CSC 466 Lab 6 - Pagerank Report

Aidan Barbieux
Cal Poly San Luis Obispo
abarbieu@calpoly.edu

December 4, 2021

Contents

List of Tables and Figures

1 Implementation

This implementation of Page Rank was relatively unmodified. The most major change was adding an edge from sink nodes (those with no outgoing edges) to themselves to ensure no division by zero. This can be viewed as a page having a link to itself, and the web searcher staying on the page until becoming bored and going to a different random page. Alternatively an edge could be added to the incoming nodes, representing a back button, but this made less methodological sense and resulted in Page Ranks summing to more than 1.

Overall the implementation looked as follows:

- Read in data in either SNAP or SMALL format
- Convert node/actor names to a unique identifier (index in a list of unique names)
- Create scipy lil sparse matrix from this converted list of edges
- Fix sink nodes by connecting them to themselves with set_diag, convert to scipy csr sparse matrix (more efficient for further computation)
- Scale adjacency matrix by outgoing edges (divide each row by it's sum)
- Iterate Page Rank until maximal difference between two ranks on previous iteration is less than epsilon
- Print results by converting id to name

2 Results

2.1 Settings

All runs were done using a d of 0.85 and an epsilon of 0.00001 where maximal difference in rank was used as the stoppage condition (found by plotting 100 iterations and seeing where the difference leveled out)

2.2 Output

All output is included in the appendix, the SNAP datasets chosen were amazon0505, p2p-Gnutella, and wiki-Vote.

2.3 Observations

It is hard to say what ranking would be correct, but given the ranking of the NCAA_football dataset and the lesmis dataset I believe the ranks to be correct. The football dataset produced ranks consistent with general team performance (google) and in line roughly with the handout results. The results from the lesmis dataset were roughly in order of main characters (google). The shape of

the results for the datasets showed a few main actors with a roughly exponential drop in rank values after that (seen in amazon and football datasets: Figure 4, Figure 3) the only exception was the dolphin dataset which showed a roughly linear drop in ranks: Figure: 5.

3 Summary

I was happy with the consistency of Page Rank on the datasets, with the one exception of the dolphin dataset which seemed to produce different results. There is not another way for me to assess ranks without information about the actors other than observing that actors with a high pagerank tended to have the most incoming edges, which correlates with more prestige if some of those edges are also prestigious, leading me to believe the ranking is good.

4 Performance Evaluation

Process and read time is shown in Figure 2. Dataset size was chosen to be the number of edges, as this takes into account the number of nodes and connectivity naturally. This timing was on a slow computer as otherwise many of the results were 0 seconds. On my personal computer the longest process time was for the amazon dataset which took 2 seconds to process and 18 seconds to read the data (with 24 iterations of page rank). Note that the graph is in logarithmic scale for both x and y, meaning that the runtime complexity is nearly linear it seems, though with so few datapoints and no replication it is hard to tell. Number of iterations proved to not be interesting and was thus not graphed, all were between 10 and 30. This difference is seen in Figures 6 8 and 7, where maximal page rank difference is plotted against iteration, showing that each had a different shape and converged at different points, but all had a tail of minimal difference past 30 iterations and all converged after 10 iterations.

5 Appendix

5.1 README

Figure 1: README

CSC 466 Lab 6 - PageRank ¶

Group Members

1. Aidan Barbieux - abarbieu@calpoly.edu

Instructions

pageRank.ipynb

contains code for pageRank.py in blocks for easy data manipulation and code needed to create graphics

pageRank_analysis.ipynb

uses pageRank.py to analyse runtime and create more graphics

pageRank.py

```
usage: pageRank.py [-h] [--d D] [--epsilon EPSILON] datafile {SMALL,SNAP}

Page Rank

positional arguments:
  datafile           .csv or .txt file with data in SNAP or SMALL format given
                     by lab spec
  {SMALL,SNAP}       SMALL or SNAP to determine in which type the data is
                     given

optional arguments:
  -h, --help         show this help message and exit
  --d D              probability of staying on the same page, usually between
                     0.7 and 0.95
  --epsilon EPSILON  maximal difference between pagerank iterations at which
                     to finish
```

Figure 2: Timing of reading and processing for all datasets (including soc-livejournal) on slow computer (otherwise it was 0.0sec). Dataset size is the number of edges, which already takes into account the number of actors

