Lab 3: PageRank and Link Analysis

Nicole Martin and Jeff McGovern
CSC 466, Section 03
October 24, 2015

Abstract

Contents

1 Implementation Overview **1**
1.1 Language . . . 1
1.2 Parsing . . . 2
1.3 Graph Structure . . . 2
1.4 Data Interpretation . . . 2

2 Results **3**
2.1 Small Datasets . . . 3
2.1.1 STATES . . . 3
2.1.2 NCAA-FOOTBALL . . . 5
2.1.3 KARATE . . . 7
2.1.4 DOLPHINS . . . 8
2.1.5 LES-MISERABLES . . . 11
2.1.6 POLITICAL-BLOGS . . . 13
2.2 SPAN Datasets . . . 15
2.2.1 WIKI-VOTE . . . 15
2.2.2 P2P-GNUTELLA05 . . . 16
2.2.3 SLASHDOT-ZOO-NOV6-2008 . . . 17
2.2.4 AMAZON-MAY03 . . . 17
2.2.5 LIVEJOURNAL1 . . . 17

3 Performance Evaluation **17**

4 Conclusion **19**

Appendix A **20**
A.1 `Readme` . . . 20

1 Implementation Overview

1.1 Language

PageRank is implemented in Python 3.5 on an Intel i7 4770k processor. Notable packages include `argparse` for parsing complicated commadn line arguments, `bisect` for sorting lists upon insertion, and `time` for timing.

1.2 Parsing

Parsing the file consists of the following three steps:

1. Open the file
2. Split on `,` for `csv` files or whitespace for the SPAN `txt` files
3. Insert into graph

1.3 Graph Structure

The structure of the graph is a sparse digraph adjacency matrix, with an option to interpret a given graph as undirected edges. In Python, this looks like a dictionary of dictionaries, or a hash table of hash tables. For the smaller datasets, this is fast and memory allocation in manageable. Lookup is $O(1)$ making the already $O(n^2)$ PageRank algorithm not-worse.

For the larger SPAN datasets, this became unwieldy and unmanageable with only 8 GB of memory, and thus an initialized list of dictionaries is used in lieu of an overall dictionary. We hypothesize that this is due to Python's garbage collection mechanism not freeing memory immediately. Instead of hashing on the string label of the node, each node number in the SPAN datasets becomes the index into the array.

1.4 Data Interpretation

Each graph had to be interpreted in such a way that PageRank could be meaningfully applied. Below describes how our PageRank calculator interpreted each dataset.

1.4.0.1 STATES States are nodes and each state connected to another state is represented as two directed edges.

1.4.0.2 NCAA-FOOTBALL Nodes are teams and if a team loses to another team, an edge points from the losing team to the winning team.

1.4.0.3 KARATE Nodes are karate students and two directed edges represent an interaction between two karate students.

1.4.0.4 DOLPHINS Nodes are dolphins and two directed edges represent an interaction between two dolphins.

1.4.0.5 LES-MISERABLES Nodes are characters and two directed edges represent an interaction between two characters.

1.4.0.6 POLITICAL-BLOGS Nodes are political blogs and an edge represents a blog citing another blog.

2 Results

2.1 Small Datasets

2.1.1 STATES

STATES represents the land borders between the 48 contiguous states in the USA and the District of Columbia. The STATES dataset is an undirected graph, as land borders are mutual. The supplied CSV had to be fixed as West Virginia (WV) was mislabeled as MV in some edges and Nebraska (NE) was mislabeled as NB in some places.

2.1.1.1 Settings Connections to bordering states are represented as undirected edges. PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5. All 49 results are shown in order of descending PageRank.

2.1.1.2 Output The parse time was 0.0011491775 seconds. The PageRank computation time was 0.013484954 seconds and it required 13 iterations.

Table 1: PageRanks of States based on their borders.

