James Abundis
jlabundi@calpoly.edu
Thomas Fahrner
tfahrner@calpoly.edu
Lec: CSC 466-01
Lab: CSC 466-02

Lab 5

PageRank Implementation:

We implemented the standard PageRank. At first, we built a graph data structure with heavy redundancy. While this was effective for the smaller datasets, we ran into performance issues with the large datasets. We changed our algorithm to adopt a Node data structure shown here:

```
class Node:
    def __init__(self, name):
        self.name = name
        self.num_out = 0
        self.in_node_names = set()
```

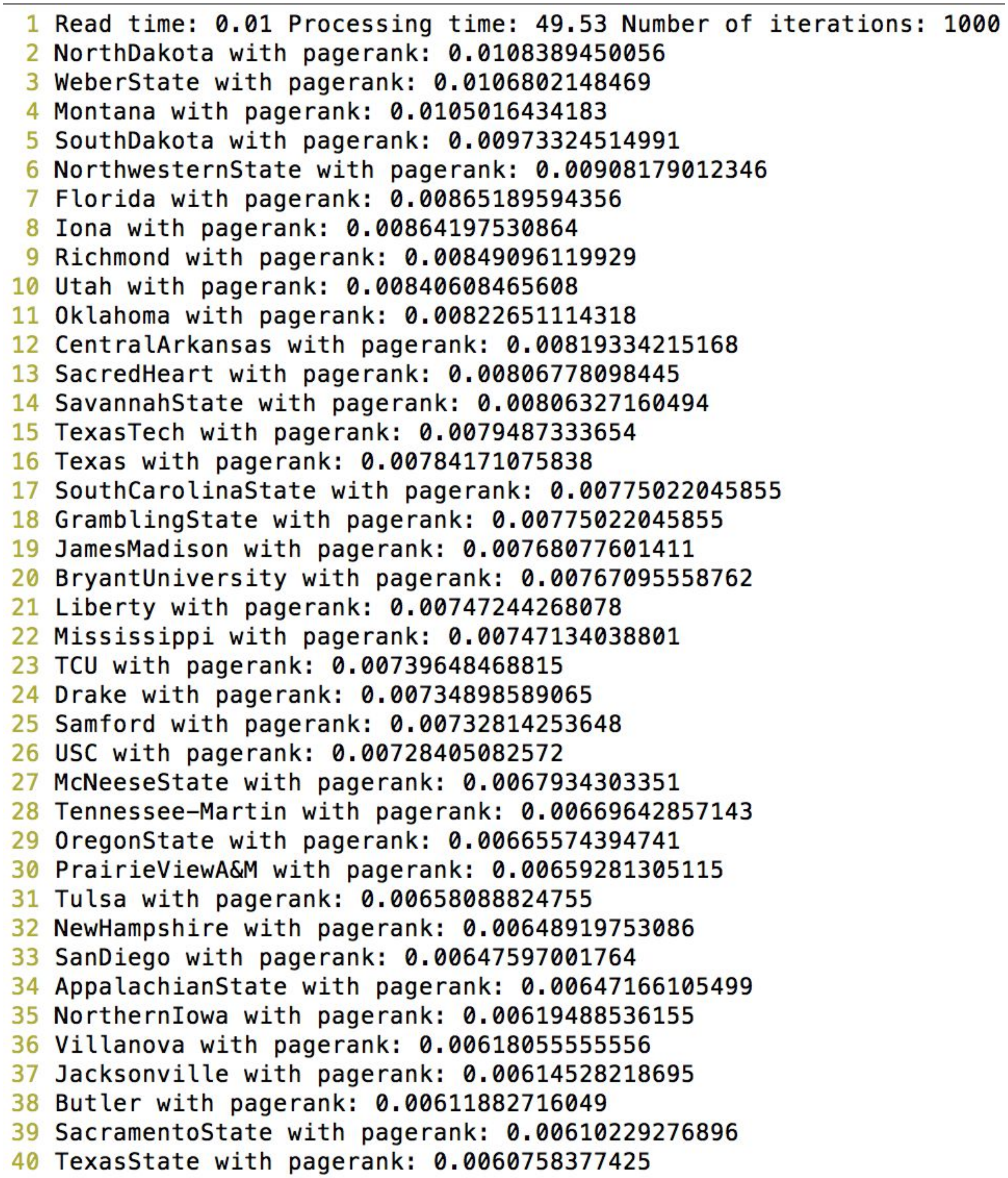
This data structure allowed us to perform very efficient iterations of PageRank. Our algorithm for building the graph representation handles the discrepancy between directed/undirected graphs.

The lack of ground truth made it difficult to determine the values of the parameters (d: probability of traveling to a connected node; epsilon/number of iterations: stopping condition). We split this parameter search into two parts for the small/large datasets respectively.