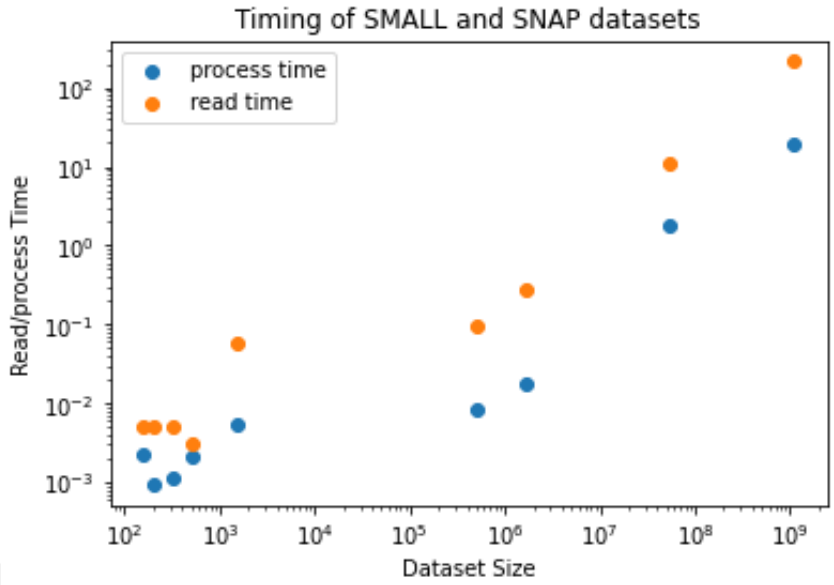

Figure 3: shape of ranks for amazon 0505 dataset

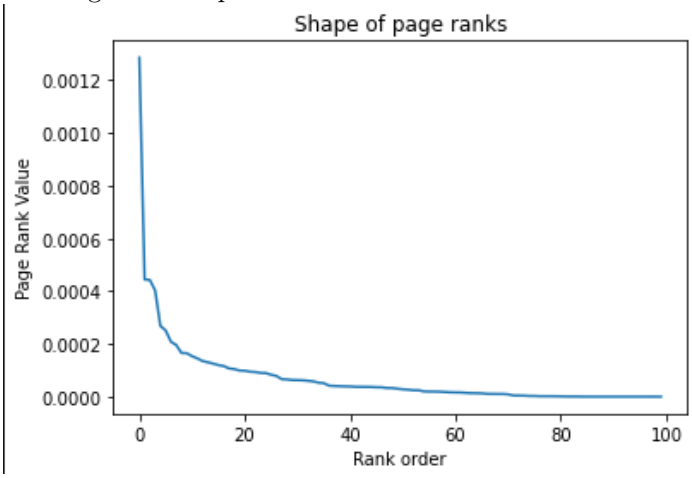

Figure 4: shape of ranks for football dataset

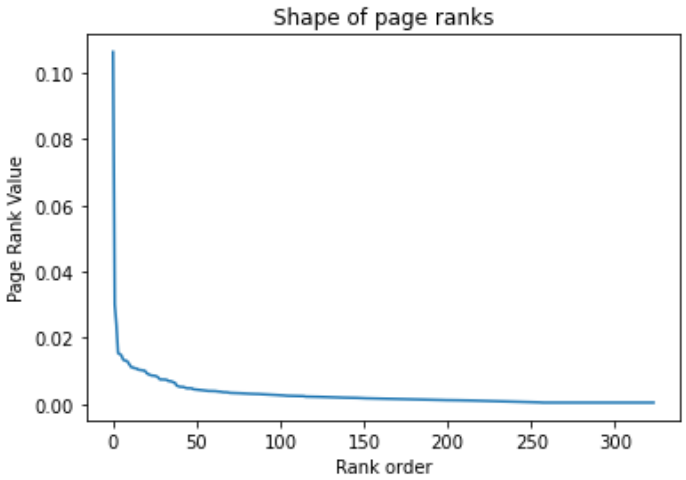

Figure 5: shape of ranks for dolphin dataset

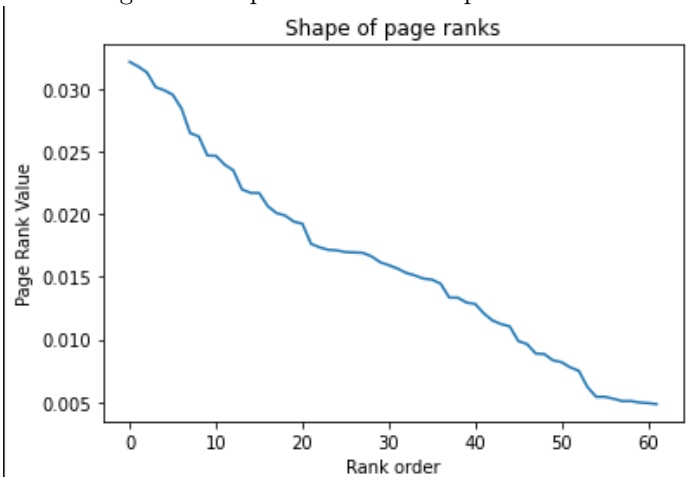

Figure 6: maximal rank difference for the amazon0505 data set by iteration

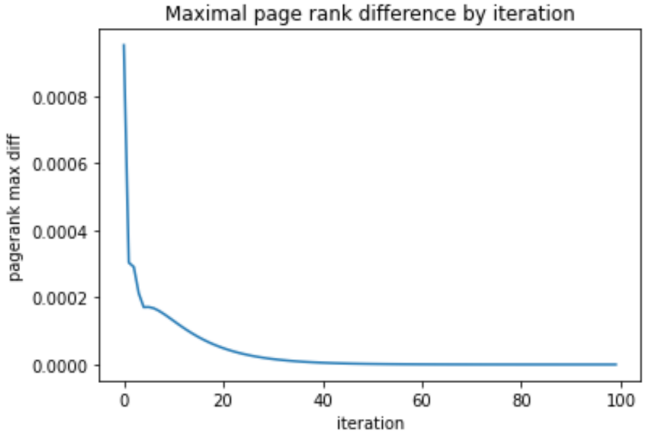

Figure 7: maximal rank difference for the wiki-Vote dataset by iteration

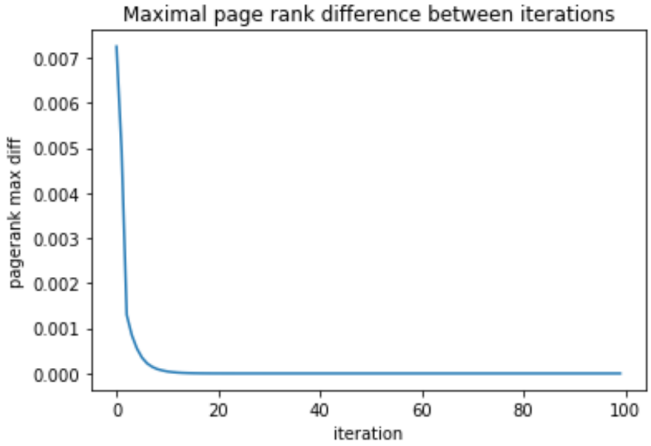

Figure 8: maximal rank difference for the NCAA_football dataset by iteration

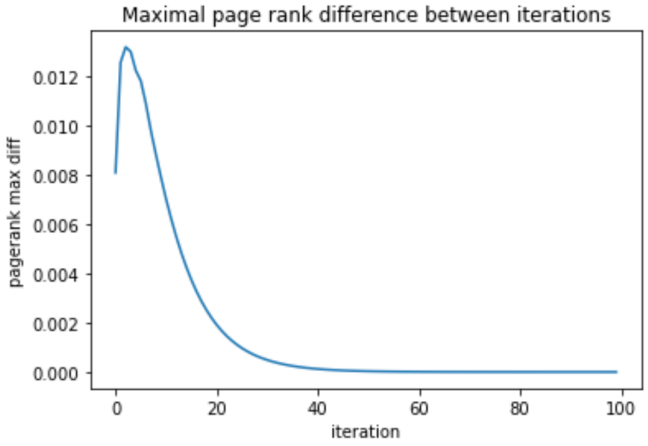

Table 9: NCAA_football top 50 Page Ranks

rank	actor	pagerank
0	Utah	0.106268
1	Mississippi	0.0300093
2	Florida	0.02394
3	Oklahoma	0.0153808
4	Texas Tech	0.0151081
5	Wake Forest	0.0147016
6	Alabama	0.0134531
7	Virginia Tech	0.0131276
8	Oregon State	0.013109
9	Texas	0.0125917
10	Vanderbilt	0.0117777
