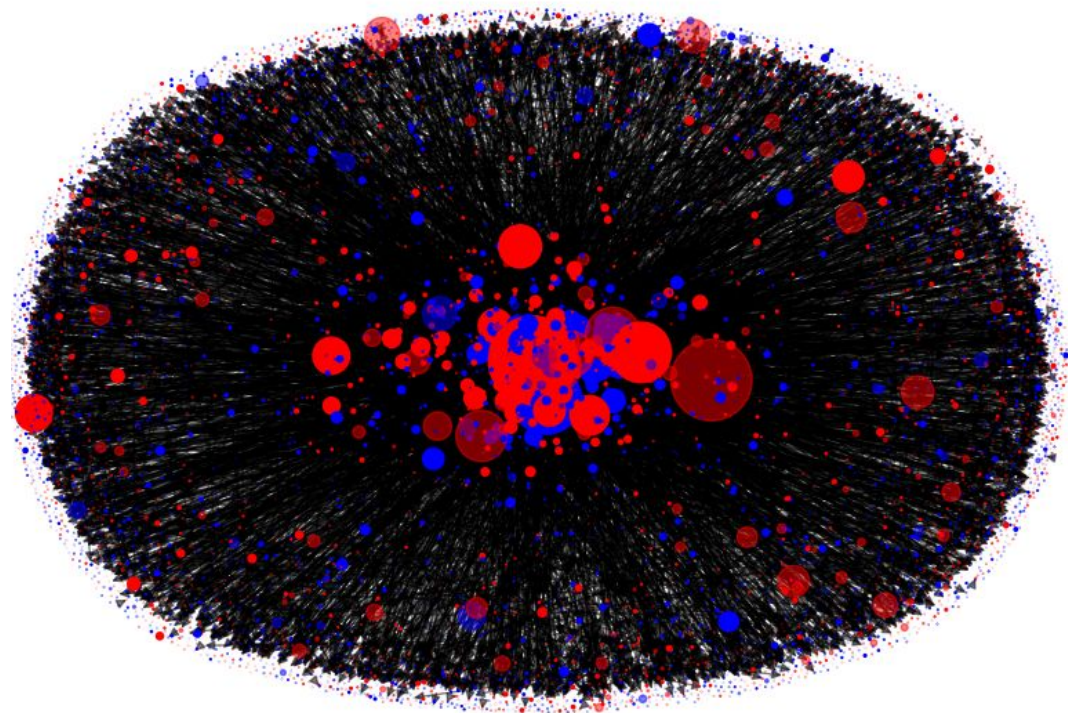


Relationship of explicit context and language and views



(who consumes more explicit material), among all twitch users (+ relationship with language & views)

Network Graph of Twitch Users

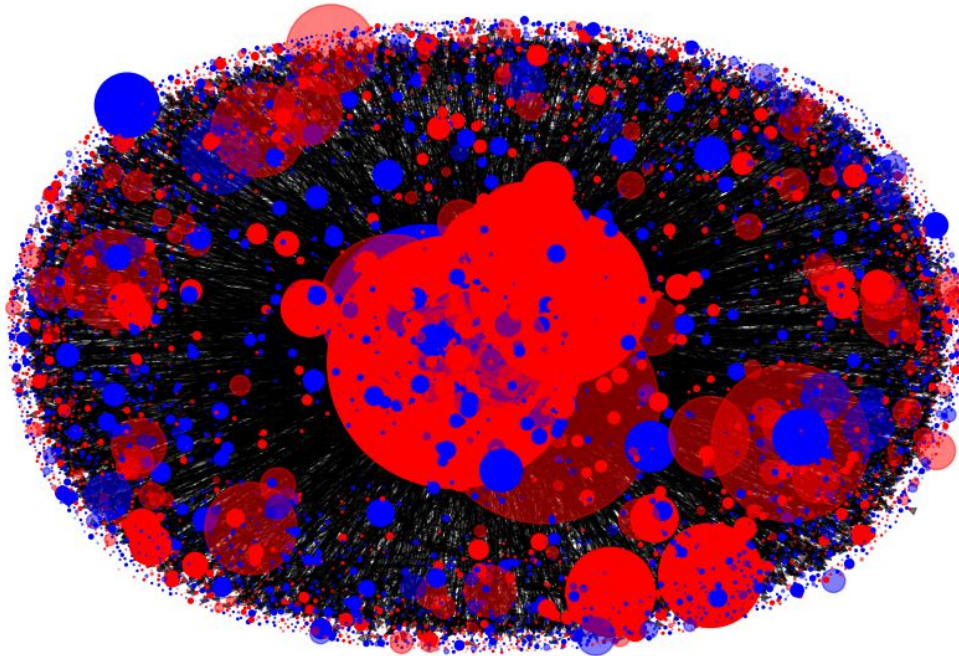


- **Red** = explicit
- **Blue** = not explicit
- **Darker** = english
- **More transparent** = not english
- **Size** = viewers
- **Nodes**= followings