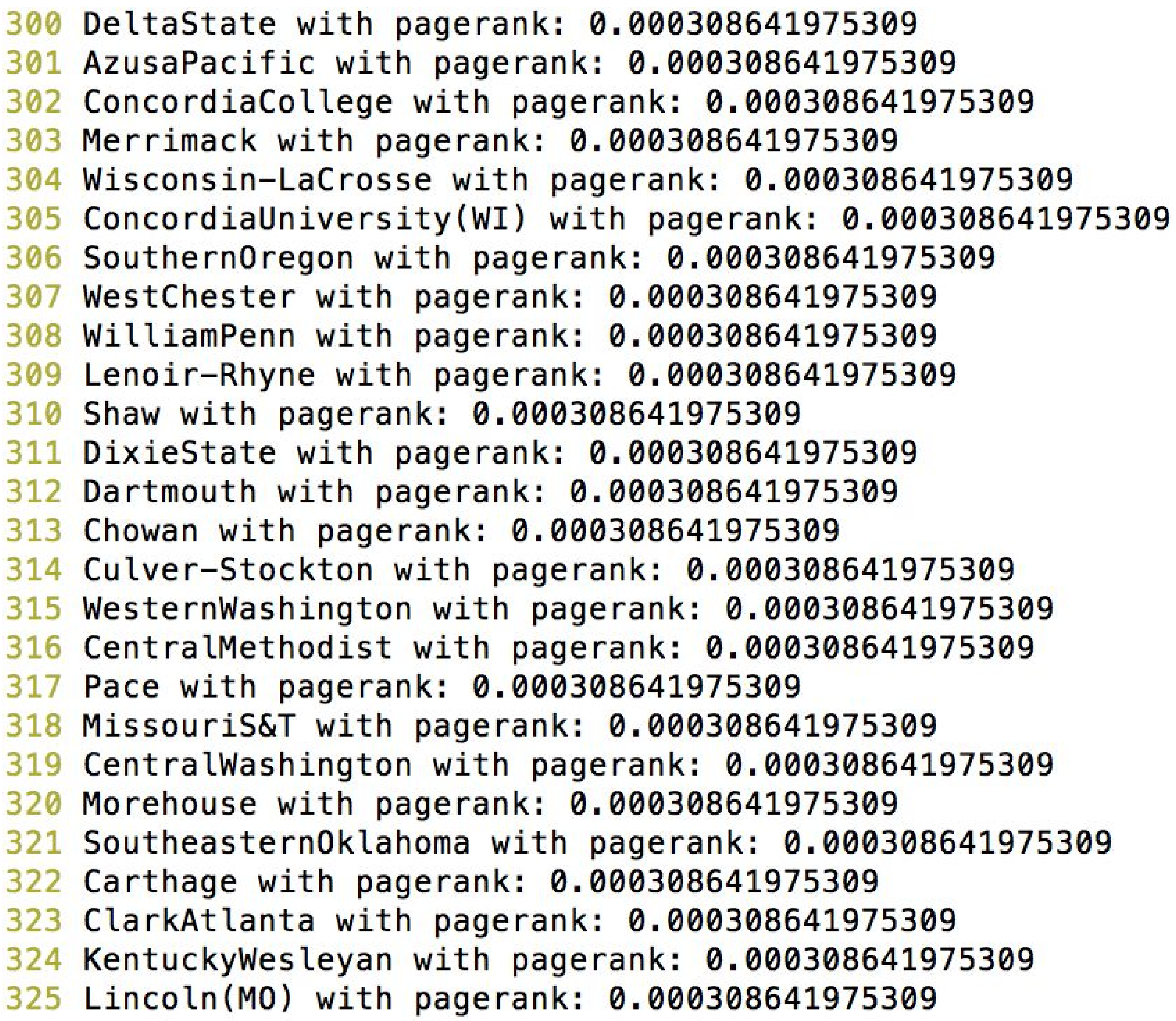
- Small datasets:
    - d = 0.9; we found this worked well with a quick sanity check on the results
    - num_iterations = 1000; this was effectively limitless because performance was not an issue on these datasets. There was no need to parameter search and instead we chose to just guarantee PageRank convergence.
- Large datasets:
    - d = 0.9; we figured that this parameter would suit the large datasets just as well. While it is contextually dependent, we did not find a strong argument to alter this parameter.
    - num_iterations = n; this was dependent on the performance evaluation (see below)

Results:

NCAA Football:



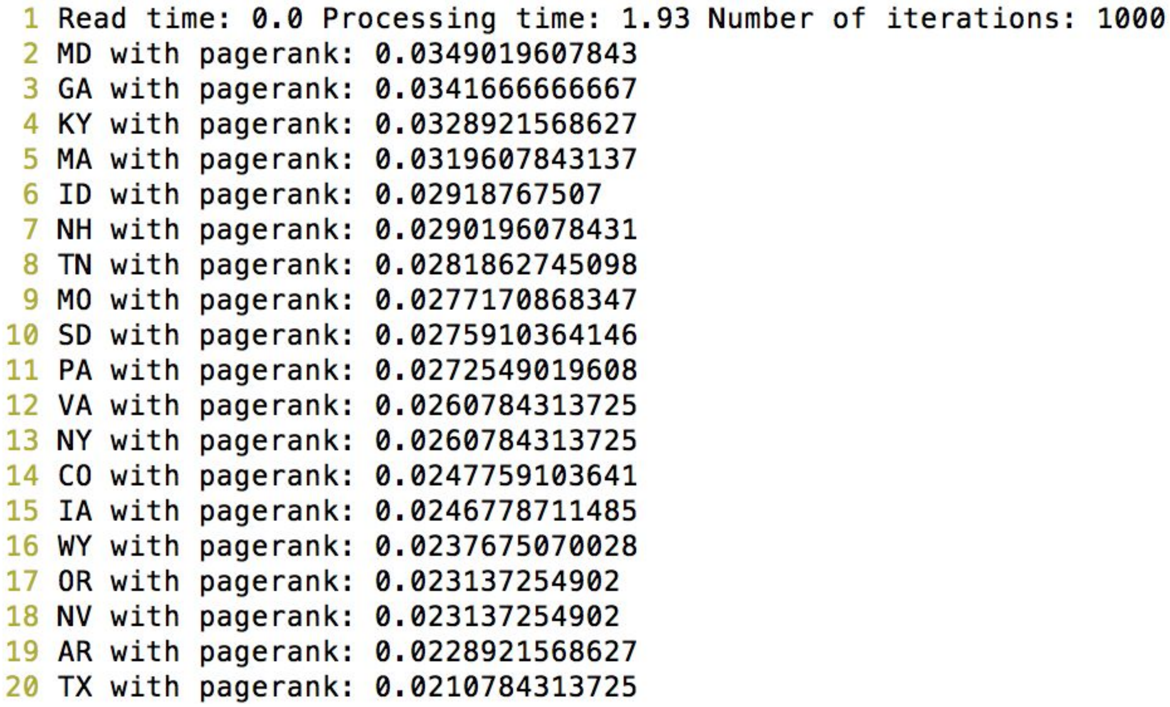
```
Read time: 0.01 Processing time: 49.53 Number of iterations: 1000
NorthDakota with pagerank: 0.0108389450056
WeberState with pagerank: 0.0106802148469
Montana with pagerank: 0.0105016434183
SouthDakota with pagerank: 0.00973324514991
NorthwesternState with pagerank: 0.00908179012346
Florida with pagerank: 0.00865189594356
Iona with pagerank: 0.00864197530864
Richmond with pagerank: 0.00849096119929
Utah with pagerank: 0.00840608465608
Oklahoma with pagerank: 0.00822651114318
CentralArkansas with pagerank: 0.00819334215168
SacredHeart with pagerank: 0.00806778098445
SavannahState with pagerank: 0.00806327160494
TexasTech with pagerank: 0.0079487333654
Texas with pagerank: 0.00784171075838
SouthCarolinaState with pagerank: 0.00775022045855
GramblingState with pagerank: 0.00775022045855
JamesMadison with pagerank: 0.00768077601411
BryantUniversity with pagerank: 0.00767095558762
Liberty with pagerank: 0.00747244268078
Mississippi with pagerank: 0.00747134038801
TCU with pagerank: 0.00739648468815
Drake with pagerank: 0.00734898589065
Samford with pagerank: 0.00732814253648
USC with pagerank: 0.00728405082572
McNeeseState with pagerank: 0.0067934303351
Tennessee-Martin with pagerank: 0.00669642857143
OregonState with pagerank: 0.00665574394741
PrairieViewA&M with pagerank: 0.00659281305115
Tulsa with pagerank: 0.00658088824755
NewHampshire with pagerank: 0.00648919753086
SanDiego with pagerank: 0.00647597001764
AppalachianState with pagerank: 0.00647166105499
NorthernIowa with pagerank: 0.00619488536155
Villanova with pagerank: 0.00618055555556
Jacksonville with pagerank: 0.00614528218695
Butler with pagerank: 0.00611882716049
SacramentoState with pagerank: 0.00610229276896
TexasState with pagerank: 0.0060758377425
```