11	Boston College	0.0110844
12	James Madison	0.0110482
13	Richmond	0.0107975
14	Georgia Tech	0.0107667
15	North Carolina Pembroke	0.010418
16	Montana	0.0103147
17	USC	0.0102598
18	Virginia	0.0101443
19	South Carolina	0.0100715
20	North Carolina	0.00940904
21	Florida State	0.0090454
22	Duke	0.00887004
23	Maryland	0.00864738
24	Miami (FL)	0.00858325
25	North Carolina State	0.00854738
26	West Virginia	0.00836997
27	Clemson	0.00802929
28	Weber State	0.0074548
29	Georgia	0.00744681
30	East Carolina	0.0074376
31	Villanova	0.0074001
32	TCU	0.00734758
33	Pittsburgh	0.00699556
34	Cincinnati	0.00690889
35	Penn State	0.00684329
36	LSU	0.00657866
37	Iowa	0.00645222
38	Oregon	0.00558221
39	California	0.005388
40	Franklin	0.00522238
41	Tulsa	0.00517981
42	Rutgers	0.00517343
43	Appalachian State	0.00500158
44	Connecticut	0.00479101
45	Boise State	0.00478814
46	Navy	0.00477159
47	Northwestern	0.00476417
48	Brown	0.00448872
49	New Hampshire	0.00443932