Standardized relationship of explicit context and language and views

(who consumes more explicit material), among all twitch users (+ relationship with language & views)

Network Graph of Twitch Users



- **Red** = explicit
- **Blue** = not explicit
- **Darker** = english
- **More transparent** = not english
- **Size** = viewers
- **Nodes**= followings

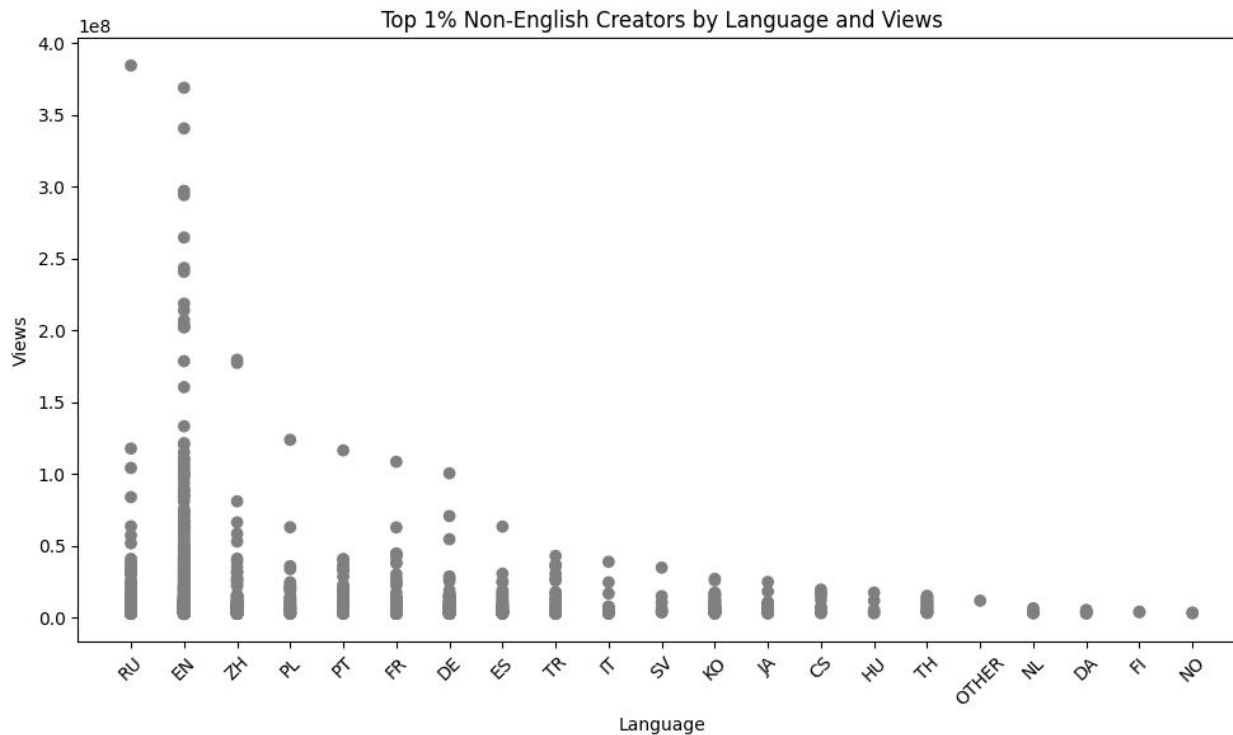
Conclusion:

- It is shown that the smaller people follow the high concentration middle

Overall Average Degree Centrality of the Top 1%:
0.0031644629326994874



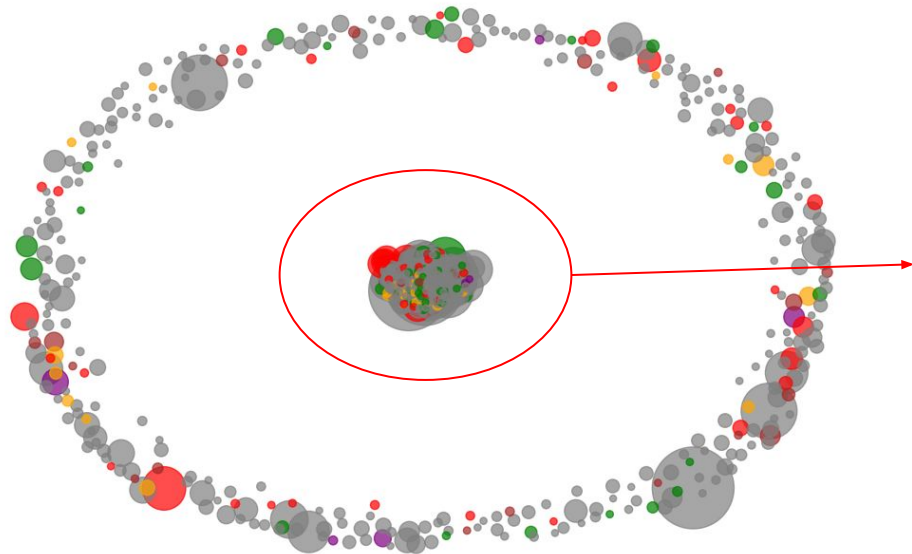
Top 1% of creators, by views and language:





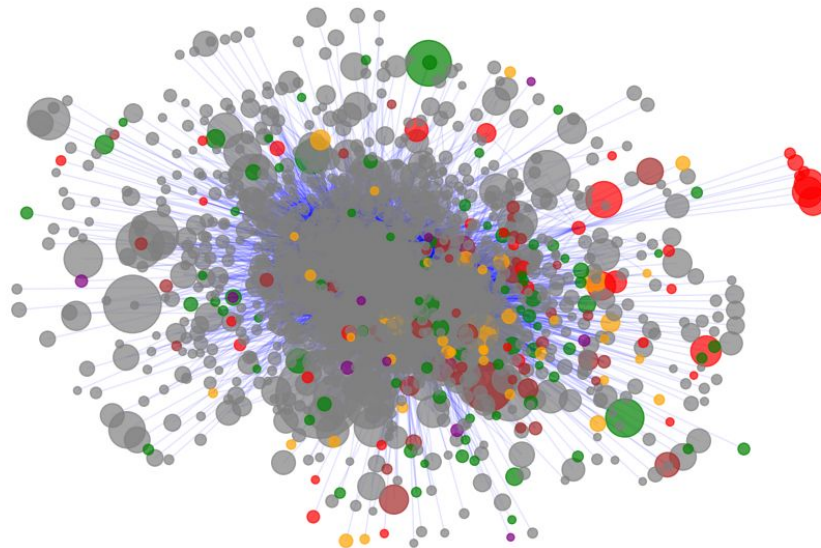
Connection between top 1% of top non-english creators

Network Plot of Top 1% Non-English Twitch Users by Language and Views



Overall Average Degree Centrality of the top 1% non-english (overall): 0.007850344182884453

Network Plot of Highly Connected Component of Top 1% Non-English Twitch Users by Language and Views

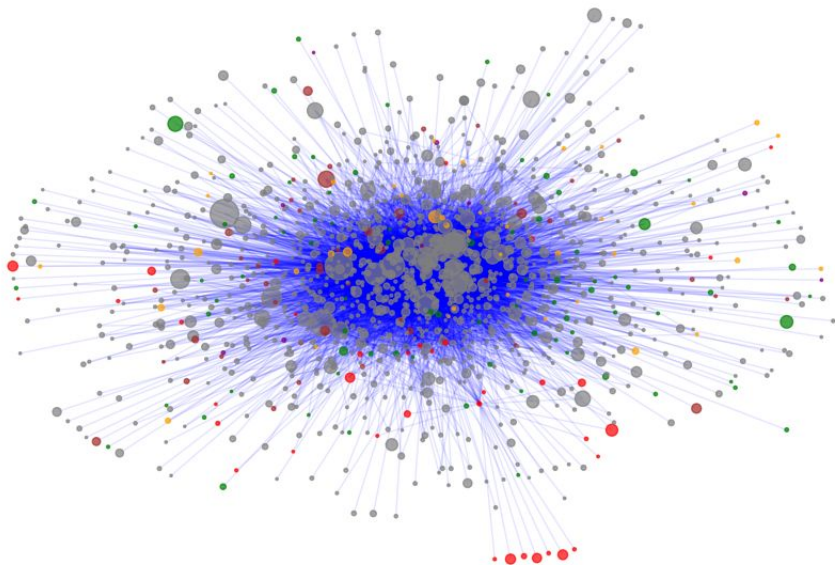


Overall Average Degree Centrality of Sub-Group: 0.012752798195165657



High-concentration top 1% of top non-english creators

Network Plot of Highly Connected Component of Top 1% Non-English Twitch Users by Language and Views



Language and colors relation

- 'DE' (Germany): 'green'
- 'FR' (France): 'red'
- 'ES' (Spain): 'orange'
- 'IT' (Italy): 'purple'
- 'PT' (Portugal): 'brown'
- Default color for languages not explicitly mapped: "gray"

Overall Average Degree Centrality:
0.012752798195165657



Overall, english users are more tight-knit than non-english users:

Overall Average Degree Centrality for:

English Users only: 0.022665337321369584

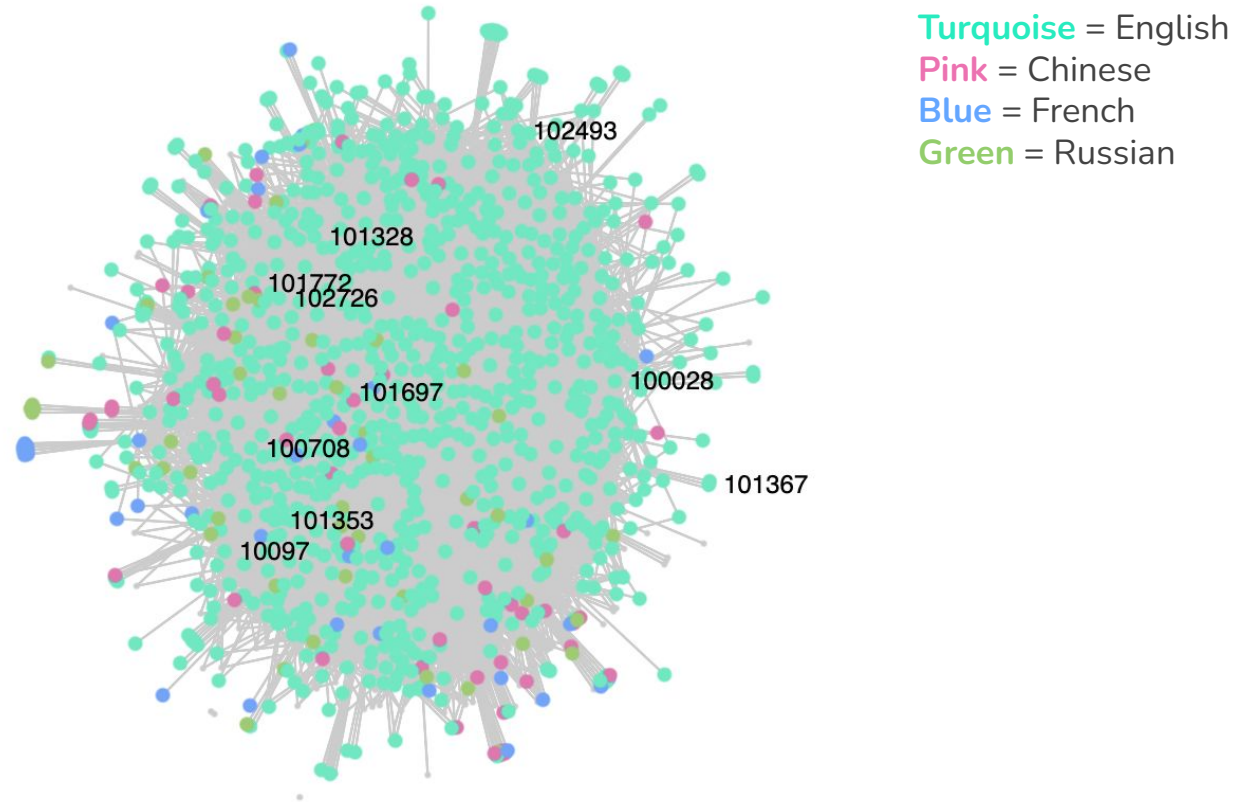
Top 1% non-english : 0.007850344182884453