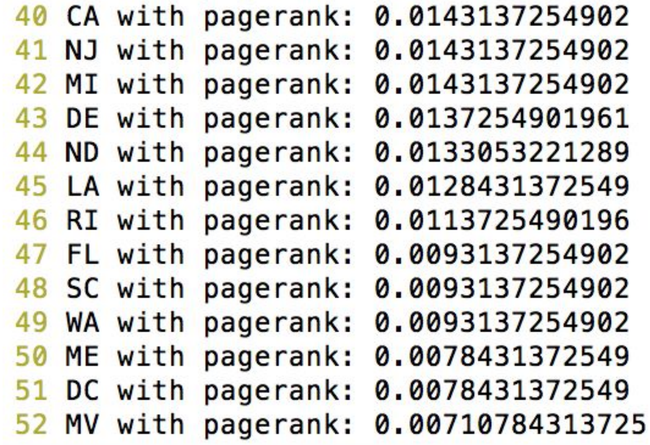
```
300 DeltaState with pagerank: 0.000308641975309
301 AzusaPacific with pagerank: 0.000308641975309
302 ConcordiaCollege with pagerank: 0.000308641975309
303 Merrimack with pagerank: 0.000308641975309
304 Wisconsin-LaCrosse with pagerank: 0.000308641975309
305 ConcordiaUniversity(WI) with pagerank: 0.000308641975309
306 SouthernOregon with pagerank: 0.000308641975309
307 WestChester with pagerank: 0.000308641975309
308 WilliamPenn with pagerank: 0.000308641975309
309 Lenoir-Rhyne with pagerank: 0.000308641975309
310 Shaw with pagerank: 0.000308641975309
311 DixieState with pagerank: 0.000308641975309
312 Dartmouth with pagerank: 0.000308641975309
313 Chowan with pagerank: 0.000308641975309
314 Culver-Stockton with pagerank: 0.000308641975309
315 WesternWashington with pagerank: 0.000308641975309
316 CentralMethodist with pagerank: 0.000308641975309
317 Pace with pagerank: 0.000308641975309
318 MissouriS&T with pagerank: 0.000308641975309
319 CentralWashington with pagerank: 0.000308641975309
320 Morehouse with pagerank: 0.000308641975309
321 SoutheasternOklahoma with pagerank: 0.000308641975309
322 Carthage with pagerank: 0.000308641975309
323 ClarkAtlanta with pagerank: 0.000308641975309
324 KentuckyWesleyan with pagerank: 0.000308641975309
325 Lincoln(MO) with pagerank: 0.000308641975309
```

Here we show the highest and lowest ranked results. If we look at North Dakota in the dataset, it often appears first (they won a lot of games). The same follows with Weber State and the others ranked highly. If we look at some of the low ranked teams, they often had just one game represented in the dataset. With one game, often a loss, we expect their PageRank to be very low. Overall, a cursory look indicates that PageRank performed well on this dataset.

State Borders:



```
 1 Read time: 0.0 Processing time: 1.93 Number of iterations: 1000
 2 MD with pagerank: 0.0349019607843
 3 GA with pagerank: 0.0341666666667
 4 KY with pagerank: 0.0328921568627
 5 MA with pagerank: 0.0319607843137
 6 ID with pagerank: 0.02918767507
 7 NH with pagerank: 0.0290196078431
 8 TN with pagerank: 0.0281862745098
 9 MO with pagerank: 0.0277170868347
10 SD with pagerank: 0.0275910364146
11 PA with pagerank: 0.0272549019608
12 VA with pagerank: 0.0260784313725
13 NY with pagerank: 0.0260784313725
14 CO with pagerank: 0.0247759103641
15 IA with pagerank: 0.0246778711485
16 WY with pagerank: 0.0237675070028
17 OR with pagerank: 0.023137254902
18 NV with pagerank: 0.023137254902
19 AR with pagerank: 0.0228921568627
20 TX with pagerank: 0.0210784313725
```

