Networks and Markets

Hw4 submission

Open Ended Project

By:

Team 22

Ariel Chiskis 322442112

Alon Polsky 206530461

Anna Petrenko 320460306

The topical issue we choose is the asymmetry that relationships have in social networks. Not every node has the same influence over the world, and that is not only because a Node doesn't have the same number of relationships or because the other participants of its relationships have little influence, but only because it has little to no influence on the other participants of its relationships, even though there's some influence.

For example, read this story about a post of Kylie Jenner. It has about 286 thousand likes, it'll be completely naïve to think that all of the interactions users had, when they liked this post, had the same significance, as Kylie's interaction, when she posted that post, and it would be even more naïve to think that Kylie influence all of this people as much as their mom influence them.

To be more precise, we tried to fix the imprecision that Scaled PageRank algorithm has for capturing the diverging powers that people have within relationships.

At this point it would be best if you read our answer for question 8(d) in Part 4 of this assignment, the algorithm we implemented is exactly the one suggested in that answer and the terminology we are going to use is defined in that answer.

The data-set we used is a Social Network of Hyperlinks between subreddits, in this Social Network, every node represents a subreddit. It's a directed graph, where there's an edge from a to b for every link from a to b.

The data-set 55863 nodes and 858490 edges between them.

That network has a lot of sinks (43695), so we used the next equation for running the refined PageRank algorithm:

$$score(p) = \frac{\epsilon}{n} + (1-\epsilon)\left(\sum_{n_{p',p}>0} \frac{n_{p',p}}{n_{p'}} Score(p') + \sum_{n_{p'}=0} \frac{1}{number\ of\ nodes} Score(p')\right)$$

The ϵ we used is $\frac{1}{7}$, for every nodes $,p, p'$, n_p is the number of edges coming out of p, $n_{p',p}$ is the number of edges from p' to p.

We ran Scaled PageRank for comparison too, each of the algorithms ran for 10 iterations.

The results we got from Scaled PageRank:

- Average influence: $6.57229 \cdot 10^{-6}$
- Standard deviation: $4.80593 \cdot 10^{-5}$
- Minimal influence $2.12648 \cdot 10^{-6}$
- Maximal influence 0.00574
- Sum of influences: 0.44152
- Runtime: 28 minutes and 59 seconds.

The results we got from the improved PageRank are:

- Average influence: $1.45679 \cdot 10^{-5}$

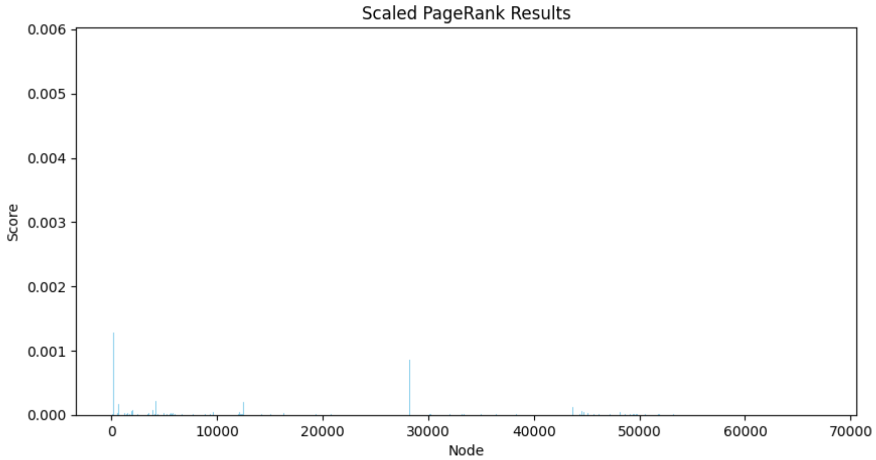
- Standard deviation: 0.00013
- Minimal influence: $4.46581 \cdot 10^{-6}$
- Maximal influence: 0.01812
- Sum of influences: 0.97867
- Runtime: 38 minutes and 48 seconds.