Overall Average Degree Centrality for all Users: 0.012752798195165657

Silo effect analysis

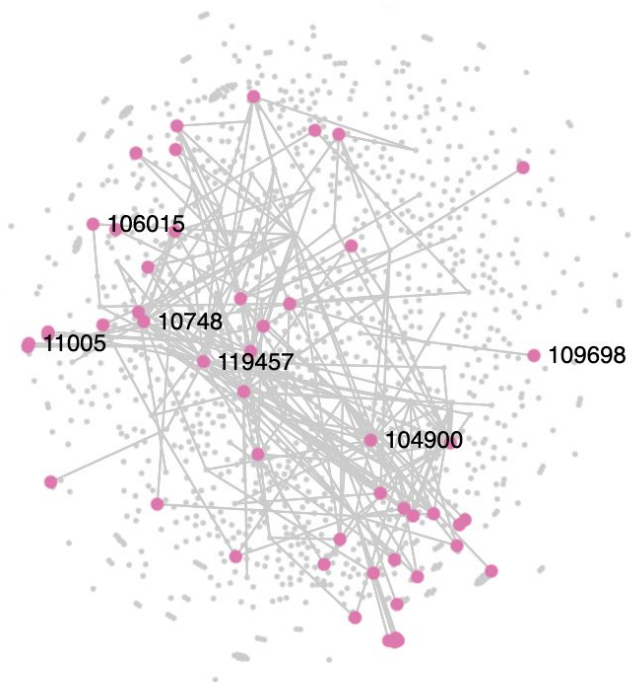
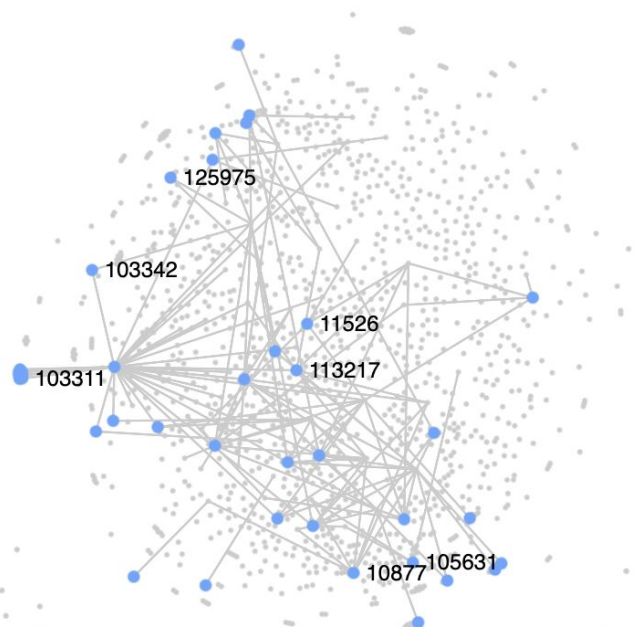
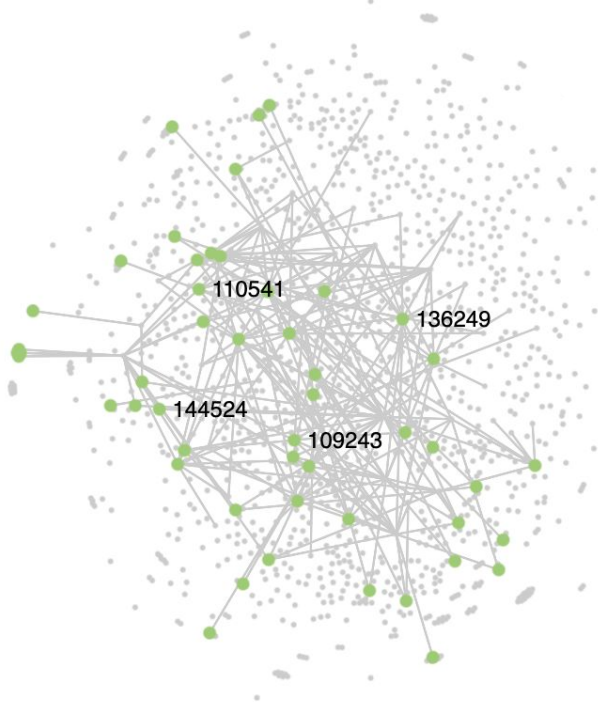


Social network of following between top 1% of twitch streamers, by top 4 languages:





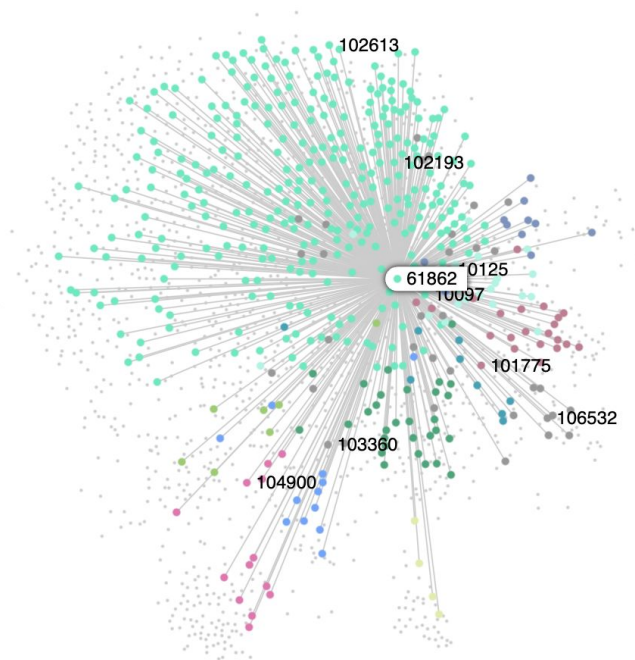
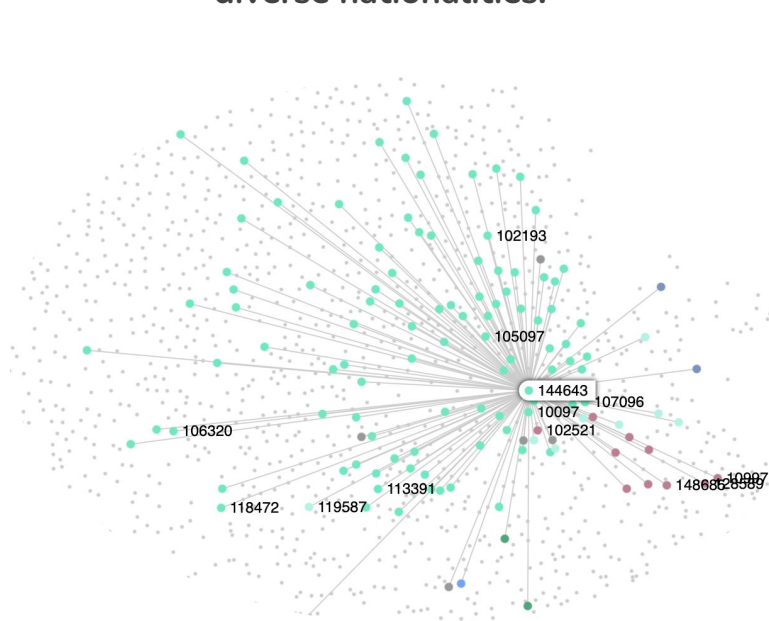
Green = Russian
Blue = French
Pink = Chinese