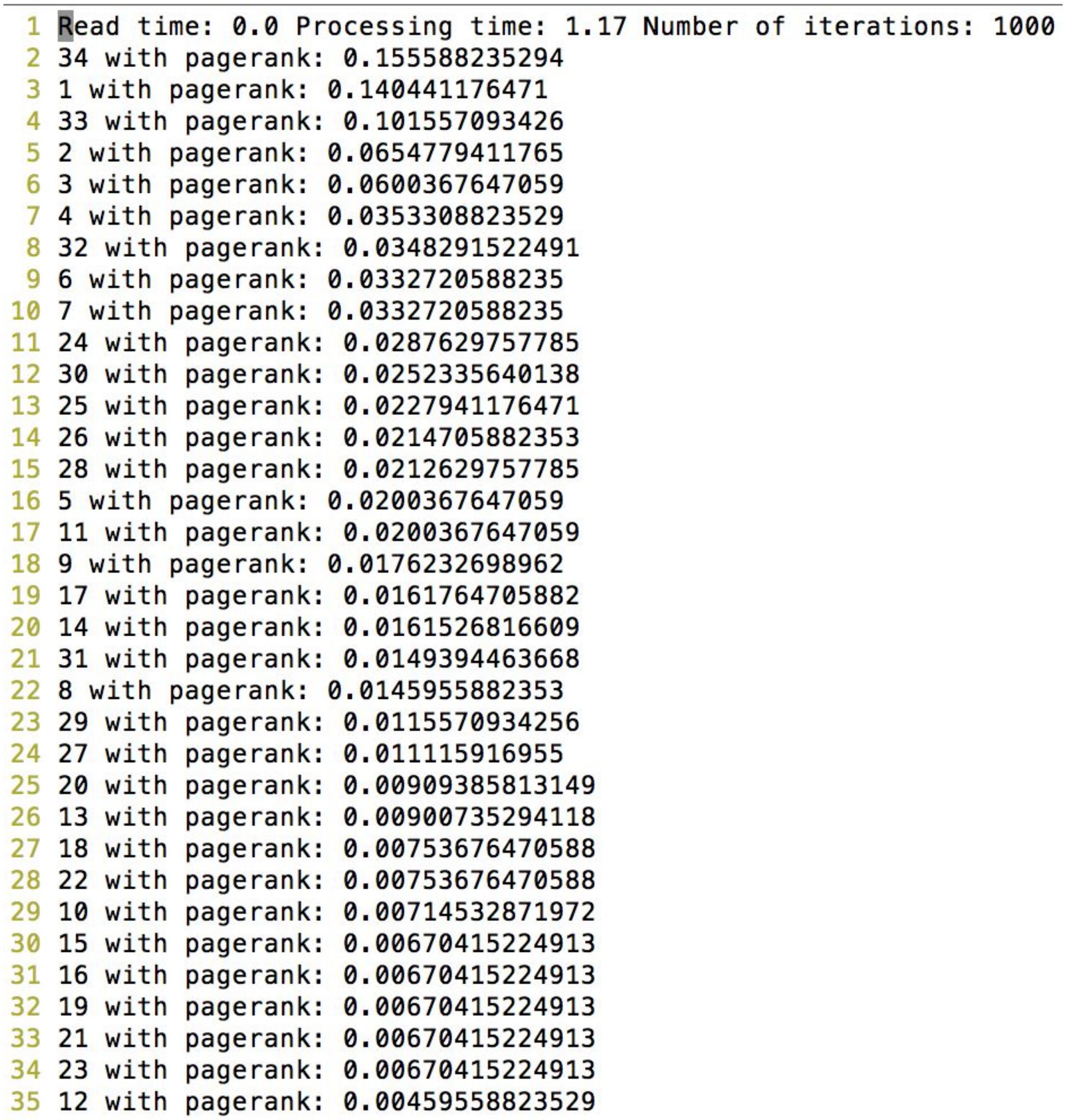


```
40 CA with pagerank: 0.0143137254902
41 NJ with pagerank: 0.0143137254902
42 MI with pagerank: 0.0143137254902
43 DE with pagerank: 0.0137254901961
44 ND with pagerank: 0.0133053221289
45 LA with pagerank: 0.0128431372549
46 RI with pagerank: 0.0113725490196
47 FL with pagerank: 0.0093137254902
48 SC with pagerank: 0.0093137254902
49 WA with pagerank: 0.0093137254902
50 ME with pagerank: 0.0078431372549
51 DC with pagerank: 0.0078431372549
52 MV with pagerank: 0.00710784313725
```

Click on a state to learn More
Washington
Oregon
Idaho
Montana
North Dakota
Minnesota
Wyoming
South Dakota
Wisconsin
Michigan
New Hampshire
Vermont
Maine
Massa chuset
New York
Rho Islar
Connecticu
Pennsylvania
New Jersey
Delaware
Maryland
Nevada
California
Utah
Colorado
Nebraska
Iowa
Illinois
Indiana
Ohio
West Virginia
Virginia
Kentucky
Kansas
Missouri
Tennessee
North Carolina
South Carolina
Arizona
New Mexico
Oklahoma
Arkansas
Mississippi
Alabama
Georgia
Texas
Louisiana
Florida
Alaska
Hawaii

Here we show the highest and lowest ranked results. It's clear from the top results that states with a high number of borders receive a high PageRank. Conversely, the low ranked results also make sense. PageRank seems to work well on this dataset.