Filename : stateborders.csv
Damping : 0.500000000
Epsilon : 0.000000100

RESULT	NODE	PageRank
1	TN	0.027495698
2	MA	0.026949944
3	GA	0.025374823
4	MO	0.025259187
5	ID	0.025134611
6	PA	0.024953427
7	NY	0.024343422
8	VA	0.024134314
9	KY	0.024060355
10	SD	0.023307097
11	NH	0.023126294
12	AR	0.022995094
13	MD	0.022881875
14	OK	0.022467746
15	NV	0.022427016
16	IA	0.022351980
17	WY	0.022328398
18	CO	0.022313749
19	NE	0.021470854
20	OH	0.021261319
21	OR	0.021207518
22	UT	0.020736515
23	AL	0.020698044
24	WV	0.020427641
25	NC	0.020290715

Table 1: PageRanks of States based on their borders.

Filename	: stateborders.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
26	IL	0.020145846
27	AZ	0.019800016
28	CT	0.019365074
29	WI	0.019339070
30	MS	0.019246342
31	MN	0.019245921
32	VT	0.019187804
33	TX	0.019163708
34	MT	0.018921125
35	IN	0.018915394
36	NM	0.018806337
37	NJ	0.017636397
38	CA	0.017572727
39	DE	0.017511122
40	KS	0.017303807
41	MI	0.017112018
42	LA	0.016921595
43	ND	0.016917218
44	RI	0.016126596
45	FL	0.015328821
46	SC	0.015277904
47	WA	0.014949573
48	DC	0.014503461
49	ME	0.014058463

2.1.1.3 Observations Since this dataset contains only states that border each other, it has 49 nodes — 1 extra for Washington DC (DC) — and 107 undirected edges in the graph. Two states were not included in the graph since they do not have any land borders with any other states: Alaska (AK) and Hawaii (HI). If those states had been included, they would have a PageRank of (1-d)/N. The state with the fewest edges was Maine (ME), with only 1 state border and therefore one edge in the graph. The two states that had the most edges were Tennessee (TN) and Missouri (MO), both with 8 land borders to other states.

Overall we would expect landlocked states to have a higher PageRank than coastal states since all the borders of landlocked states boarder either another state, Canada, or Mexico. Similarly, we would expect landlocked states that border Canada and Mexico to have lower PageRanks than landlocked states that only border other states.

Tennessee (TN) is the state with the highest PageRank, which is reasonable given that is is a landlocked state with one of the highest numbers of state borders, and that those borders go to other landlocked states, including Missouri (MO) which is the other state with the highest number of state borders. TN and MO also both share borders with Kentucky (KY) and Arkansas (AR), which results in both those states receiving a PageRank boost from bordering the two states with the most borders.

Massachusetts (MA) having the 2nd highest PageRank initially seems odd since it is a coastal state and only borders 5 other states, but that makes sense given the states it borders and their respective PageRanks. New York (NY) has the 7th highest PageRank, and both NY and MA share borders with Vermont (VT) and Connecticut (CT) which both only border 3 states, resulting in NY's high PageRank influencing MA's PageRank even more so.

Georgia (GA) having a higher PageRank than MO initially looks odd given that MO borders 8 states while GA only borders 5. However, since both Florida (FL) and South Carolina (SC) only border 2 other states they contribute more to GA overall prestige score. In contrast the state bordering MO with the fewest borders is Kansas (KS) with 4 borders. Since the states bordering MO all border many other states, they do not contribute as much of their prestige to MO, which resulted in MO having a slightly lower PageRank than GA despite how many states MO borders.

The states with the lowest PageRank values also make sense. ME has the lowest PageRank overall and it is both a coastal state and shares a border with Canada. Considering that ME only borders one other state, and that the state it borders only borders 3 other states in turn, it understandably garners a low PageRank.

2.1.2 NCAA-FOOTBALL

NCAA-FOOTBALL represents every game between Division I teams during the 2009 NCAA football season and the team scores at the end of the game. NCAA-FOOTBALL is a directed graph, where the losing team gives its prestige to the winning team, thereby resulting in a directed edge from the losing team to the winning team.

2.1.2.1 Settings PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5. .

2.1.2.2 Output The parse time was 0.0085580348 seconds. The PageRank computation time was 0.038016080 seconds and it required 13 iterations.

Table 2: Top 25 PageRanks of NCAA Football teams based on their wins against other teams.