Table 10: dolphins top 50 Page Ranks

rank	actor	pagerank
0	Grin	0.0321313
1	Jet	0.031753
2	Trigger	0.0312874
3	Web	0.0301203
4	SN4	0.0298639
5	Topless	0.0295013
6	Scabs	0.0284122
7	Patchback	0.0264473
8	Gallatin	0.0261803
9	Beescratch	0.0246627
10	Kringel	0.0246325
11	SN63	0.0239303
12	Feather	0.0234795
13	SN9	0.02196
14	Stripes	0.0216831
15	Upbang	0.0216665
16	SN100	0.0206118
17	DN21	0.0200717
18	Haecksel	0.0198752
19	Jonah	0.0193867
20	TR99	0.0192238
21	SN96	0.0176133
22	TR77	0.0173349
23	Number1	0.0171396
24	Double	0.0170925
25	Beak	0.0169592
26	MN105	0.0169315
27	MN83	0.016898
28	Hook	0.0166202
29	SN90	0.0161511
30	Shmuddel	0.0159145
31	DN63	0.0156477
32	PL	0.0153019
33	Fish	0.015103
34	Oscar	0.0148445
35	Zap	0.0147628
36	DN16	0.0144411
37	Bumper	0.0133337
38	Ripplefluke	0.0133196
39	Knit	0.012933
40	Thumper	0.0128265
41	TSN103	0.0120682
42	Mus	0.0115108
43	Notch	0.0112162
44	Zipfel	0.0110358
45	MN60	0.00986039
46	CCL	0.00962589
47	TR88	0.00887464
48	TR120	0.00882383
49	Wave	0.00833248

Table 11: karate Page Ranks

rank	actor	pagerank
0	34	0.100911
1	1	0.0970005
2	33	0.0716872
3	3	0.057078
4	2	0.052877
5	32	0.0371577
6	4	0.0358607
7	24	0.0315213
8	9	0.0297661
9	14	0.0295374
10	7	0.0291144
11	6	0.0291144
12	30	0.0262874
13	28	0.0256389
14	31	0.0245901
15	8	0.0244914
16	11	0.0219801
17	5	0.0219801
18	25	0.021075
19	26	0.021005
20	20	0.0196054
21	29	0.019573
22	17	0.0167857
23	27	0.0150435
24	13	0.0146456
25	18	0.0145594
26	22	0.0145594
27	15	0.0145359
28	19	0.0145359
29	21	0.0145359
30	23	0.0145359
31	16	0.0145359
32	10	0.0143094
33	12	0.00956523

Table 12: lesmis top 50 Page Ranks

rank	actor	pagerank
0	Valjean	0.0754315
1	Myriel	0.0427872
2	Gavroche	0.0357662
3	Marius	0.0308941
4	Javert	0.0303027
5	Thenardier	0.0279262
6	Fantine	0.0270226
7	Enjolras	0.0218812
8	Cosette	0.0206111
9	MmeThenardier	0.019501
10	Bossuet	0.0189587
11	Courfeyrac	0.0185775
12	Eponine	0.0177936
13	Mabeuf	0.0174773
14	Joly	0.017199
15	Bahorel	0.017199
16	Babet	0.0166917
17	Gueulemer	0.0166917
18	Claquesous	0.0165609
19	MlleGillenormand	0.0162601
20	Feuilly	0.0158913
21	Combeferre	0.0158913
22	Tholomyes	0.0156472
23	Bamatabois	0.0155767
24	Montparnasse	0.0151708
25	Gillenormand	0.0149574
26	Grantaire	0.0144559
27	Prouvaire	0.0131452
28	Blacheville	0.012618
29	Listolier	0.012618
30	Favourite	0.012618
31	Dahlia	0.012618
32	Zephine	0.012618
33	Fameuil	0.012618
34	Brevet	0.012425
35	Chenildieu	0.012425
36	Champmathieu	0.012425
37	Cochepaille	0.012425
38	Judge	0.012425
39	Brujon	0.0118665
40	Fauchelevent	0.0116382
41	MmeHucheloup	0.0106893
42	MlleBaptistine	0.0102777
43	MmeMagloire	0.0102777
44	Simplice	0.0090737
45	LtGillenormand	0.00871351
46	MmeBurgon	0.00780544
47	Pontmercy	0.007368
48	Woman2	0.00683691
49	Toussaint	0.00683691