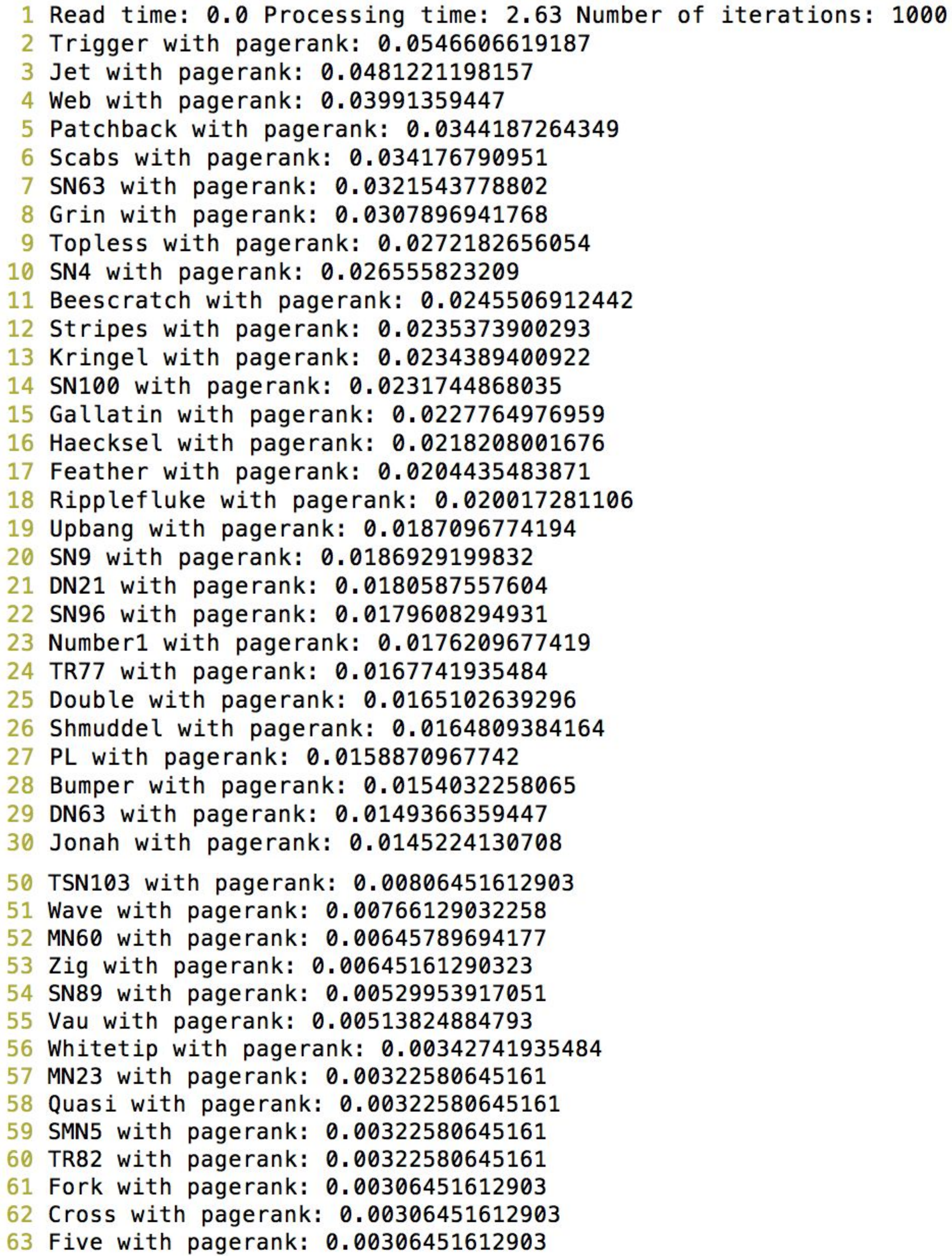
Karate:



```
Read time: 0.0 Processing time: 1.17 Number of iterations: 1000
34 with pagerank: 0.155588235294
1 with pagerank: 0.140441176471
33 with pagerank: 0.101557093426
2 with pagerank: 0.0654779411765
3 with pagerank: 0.0600367647059
4 with pagerank: 0.0353308823529
32 with pagerank: 0.0348291522491
6 with pagerank: 0.0332720588235
7 with pagerank: 0.0332720588235
24 with pagerank: 0.0287629757785
30 with pagerank: 0.0252335640138
25 with pagerank: 0.0227941176471
26 with pagerank: 0.0214705882353
28 with pagerank: 0.0212629757785
5 with pagerank: 0.0200367647059
11 with pagerank: 0.0200367647059
9 with pagerank: 0.0176232698962
17 with pagerank: 0.0161764705882
14 with pagerank: 0.0161526816609
31 with pagerank: 0.0149394463668
8 with pagerank: 0.0145955882353
29 with pagerank: 0.0115570934256
27 with pagerank: 0.011115916955
20 with pagerank: 0.00909385813149
13 with pagerank: 0.00900735294118
18 with pagerank: 0.00753676470588
22 with pagerank: 0.00753676470588
10 with pagerank: 0.00714532871972
15 with pagerank: 0.00670415224913
16 with pagerank: 0.00670415224913
19 with pagerank: 0.00670415224913
21 with pagerank: 0.00670415224913
23 with pagerank: 0.00670415224913
12 with pagerank: 0.00459558823529
```

Karate student #34 is friends with about half the class - what a popular guy! His PageRank reflects this. Student #12 has one friend - #1. While #1 is very popular, his friend #12 is not. PageRank seems to work well on this dataset.