Filename : NCAA_football.csv
Damping : 0.500000000
Epsilon : 0.000000100

RESULT	NODE	PageRank
1	Montana	0.007595809
2	Tulsa	0.007193120
3	Rice	0.006869487
4	Boise State	0.006833071
5	Ball State	0.006423026
6	Utah	0.006268279
7	TCU	0.006039433
8	Weber State	0.005943094
9	Richmond	0.005896953
10	New Hampshire	0.005861479
11	Alabama	0.005849003

Table 2: Top 25 PageRanks of NCAA Football teams based on their wins against other teams.

Filename	: NCAA_football.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
12	Texas Tech	0.005775430
13	Florida A&M	0.005665455
14	Troy	0.005645406
15	Appalachian State	0.005625847
16	South Carolina State	0.005609834
17	Maine	0.005461631
18	Texas	0.005460256
19	Penn State	0.005432817
20	Oklahoma	0.005427830
21	Bethune-Cookman	0.005337554
22	Houston	0.005316895
23	Florida	0.005306193
24	Cincinnati	0.005253648
25	Western Michigan	0.005170865

Table 3: Top 10 PageRanks of NCAA Football teams based on their wins against other teams, with a smaller damper constant.

Filename	: NCAA_football.csv	
Damping	: 0.100000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
1	Tulsa	0.003999626
2	Montana	0.003975512
3	Rice	0.003946057
4	Ball State	0.003835217
5	Boise State	0.003811216
6	New Hampshire	0.003753150
7	TCU	0.003722277
8	Maine	0.003718225
9	Weber State	0.003692130
10	Troy	0.003664002

The PageRank for teams that never won a game was 0.001543210 with a damping of 0.5 and epsilon of 0.000000100.

2.1.2.3 Observations This dataset has 324 nodes corresponding to the different NCAA Division 1 football teams in the 2009 season and 1,537 edges in its graph corresponding to the 1,537 NCAA Division I football games played in 2009 before the bowls and championships started.

Analysis of these rankings is difficult because NCAA teams play teams within their own conferences for the most part, with non-conference games that are pre-negotiated by the teams years in advance. As a result, the structure of this graph would contain many node clusters with few linkages between them.

PageRank was run with an epsilon of 0.000000100 and 2 different damping values: 0.5 and 0.1. In both cases the top 5 results were Tulsa, Montana, Rice, Ball State, and Boise State, though in different order.

With both settings Rice had the #3 PageRank. This makes sense given that Rice had a good winning ration at 10-3, but with its loss to Tulsa it gave a lot of prestige to Tulsa, which boosted that school's rankings.

Tulsa had an 11-3 record, which included wins against Rice and Ball State. With a damper of 0.5 it had the 2nd highest PageRank, and with a damper of 0.1 it had the highest PageRank. This makes sense since it won against #3 and #5 schools when the damper was set to 0.5, and won against the #3 and #4 schools with the damper set to 0.1. In both cases, Tulsa had a significant boost in its PageRank by winning against similarly highly ranked schools.

Ball State was ranked #5 and #4 in the two scenarios had a 12-2 record with a loss to Tulsa. In contrast Neither Montana nor Boise state played the other top 5 teams. Montana had a 14-2 record and Boise State had a 12-1 record. Both teams benefited from having high win ratios though the teams they won against did not have similarly high prestige.

Unsurprisingly the teams that never won a game were tied for the lowest PageRank, as they never gained any prestige from winning a game.

2.1.3 KARATE

KARATE represents a small social network of members of a university karate club. This dataset is an undirected graph, as all friendships in this social network are mutual.

2.1.3.1 Settings PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5.

2.1.3.2 Output The parse time was 0.0008490085 seconds. The PageRank computation time was 0.015735387 seconds and it required 15 iterations.

Table 4: PageRanks of Karate students based on their interaction with other Karate students.

Filename : karate.csv
Damping : 0.500000000
Epsilon : 0.000000100

RESULT	NODE	PageRank
1	34	0.079973839
2	1	0.076404061
3	33	0.058828626
4	3	0.045020395
5	2	0.044320974
6	32	0.033867544
7	4	0.032508735
8	7	0.030715787

Table 4: PageRanks of Karate students based on their interaction with other Karate students.