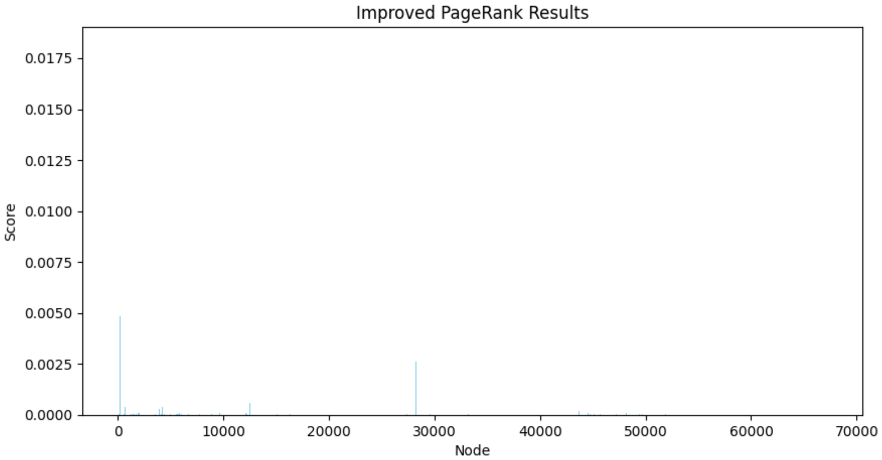
The most noticeable difference in these statistics is the difference between the sum of influences, Scaled PageRank's sum is 0.44152 while the sum of influences of the improved version is close to 1, 0.97867, this is a symptom of the term we added to the improved version's definition of a score:

$$\sum_{n_{p'}=0} \frac{1}{number\ of\ nodes} Score(p')$$

and Scaled PageRank's inability to cope with sinks.
By the definition of the algorithm and pure math the sum of scores in the result of the improved version should have stayed 1, but due to floating-point imprecision we are not able to get exactly 1.
The fact that the standard deviation of the improved version was more than two times bigger is probably a symptom of the fact that we had less loss of influence (the sum of scores stayed close to one) so scores had a larger spectrum they could vary over, accordingly, we would also like to say that this higher variance is a symptom of the refined version's better ability to express influence, but we can only say that as an hypothesis.
We wanted to under the difference between the results better, to the that we had to find subreddits that had very different scores in the two results.
That was the hardest part for 2 reasons:

- The amount of influence that subreddit a, has in a relationship between subreddits is not a function of how many subreddits had hyperlinks pointing towards a, but it is most certainly corelated to that number.
  So, most of the scores were relatively big, small or mid in one result iff they had that property in the other.
  Plots to demonstrate that:





- Because of Scaled PageRank's lack of ability to cope with sinks, its scores were much smaller than the results of the improved algorithm.

We eventually came up with the following method to determine the difference between scores;
Let $(r_1, \dots, r_n)$ be one of the results and $(r_1', \dots, r_n')$ be the other, We used:

$$\left| \frac{r_i}{\max\{r_i\}_{i=1}^{n}} - \frac{r_i'}{\max\{r_i\}_{i=1}^{n'}} \right|$$

That is, we used the distance between normalized results to measure how different are the scores, this method still isn't perfect because it does not solve the first problem. At the last moment we came with a measurement that could handle that problem better, presumably:

$$\frac{\max\{r_i, r_i'\}}{\min\{r_i, r_i'\}}$$

But, that idea came to our minds only in the last moment and we did not have time to implement it.
This is the results we got, using this technique:

- 10 nodes with biggest ditance between normalized Scaled pageRank score and normalized improved PageRank score (from highest difference to lowest):

  pics, todayilearned, wtf, movies, askreddit, mildlyinteresting, writingprompts, explainlikeimfive, gifs, showerthoughts

What caught our eye in that is that only two of those subreddits: movies and writingprompts are topical and are not subreddits for shitposting/random-info, also all of the subreddits had lower normalized improved PageRank score than the Scaled PageRank score, except for askreddit that had a higher normalized improved PageRank score. (The exact scores and distances can be found in the program's output, we decided to not include it to make this write-up more reader friendly.)

This led us to a hypothesis: subreddits that are more topical/theme-related and are focused around a certain subject that a community can have genuine interest in will generally have higher improved PageRank score, and subreddits that are for shitposting/random-info focused will have much lower normalized improved PageRank than scaled PageRank score, because people scatter random memes and shitposts all around the internet, but their sources vary and meme pages don't have well founded communities around them, that probably makes shitposting/random-info subreddits much less influential if you capture strength of links between subreddits, like the improved algorithm does and Scaled PageRank fails to do.

We also computed the 10 most influential subreddits according to Scaled PageRank (from most influential to least):

iama, askreddit, pics, funny, videos, todayilearned, the_donald, gaming, music, gifs

and the 10 most influential subreddits by the improved algorithm:

iama, askreddit, pics, funny, videos, the_donald, gaming, music, todayilearned, worldnews

To try to back-up our hypothesis.

As you can see, the_donald, a subreddit of memes that promote Donald Trump, a subreddit that is very much focused on a certain theme and has a well-founded community around it, is ranked much higher in the improved algorithm.

Also games and music that are theme related got ranked higher than todayilearned, that is about random-data, similarly, worldnews got ranked higher than gifs.

We understand that this is very far from proving our hypothesis, but this is the best we could do with our time and budget.

A quick note: It can take the script four hours to run at worst, on our set-up. We ran it on Linux, with python 3.11.6.