Neil Kumar
4/26/18
Professor Chaintreau

Social Networks Data Challenge 2 Part 2 README

Code for new algorithm: “datachallenge2_part2.py”

New Algorithm: $score(u, v) = \log_{10}(|\Gamma(u) \cap \Gamma(v)| + 1)(\sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\log(|\Gamma(z)|})$

I chose the above algorithm because I read “The Link Prediction Problem for Social Networks” by Liben-Nowell and Kleinberg and noticed that both Adamic-Adar and the “Common Neighbors” ($|\Gamma(u) \cap \Gamma(v)|$) heuristics performed well. I wanted to mix the two heuristics because to me it makes sense logically to connect two individuals who have many friends in common, so I tested variants of their combination until I found the above algorithm that worked well. The number of friends in common is logged by 10 to give more weight to the Adamic-Adar score, helping the algorithm perform better, and I added 1 inside the log to avoid giving a score of 0 to nodes that only had 1 node in common.

New Dataset: “polblogs.gml” (raw data) and “polblogs.txt” (information about data)

Link: http://www-personal.umich.edu/~mejn/netdata/
The above gml file contains data about political blogs where the “edges” are links to the other blogs in the directed graph. The dataset contains 1490 nodes (blogs) and 19090 edges (links). Each node has a “value” associated with it, where a value of “0” means the blog is liberal and a value of “1” means the blog is conservative. I calculated fairness with regards to whether “liberal” blogs are fairly represented in the resulting predictions.

Performance of New Algorithm vs. Adamic-Adar:

Running the above .py file prints the following results:

```
Fairness(IG Data_Adamic-Adar): 0.0556605348379
Fairness(IG Data_My Algorithm): 0.0794663341529


Fairness(Blog Data_Adamic-Adar): 1.22298221614
Fairness(Blog Data_My Algorithm): 1.19972640219


Accuracy(IG Data_Adamic-Adar): 0.00063824355374
Accuracy(IG Data_My Algorithm): 0.000574419198366


Accuracy(Blog Data_Adamic-Adar): 0.0429544264013
Accuracy(Blog Data_My Algorithm): 0.0437401781037
```

Although my algorithm performs worse with regards to fairness and accuracy on the Instagram data, it performs better in both regards on the blog data. The accuracy increased about 2% on the blog data, and the fairness improved (decreased) on the blog data as well. I believe these discrepancies are due to the nature of the graphs themselves.