Filename	: karate.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
9	6	0.030715787
10	24	0.030404437
11	30	0.027679177
12	9	0.027234766
13	14	0.026868036
14	28	0.026510095
15	11	0.025119580
16	5	0.025119580
17	25	0.024963474
18	26	0.024729201
19	31	0.024694997
20	8	0.024515865
21	17	0.022384830
22	29	0.022131369
23	20	0.021907955
24	27	0.020517950
25	13	0.019802570
26	22	0.019555784
27	18	0.019555784
28	23	0.019509244
29	19	0.019509244
30	15	0.019509244
31	16	0.019509244
32	21	0.019509244
33	10	0.019309072
34	12	0.017093509

2.1.3.3 Observations This was a very small graph, with only 34 nodes and 78 edges representing members of the karate club and mutual friendships respectively. The most social people in the network were #34 with 17 friendships and #1 with 16 friendships. The least social person in the club was #12 with only 1 friend.

The nodes with the highest and second-highest PageRanks were #34 and #1 respectively, which corresponds to the club members with the most friendships and is consistent with our expectations of the PageRank algorithm. The person with the fewest friends, #12, also unsurprisingly had the lowest PageRank.

2.1.4 DOLPHINS

DOLPHINS a small social network of dolphins observed by researchers. It is an undirected graph of mutual dolphin friendships.

2.1.4.1 Settings

PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5.

2.1.4.2 Output

The parse time was 0.0030930042 seconds. The PageRank computation time was 0.020778417 seconds and it required 12 iterations.

Table 5: PageRanks of observed social interactions between dolphins

Filename	: dolphins.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
1	Trigger	0.029643421
2	Jet	0.028729336
3	Web	0.026117101
4	Grin	0.024610818
5	Scabs	0.024123720
6	Patchback	0.023660385
7	SN4	0.023149422
8	Topless	0.023047488
9	SN63	0.022238950
10	Gallatin	0.021455867
11	Beescratch	0.020952327
12	Kringel	0.020436276
13	Stripes	0.020193012
14	Feather	0.020047642
15	SN100	0.018938709
16	SN9	0.018504134
17	Upbang	0.018441378
18	Haecksel	0.018314271
19	DN21	0.018067878
20	Number1	0.017010041
21	SN96	0.016938773
22	Jonah	0.016921184
23	TR99	0.016627969
24	TR77	0.016514054
25	Shmuddel	0.016332147
26	Double	0.016285683
27	Ripplefluke	0.016221531
28	Beak	0.015880190
29	PL	0.015610846
30	DN63	0.015525973
31	MN83	0.015451038
32	Fish	0.015378191
33	MN105	0.015375856
34	DN16	0.015154557
35	Zap	0.015123440

Table 5: PageRanks of observed social interactions between dolphins

Filename	: dolphins.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
36	Bumper	0.015105026
37	Hook	0.015061392
38	SN90	0.014915194
39	Oscar	0.014803775
40	Thumper	0.014111140
41	Zipfel	0.013963419
42	Knit	0.013804958
43	Mus	0.013584798
44	Notch	0.013339171
45	TSN103	0.012950873
46	TR88	0.012879415
47	TR120	0.012726728
48	CCL	0.011959448
49	MN60	0.011947070
50	TSN83	0.011834109
51	Wave	0.011464499
52	SN89	0.010868234
53	Vau	0.010854854
54	Zig	0.010768115
55	Quasi	0.009660595
56	MN23	0.009660595
57	Five	0.009546690
58	Cross	0.009546690
59	TR82	0.009515468
60	Whitetip	0.009454450
61	SMN5	0.009378982
62	Fork	0.009270703

2.1.4.3 Observations This graph was small, with 62 nodes and 159 undirected edges corresponding the 62 dolphins and the 159 mutual friendships between them. Nine of the dolphins in the graph only had one friend: Cross, Five, Fork, MN23, Quasi, SMN5, TR82, Whitetip, and Zig. The most social dolphin was Grin with 12 friends, followed by SN4 and Topless with 11 friends each, and Scabs and Trigger with 10 friends each.