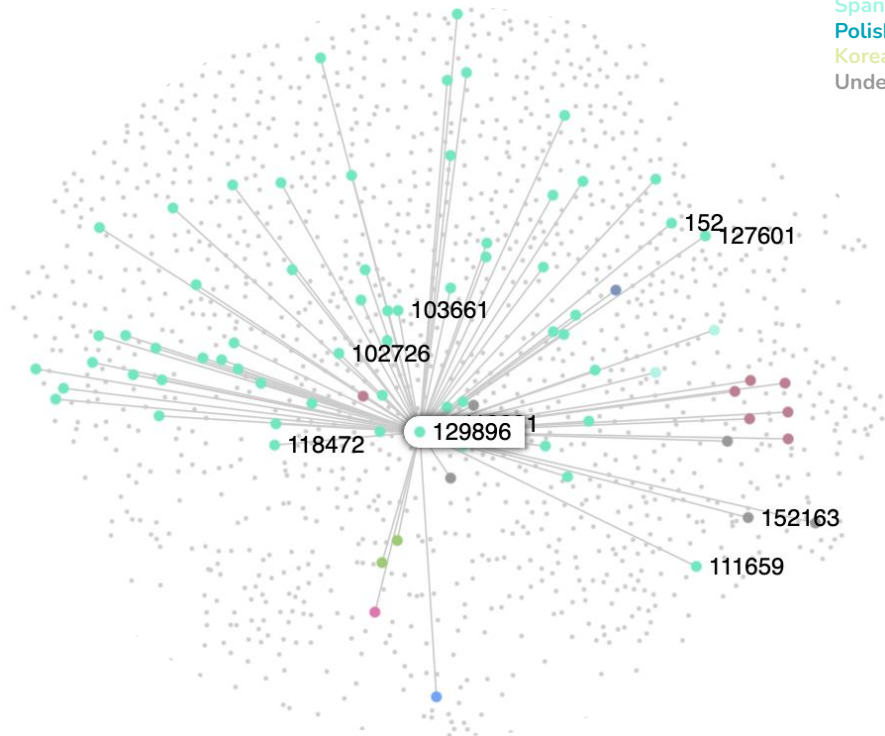
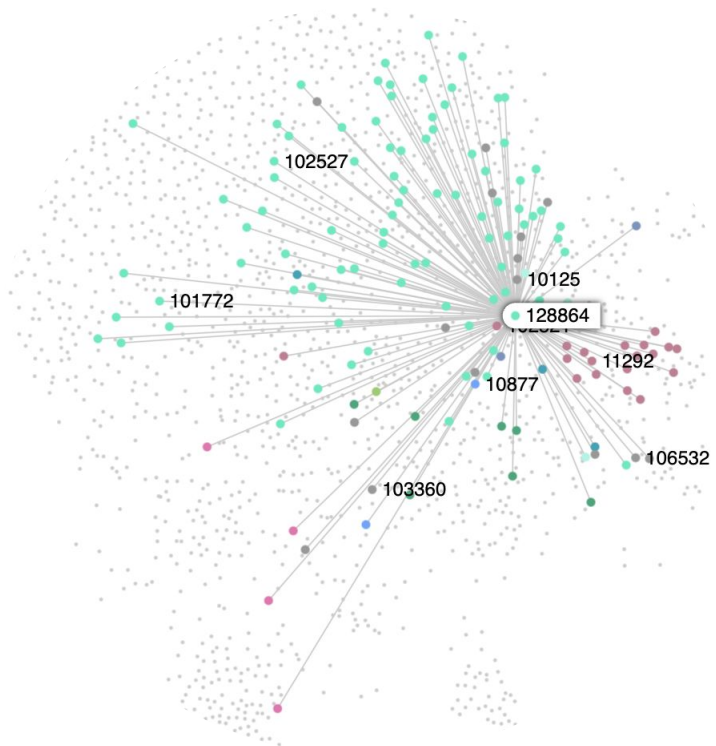
Following of top 10 English-speaking users:

Creators range in the number of following.

While english is the far majority, some have relatively follow relatively diverse nationalities:

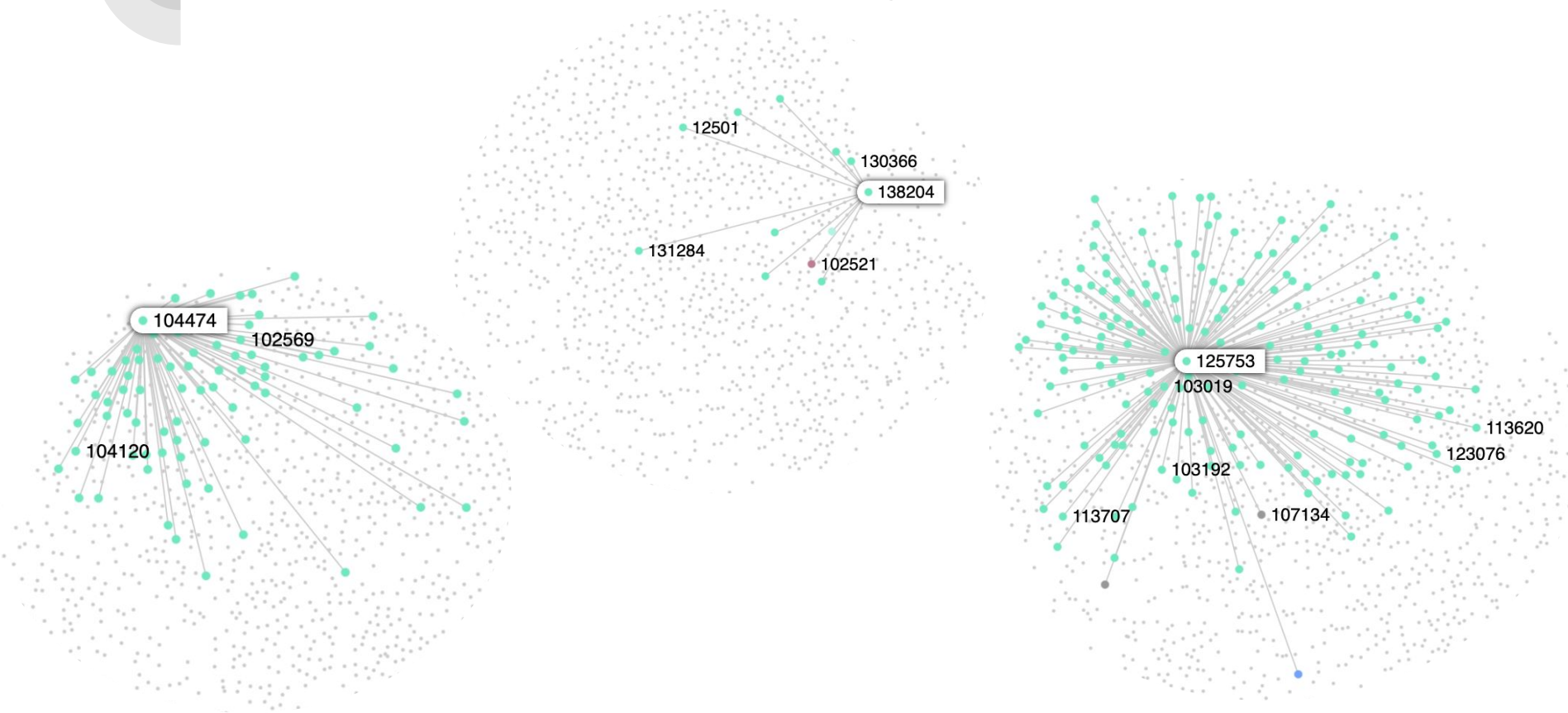


Legend:
English
Chinese
Russian
German
French
Portuguese
Spanish
Polish
Korean
Undetermined



- Legend:**
- English
 - Chinese
 - Russian
 - German
 - French
 - Portuguese
 - Spanish
 - Polish
 - Korean
 - Undetermined

Yet, some creators only follow other english creators:





Further analysis,

Who do the top 1% of English-speaking users' follow?