Table 13: amazon0505 top 50 Page Ranks

rank	actor	pagerank
0	593	0.00173076
1	595	0.00143695
2	591	0.00142064
3	11042	0.00140948
4	89	0.00128614
5	590	0.00101795
6	972	0.000986295
7	976	0.000835935
8	974	0.000816369
9	975	0.000762586
10	978	0.000747219
11	120	0.000732212
12	977	0.000713468
13	634	0.000712168
14	2612	0.000694737
15	598	0.00062331
16	597	0.000551489
17	585	0.000537382
18	162	0.000528745
19	4455	0.000474783
20	596	0.000474423
21	88	0.000444805
22	44	0.000442959
23	1196	0.000436452
24	4458	0.000432857
25	594	0.000407363
26	39	0.000403304
27	4460	0.000390197
28	605	0.000389397
29	157	0.000384029
30	587	0.000375747
31	2611	0.000359857
32	4461	0.000347681
33	158	0.000334256
34	4454	0.000325776
35	4459	0.00032537
36	578	0.000320785
37	2264	0.000309737
38	10999	0.000301942
39	121	0.00029999
40	7241	0.000299413
41	1850	0.000297991
42	592	0.000283213
43	2613	0.000276207
44	80633	0.000271623
45	8548	0.000270982
46	37	0.000268481
47	582	0.000267868
48	12642	0.000260919
49	830	0.000259928

Table 14: p2pGnutella top 50 Page Ranks

rank	actor	pagerank
0	389	0.00152297
1	679	0.00117428
2	672	0.000887854
3	1297	0.000887348
4	324	0.000862699
5	1929	0.000842504
6	4057	0.000743227
7	1066	0.000700692
8	737	0.000693811
9	343	0.000687549
10	3543	0.000675097
11	1110	0.000644355
12	1019	0.000640216
13	1681	0.000638478
14	1673	0.000632304
15	919	0.000616628
16	20	0.000612178
17	1606	0.000597015
18	5356	0.00057494
19	485	0.000570306
20	1465	0.000556148
21	1027	0.000552718
22	891	0.000548896
23	744	0.000546504
24	395	0.0005463
25	3773	0.000544878
26	1139	0.000543968
27	827	0.000538983
28	1211	0.000531912
29	660	0.000530458
30	3171	0.000528408
31	5051	0.000527069
32	869	0.000524056
33	161	0.000523744
34	4506	0.000520956
35	1471	0.000511013
36	1796	0.000510139
37	955	0.000507011
38	917	0.000505822
39	2922	0.000504915
40	181	0.000500448
41	640	0.00050013
42	1902	0.000498876
43	916	0.000489924
44	3933	0.000486462
45	825	0.000485598
46	2921	0.000485546
47	2833	0.000482982
48	2302	0.000480125
49	1256	0.000479639

Table 15: Top 50 Page Ranks from wiki-Vote

rank	actor	pagerank
0	2625	0.00913368
1	2470	0.00702433
2	7553	0.0060355
3	1186	0.0056651
4	7620	0.00537428
5	5412	0.00533591
6	7632	0.00530533
7	4875	0.0052122
8	6832	0.00491766
9	2066	0.00477141
10	8293	0.00474563
11	214	0.00459574
12	4735	0.00450371
13	271	0.00369063
14	5210	0.00356158
15	8163	0.00350675
16	1842	0.00343728
17	1026	0.00326302
18	3537	0.00318332
19	3117	0.00318138
20	2643	0.00311999
21	299	0.00307235
22	7699	0.00296291
23	3755	0.00295299
24	5459	0.00291944
25	1633	0.00282958
26	7890	0.00282107
27	4256	0.00281394
28	1726	0.00277207
29	4247	0.00268769
30	8294	0.00253837
31	4402	0.00252786
32	3408	0.00252407
33	7809	0.00249076
34	5288	0.0024268
35	7961	0.00235938
36	3650	0.00228157
37	1412	0.00226617
38	7478	0.00225942
39	1956	0.00223469
40	5963	0.00222512
41	7414	0.00222291
42	7214	0.0021639
43	1453	0.00214538
44	7803	0.00214386
45	8295	0.00209878
46	4011	0.00203875
47	3265	0.00203691
48	2877	0.00202496
49	7908	0.00201242