With this in mind, the fact that the 3 dolphins with the most friends (Grin, SN4, and Topless) did not have the highest PageRanks initially seems odd with respective ranks of #4, #7, #8. The dolphin with the highest PageRank is Trigger, who has 10 edges. Only 2 of Triggers friends are in the top 10, Patchback and Topless with 9 and 11 friends respectively. Jet similarly had 9 friends with only 2 (Web and Gallatin) in the top 10. Overall this shows that highly connected nodes that distribute their privilege a lot can have a lower than expected PageRank.

Unsurprisingly the nine dolphins with only 1 friend each had the lowest PageRanks.

2.1.5 LES-MISERABLES

LES-MISERABLES represents every instance of characters appearing in the same chapter as each other in the novel *Les Misérables* by Victor Hugo. This is represented by an undirected, edge-labeled graph.

2.1.5.1 Settings PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5.

2.1.5.2 Output The parse time was 0.0026361942 seconds. The PageRank computation time was 0.022474527 seconds and it required 17 iterations.

Table 6: PageRanks of interactions between character in the book *Les Misérables*

Filename : lesmis.csv
Damping : 0.500000000
Epsilon : 0.000000100

RESULT	NODE	PageRank
1	Valjean	0.060053858
2	Myriel	0.040976294
3	Gavroche	0.026658538
4	Javert	0.023162103
5	Marius	0.022975407
6	Thenardier	0.022442951
7	Fantine	0.021956282
8	Cosette	0.017384434
9	MlleGillenormand	0.017319376
10	MmeThenardier	0.016465736
11	Enjolras	0.016137772
12	Mabeuf	0.016007582
13	Gillenormand	0.015371424
14	Eponine	0.014826126
15	Bossuet	0.014749849
16	Courfeyrac	0.014595614
17	Fauchelevent	0.014441353
18	Tholomyes	0.013953934
19	Bahorel	0.013903747
20	Joly	0.013903747
21	Bamatabois	0.013861927
22	Gueulemer	0.013610888
23	Babet	0.013610888
24	Claquesous	0.013546174
25	Combeferre	0.013132505
26	Feuilly	0.013132505
27	Montparnasse	0.012830201
28	Grantaire	0.012524606
29	Blacheville	0.012445379
30	Zephine	0.012445379

Table 6: PageRanks of interactions between character in the book *Les Misérables*

Filename	: lesmis.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
31	Listolier	0.012445379
32	Dahlia	0.012445379
33	Fameuil	0.012445379
34	Favourite	0.012445379
35	Chenildieu	0.012290938
36	Cochepaille	0.012290938
37	Judge	0.012290938
38	Brevet	0.012290938
39	Champmathieu	0.012290938
40	MmeBurgon	0.011824155
41	Prouvaire	0.011744719
42	MmeMagloire	0.011251681
43	MlleBaptistine	0.011251681
44	Brujon	0.011225826
45	Simplice	0.010887275
46	MmeHucheloup	0.010550855
47	LtGillenormand	0.010223381
48	Pontmercy	0.010155251
49	Child2	0.009465843
50	Child1	0.009465843
51	Jondrette	0.009449545
52	MmePontmercy	0.009423147
53	MotherInnocent	0.009132757
54	Woman2	0.008799028
55	Toussaint	0.008799028
56	Anzelma	0.008617206
57	Perpetue	0.008586292
58	Champtercier	0.008542318
59	Geborand	0.008542318
60	Count	0.008542318
61	CountessDeLo	0.008542318
62	Napoleon	0.008542318
63	OldMan	0.008542318
64	Cravatte	0.008542318
65	Magnon	0.008339908
66	Gribier	0.008298676
67	BaronessT	0.008196081
68	Marguerite	0.008059464
69	Woman1	0.008008826
70	MlleVaubois	0.007730605
71	Labarre	0.007327588

Table 6: PageRanks of interactions between character in the book *Les Misérables*

Filename	: lesmis.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
72	Gervais	0.007327588
73	MmeDeR	0.007327588
74	Isabeau	0.007327588
75	Scaufflaire	0.007327588
76	MotherPlutarch	0.007221124
77	Boulatruelle	0.007194849

2.1.5.3 Observations *Les Misérables* has 60 named characters and it is completely unsurprising that the protagonist Jean Valjean has the highest PageRank. Myriel is a key character to Valjean's character growth and arc, so it is unsurprising as well that Myriel has a high PageRank since any chapter with Valjean in it is likely to include mentions of Myriel. Gavroche also plays a key role in the story as a messanger between characters. Javert spends most of the novel chasing Valjean, so again his high PageRank is unsurprising.