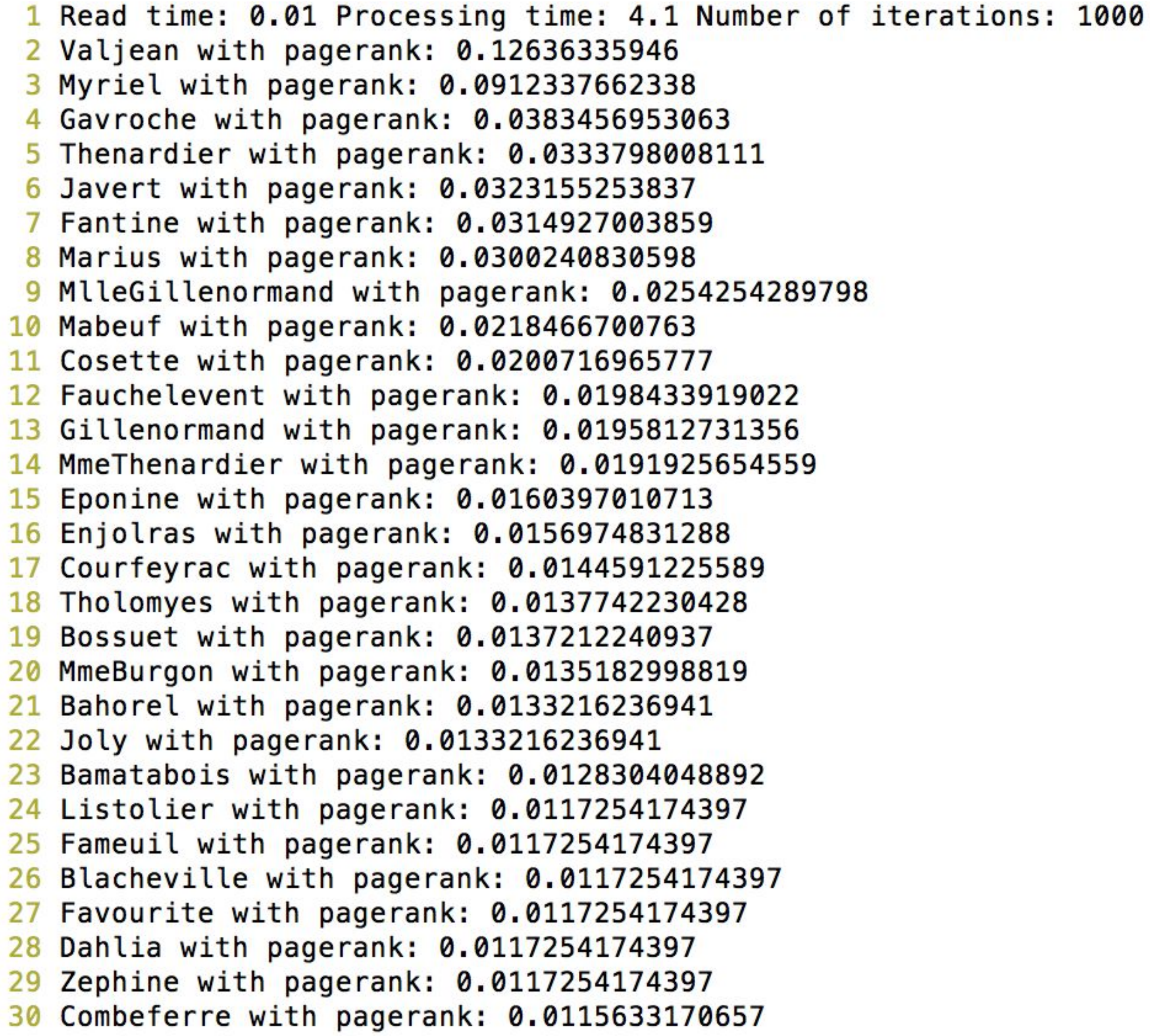
Dolphins:



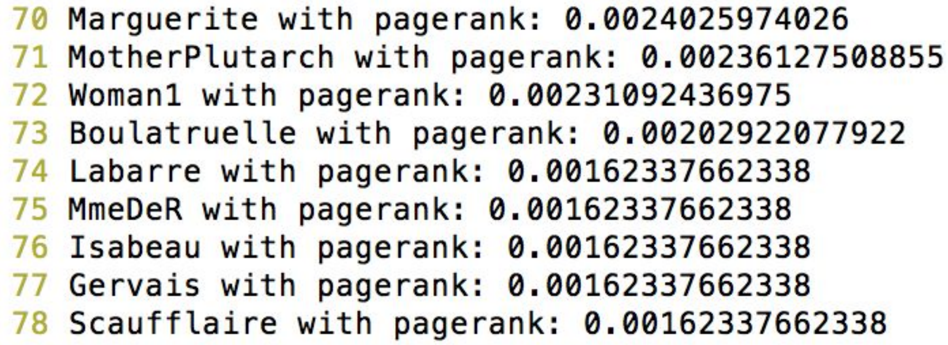
```
Read time: 0.0 Processing time: 2.63 Number of iterations: 1000
Trigger with pagerank: 0.0546606619187
Jet with pagerank: 0.0481221198157
Web with pagerank: 0.03991359447
Patchback with pagerank: 0.0344187264349
Scabs with pagerank: 0.034176790951
SN63 with pagerank: 0.0321543778802
Grin with pagerank: 0.0307896941768
Topless with pagerank: 0.0272182656054
SN4 with pagerank: 0.026555823209
Beescratch with pagerank: 0.0245506912442
Stripes with pagerank: 0.0235373900293
Kringel with pagerank: 0.0234389400922
SN100 with pagerank: 0.0231744868035
Gallatin with pagerank: 0.0227764976959
Haecksel with pagerank: 0.0218208001676
Feather with pagerank: 0.0204435483871
Ripplefluke with pagerank: 0.020017281106
Upbang with pagerank: 0.0187096774194
SN9 with pagerank: 0.0186929199832
DN21 with pagerank: 0.0180587557604
SN96 with pagerank: 0.0179608294931
Number1 with pagerank: 0.0176209677419
TR77 with pagerank: 0.0167741935484
Double with pagerank: 0.0165102639296
Shmuddel with pagerank: 0.0164809384164
PL with pagerank: 0.0158870967742
Bumper with pagerank: 0.0154032258065
DN63 with pagerank: 0.0149366359447
Jonah with pagerank: 0.0145224130708

TSN103 with pagerank: 0.00806451612903
Wave with pagerank: 0.00766129032258
MN60 with pagerank: 0.00645789694177
Zig with pagerank: 0.00645161290323
SN89 with pagerank: 0.00529953917051
Vau with pagerank: 0.00513824884793
Whitetip with pagerank: 0.00342741935484
MN23 with pagerank: 0.00322580645161
Quasi with pagerank: 0.00322580645161
SMN5 with pagerank: 0.00322580645161
TR82 with pagerank: 0.00322580645161
Fork with pagerank: 0.00306451612903
Cross with pagerank: 0.00306451612903
Five with pagerank: 0.00306451612903
```

Trigger has a lot of friends; Five has only one friend - Trigger. This is a similar social network analysis as the Karate Kids. PageRank seems to work well on this dataset.

