Twitch Popularity Analysis

Mateo Campoverde-Fordon, Rebecca Rauschmayer, Ian Vetter, Armaan Mehta

Introduction

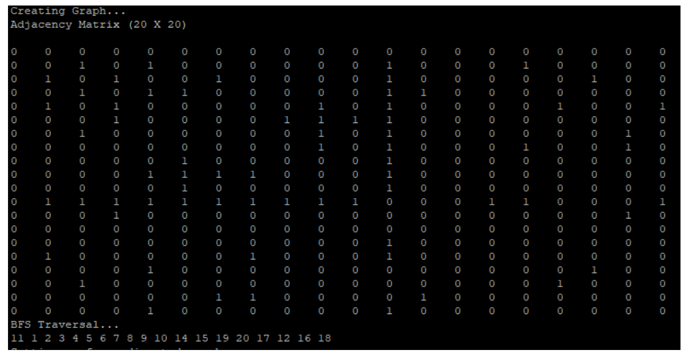
With social media growing faster than ever, everybody wants their spotlight. However, there is always the argument whether spending more time on a platform actually contributes to a higher following or view count. In this project, we aim to settle this argument by analyzing data from the popular streaming platform, Twitch. We plan to do so by using algorithms, such as PageRank, on an undirected graph we have created using the raw data provided by https://snap.stanford.edu/data/twitch_gamers.html.

The Graph Class

For our project, we created the Graph class, which describes an undirected graph of Twitch streamers. The constructor takes two inputs: a list of edges and a list of nodes. We then parsed through the data and converted the relevant strings (User ID and view count) to integers, which we stored in pairs, with User ID as the key and view count as the value. Each time we added a node, we updated the node count for the graph.

We chose to keep track of edges using an adjacency matrix. Since it is an undirected graph, if nodes A and B had a connection in the file, we connected nodes B and A in the adjacency matrix as well. This provided symmetry, however all of the edges are described in half of the matrix.

The BFS function takes in an integer start, which represents the user id the traversal starts from. We used a breadth-first traversal (BFS) for our data. One thing made clear by the BFS was how many users were connected to the starting node, as the IDs increment numerically until the edges of a node run out. This was especially useful when inputting different start points, as we could easily see the difference in the number of followers of the first user.

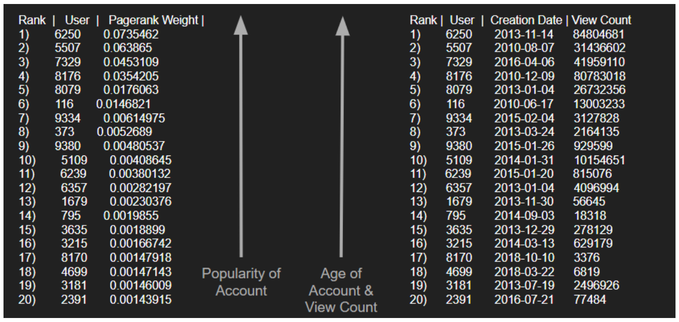

```
Creating Graph...
Adjacency Matrix (20 X 20)

0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  1  0  1  0  0  0  0  0  0  1  0  0  0  1  0  0  0  0
0  1  0  1  0  0  1  0  0  0  0  1  0  0  0  0  0  1  0  0
0  0  1  0  1  1  0  0  0  0  0  1  1  0  0  0  0  0  0  0
0  1  0  1  0  0  0  0  0  1  0  1  0  0  0  0  1  0  0  1
0  0  0  1  0  0  0  0  1  1  1  1  0  0  0  0  0  0  0  0
0  0  1  0  0  0  0  0  0  1  0  1  0  0  0  0  0  0  1  0
0  0  0  0  0  0  0  0  0  1  0  1  0  0  0  1  0  0  1  0
0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0
0  0  0  0  1  1  1  1  0  0  0  1  0  0  0  0  0  0  0  0
0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0
0  1  1  1  1  1  1  1  1  1  1  0  0  0  1  1  0  0  0  1
0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
0  1  0  0  0  0  0  1  0  0  0  1  0  0  0  0  0  0  0  0
0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
0  0  0  0  0  0  1  1  0  0  0  0  1  0  0  0  0  0  0  0
0  0  0  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
BFS Traversal...
11 1 2 3 4 5 6 7 8 9 10 14 15 19 20 17 12 16 18
```