On average, **88.02%** of who they follow are other english-speaking users.

Top 3 non-English languages for english users to follow:

Germany (DE): **2.32%** of total following

Spain (ES): **1.66%** of total following

France (FR): **1.50%** of total following

So, within popular english users, there are strong between-ties of following other english users.

As Chinese and Russian are popular languages on Twitch, but do not seem to be popular with English users, we will analyze their data.



Russian language & the silo effect:

Who do Russian-speaking users follow?

On average, **Russian users makeup 91.35%** of who Russian-speaking users follow.

Top 5 most popular non-Russian languages for Russian users to follow:

1. English (EN): **12.30%**
2. German (DE): **0.49%**
3. Spanish (ES): **0.44%**
4. French (FR): **0.36%**



Chinese language and the silo effect:

Who do Chinese-speaking users follow?

On average, **88.98%** of who they follow are other Chinese-speaking users.

Top 4 most popular non-Chinese languages for Chinese users to follow:

1. English (EN): **11%**
2. Japanese (JA): **0.9%**
3. Korean (KO): **0.49%**
4. Spanish (ES): **0.33%**



As a reference, French user data::

Who do French-speaking users follow?

On average, **87.85%** of who they follow are other French-speaking users.

Top 4 most popular non-French languages for French users to follow:

1. English (EN): **18%**
2. German (DE): **0.41%**
3. Spanish (ES): **0.32%**
4. Polish (PT): **0.22%**



So, to analyze these findings:

Users of all languages (English, Russian, Chinese, and French), **tend to primarily follow those of their own language.**

However, we have seen that **for non-English users, the top non-native language followed is English, often making up a significant (+10%) proportion of total following**, often disproportionate to the other non-native languages followed.

This difference in experience means there is a potential english user silo effect, as English users seem less inclined to diversify their following base.

Interestingly, the **top 5 languages followed also differ based on region**, as seen with Chinese users following other east asian languages, versus French users following other european languages. These regional differences are interesting nuance in a social network context.

Implications...





Possible Implications:

- Language homogeneity
 - As the top creators on Twitch are english, these large creators typically make their content accessible to primarily English speakers, because they makeup the majority of the consumer base, which could result in a lack of diversity in content languages available on the platform, potentially excluding non-English speakers from certain communities or genres
- Limited Global Reach
 - As english speakers seem less engaged with other languages than non-english speakers are with english, this imbalance towards English speakers might make non-English speakers feel marginalized and dismissed, potentially impacting potential Twitch engagement in foreign countries. This might soon be perceived as an untapped market with potential to grow the entire platform
- Cultural Diversity, Representation, and Opportunities for Growth
 - Twitch should see this disparity as an opportunity to target these consumers who feel left out
 - Language-specific features, interfaces, and content recommendations, so that those of different backgrounds can find creators who are targeting them
 - After seeing the reliability and profitability of English-speaking Twitch communities, Twitch might be able to recreate that in other parts of the world
 - Targeted ad sponsorships in different languages could bring a lot of new revenue in for Twitch, as well as provide more insights into the consumer behaviors of Twitch users from different countries



Possible Implications:

- Community Cohesion and Affinity
 - The interconnected nature of the graph reflects that those who follow one major explicit creator are very likely to be following multiple, which reflects a sense of community around these creators who presumably make content that appeals to the same tastes
 - Knowledge of this huge block that tends to exhibit the same behaviors allows for Twitch to better understand their consumer base, so they might change their tendencies when it comes to content promotion and special events
- Content Moderation
 - Twitch should also remain cognizant that there is a risk of explicit content going in a direction where it becomes too extreme and begins alienating viewers
 - Twitch must find that balance between the two extremes in order not to stray too far from their origins, and that starts with investment and attention focused on content moderation with special focus on explicitness
- Platform Growth and Market Expansion
 - The domination of one consumer group means that there lies untapped potential on the opposite side of the spectrum
 - The promotion of family-friendly creators in addition to taking care of their existing viewer base could present a major opportunity for increased traffic, as well as new partnerships with different marketers and advertisers

Thank you





Citations

B. Rozemberczki and R. Sarkar. Twitch Gamers: a Dataset for Evaluating Proximity Preserving and Structural Role-based Node Embeddings. 2021.

Sarkar, R & Rózemerczki, B 2021, 'Twitch Gamers: a Dataset for Evaluating Proximity Preserving and Structural Role-based Node Embeddings', Paper presented at Workshop on Graph Learning Benchmarks @TheWebConf 2021, 16/04/21 - 16/04/21.