Les Miserables:



```
Read time: 0.01 Processing time: 4.1 Number of iterations: 1000
Valjean with pagerank: 0.12636335946
Myriel with pagerank: 0.0912337662338
Gavroche with pagerank: 0.0383456953063
Thenardier with pagerank: 0.0333798008111
Javert with pagerank: 0.0323155253837
Fantine with pagerank: 0.0314927003859
Marius with pagerank: 0.0300240830598
MlleGillenormand with pagerank: 0.0254254289798
Mabeuf with pagerank: 0.0218466700763
Cosette with pagerank: 0.0200716965777
Fauchelevent with pagerank: 0.0198433919022
Gillenormand with pagerank: 0.0195812731356
MmeThenardier with pagerank: 0.0191925654559
Eponine with pagerank: 0.0160397010713
Enjolras with pagerank: 0.0156974831288
Courfeyrac with pagerank: 0.0144591225589
Tholomyes with pagerank: 0.0137742230428
Bossuet with pagerank: 0.0137212240937
MmeBurgon with pagerank: 0.0135182998819
Bahorel with pagerank: 0.0133216236941
Joly with pagerank: 0.0133216236941
Bamatabois with pagerank: 0.0128304048892
Listolier with pagerank: 0.0117254174397
Fameuil with pagerank: 0.0117254174397
Blacheville with pagerank: 0.0117254174397
Favourite with pagerank: 0.0117254174397
Dahlia with pagerank: 0.0117254174397
Zephine with pagerank: 0.0117254174397
Combeferre with pagerank: 0.0115633170657
```



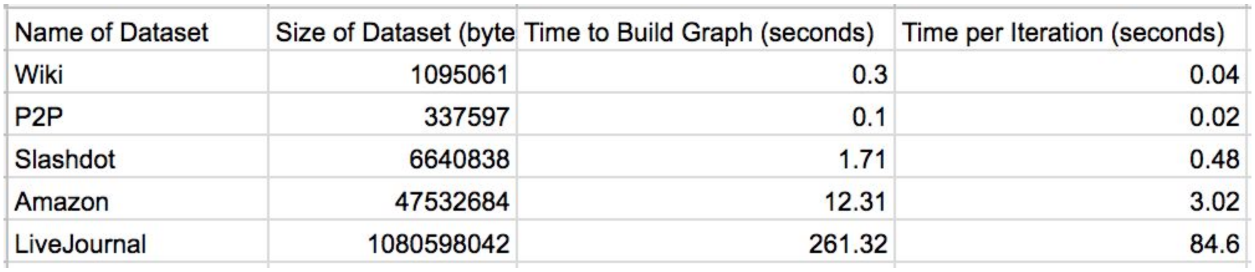
```
Marguerite with pagerank: 0.0024025974026
MotherPlutarch with pagerank: 0.00236127508855
Woman1 with pagerank: 0.00231092436975
Boulatruelle with pagerank: 0.00202922077922
Labarre with pagerank: 0.00162337662338
MmeDeR with pagerank: 0.00162337662338
Isabeau with pagerank: 0.00162337662338
Gervais with pagerank: 0.00162337662338
Scaufflaire with pagerank: 0.00162337662338
```

Valjean has a lot of co-occurrences with other characters; Scaufflaire has only one. This is a similar social network analysis as the Karate Kids. PageRank seems to work well on this dataset.

Summary:

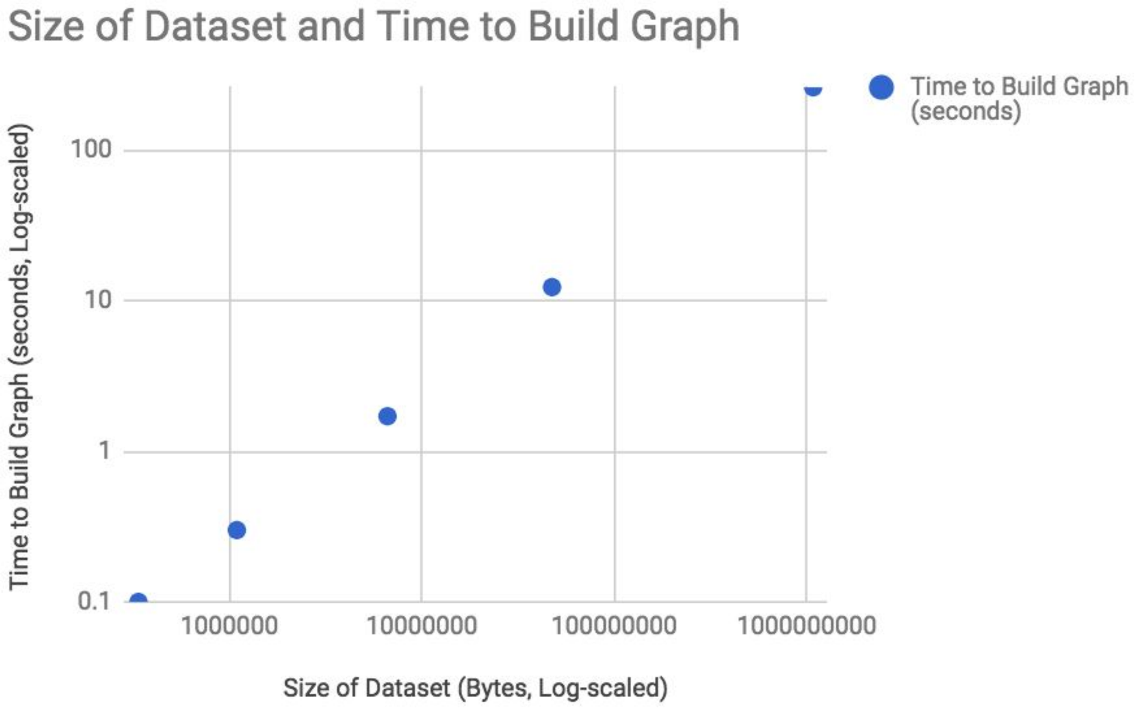
PageRank worked pretty well on every one of the small datasets. While the ranking is good, it is also important to look at the quantitative values given (the PageRank). For instance, in the Les Mes dataset, Valjean has a PageRank equal to three times the third highest ranked (Gavroche). That said, Gavroche and Valjean have a reasonably similar amount of representation in the dataset. The decay (from first to second, etc.) is less stark in the other datasets.