FIGURE 1: Breadth-First Traversal Output

The Pagerank Class

This class is used to describe the PageRank algorithm. Consisting of the primary class method *powerPageRank(Graph &graph, int iterations)*, the class goes through the following steps of PageRank. We have decided to implement a power iteration method that uses a damping factor to more accurately and efficiently emulate human interactions with follower lists and how popularity is determined. This has a big O time complexity of O(n^2) as this algorithm multiplies a 2d matrix by itself multiple times to converge to probability values. In large datasets such as these, time complexity and memory efficiency is extremely important. Below, in Figure 2, are the results truncated for a smaller dataset that we have run locally. Although the creation date is only slightly related to its popularity, it is apparent the view count is more correlated. We believe that if we ran on a complete edge dataset or if the dataset was truncated using a search algorithm, we may have seen a higher correlation.



Rank	User	Pagerank Weight
1)	6250	0.0735462
2)	5507	0.063865
3)	7329	0.0453109
4)	8176	0.0354205
5)	8079	0.0176063
6)	116	0.0146821
7)	9334	0.00614975
8)	373	0.0052689
9)	9380	0.00480537
10)	5109	0.00408645
11)	6239	0.00380132
12)	6357	0.00282197
13)	1679	0.00230376
14)	795	0.0019855
15)	3635	0.0018899
16)	3215	0.00166742
17)	8170	0.00147918
18)	4699	0.00147143
19)	3181	0.00146009
20)	2391	0.00143915

Rank	User	Creation Date	View Count
1)	6250	2013-11-14	84804681
2)	5507	2010-08-07	31436602
3)	7329	2016-04-06	41959110
4)	8176	2010-12-09	80783018
5)	8079	2013-01-04	26732356
6)	116	2010-06-17	13003233
7)	9334	2015-02-04	3127828
8)	373	2013-03-24	2164135
9)	9380	2015-01-26	929599
10)	5109	2014-01-31	10154651
11)	6239	2015-01-20	815076
12)	6357	2013-01-04	4096994
13)	1679	2013-11-30	56645
14)	795	2014-09-03	18318
15)	3635	2013-12-29	278129
16)	3215	2014-03-13	629179
17)	8170	2018-10-10	3376
18)	4699	2018-03-22	6819
19)	3181	2013-07-19	2496926
20)	2391	2016-07-21	77484

FIGURE 2: Pagerank Algorithm Output

The Forcegraph Class

This class is used to describe the force-directed graph class, which maps the nodes to coordinates and produces a PNG image. This is done by applying attractive and repelling forces on the nodes until equilibrium is reached.

The attractive force was calculated using Hooke's Law, as F = k (r - r_0) where k is the spring constant and r_0 is the spring rest length.

The repelling force was calculated using Coulomb's Law, as F = k /r^2 where k is the Coulomb constant.

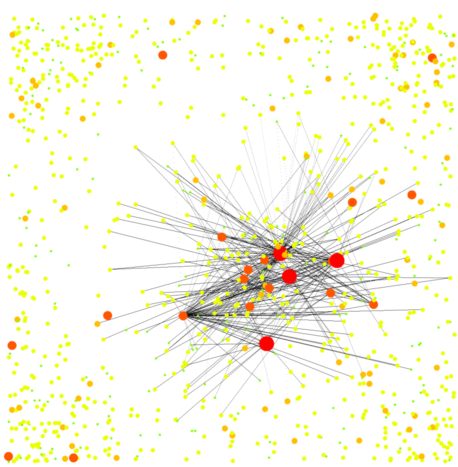


FIGURE 3: Force-Directed Graph Output

Findings

We found that the view count was not significantly related to the ranking of the user. There was no clear correlation between the creation date decreasing and the ranking based on followers increasing. There was, however, a clearer correlation between creation and view count, which makes sense, as the longer people have to view something, the more views it gets.

There was also a slight correlation between ranking and view count, but it was not as clear as the correlation between view count and creation date.

Presentation Video

https://youtu.be/-H7rJKbcLnk