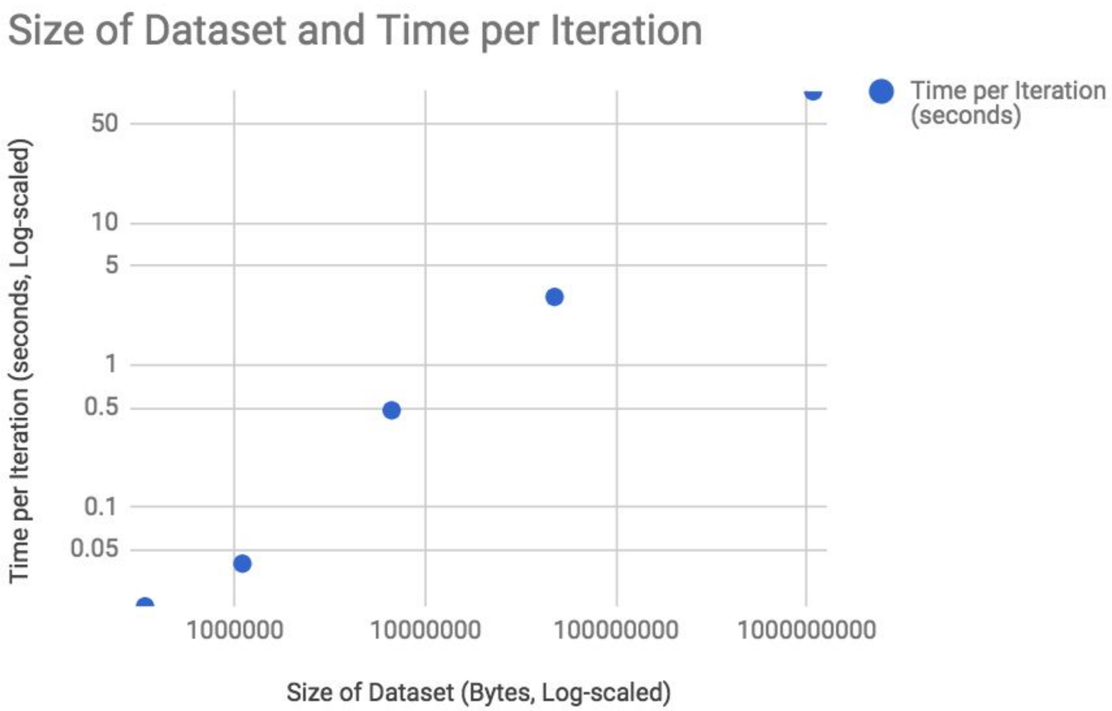
Performance Evaluation:



Name of Dataset	Size of Dataset (byte	Time to Build Graph (seconds)	Time per Iteration (seconds)
Wiki	1095061	0.3	0.04
P2P	337597	0.1	0.02
Slashdot	6640838	1.71	0.48
Amazon	47532684	12.31	3.02
LiveJournal	1080598042	261.32	84.6

Log-log Plots:

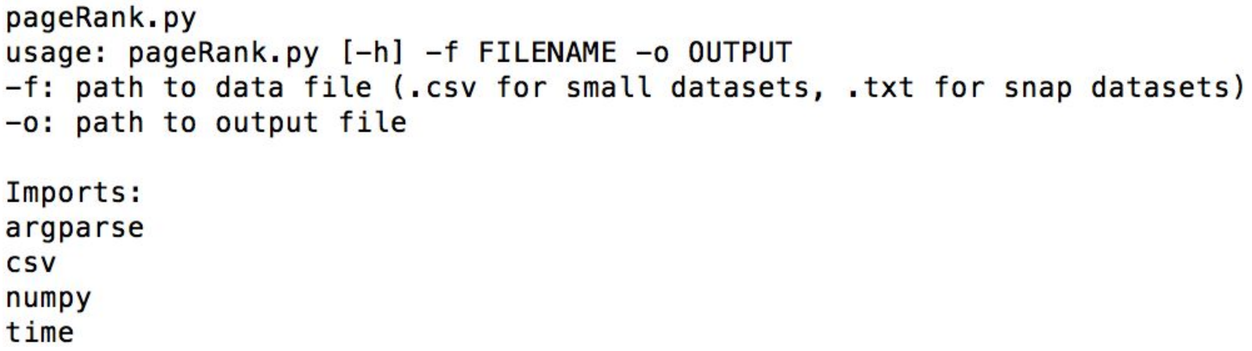




The Log-log plots suggest that the relationships size:time_to_build_graph and size:time_per_iteration scale according to a monomial. We found that the bytes of the datafiles act as a good proxy to the size of the graphs because the algorithms to both construct a graph and run PageRank iteratively scale monomially with the size of the input.

Appendix:

README:



```
pageRank.py
usage: pageRank.py [-h] -f FILENAME -o OUTPUT
-f: path to data file (.csv for small datasets, .txt for snap datasets)
-o: path to output file

Imports:
argparse
csv
numpy
time
```