Other key characters, such as Marius, Thenardier and his wife MmeThenardier, Marius, Fantine, and Cosette all appear in the top 10, which makes sense as they are key characters to the story.

2.1.6 POLITICAL-BLOGS

POLITICAL-BLOGS represents a set of political blogs that cited each other on the evening of the 2004 United States Presidential Election. This is a directed graph, as one blog citing another blog does not imply that the other blog cited the original blog.

2.1.6.1 Settings PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5.

2.1.6.2 Output The parse time was 0.0959103107 seconds. The PageRank computation time was 0.304592370 seconds and it required 14 iterations.

Table 7: PageRanks of political blogs that cited each other during the 2004 Presidential Election

Filename	: polblogs.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
1	155	0.024316675
2	963	0.018442731
3	641	0.017737135
4	1051	0.017452996
5	855	0.016882272
6	55	0.016248193

Table 7: PageRanks of political blogs that cited each other during the 2004 Presidential Election

Filename	: polblogs.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
7	1245	0.015097998
8	1437	0.013825288
9	1153	0.013316673
10	729	0.012629643
11	1112	0.011602062
12	1041	0.010285518
13	798	0.010116229
14	323	0.009695603
15	1479	0.009498175
16	1179	0.008916739
17	434	0.008882113
18	996	0.008063774
19	878	0.007958224
20	1000	0.007898844
21	483	0.007557769
22	642	0.007538489
23	741	0.007509966
24	756	0.007459060
25	1270	0.007275318
26	1306	0.007126369
27	1330	0.007085794
28	180	0.007043830
29	493	0.007014826
30	1461	0.006765612
31	826	0.006734074
32	170	0.006709852
33	1463	0.006628301
34	919	0.006428267
35	514	0.006385808
36	297	0.006260373
37	1232	0.005990919
38	687	0.005958751
39	189	0.005924997
40	990	0.005879998
41	1101	0.005864516
42	535	0.005662016
43	210	0.005612739
44	301	0.005608227
45	150	0.005557346
46	1045	0.005506802
47	1055	0.005477806

Table 7: PageRanks of political blogs that cited each other during the 2004 Presidential Election

Filename	: polblogs.csv	
Damping	: 0.500000000	
Epsilon	: 0.000000100	
RESULT	NODE	PageRank
48	979	0.005465431
49	363	0.005463737
⋮	⋮	⋮
1220	20	0.000408497
1221	80	0.000408497
1222	502	0.000408497
1223	1259	0.000408497
1224	691	0.000408497

2.1.6.3 Observations There were 1,224 nodes and 19,090 edges in the POLITICAL-BLOGS graph. The highest ranked node, 155, was cited by 338 blogs, but of the top 10 by PageRank it was only cited by nodes 55, 1153 and 729. The second highest node, 963, was cited by 240 blogs and was cited by nodes 1051 and 855. In contrast the nodes with the smallest PageRank were not cited by any other nodes, resulting in their minimal PageRank value.

2.2 SPAN Datasets

2.2.1 WIKI-VOTE

WIKI-VOTE represents a collection of internal voting results for Wikipedia administration-related elections. This is a directed graph, where each outgoing edge represents a vote for the user it is directed toward.

2.2.1.1 Settings PageRank was run on this dataset with epsilon at a default value of 0.0000001, the damper constant at a default value of 0.5.

2.2.1.2 Output The parse time was 0.3580873012 seconds. The PageRank computation time was 44.538764715 seconds and it required 4 iterations.

Table 8: PageRanks of internal voting results for Wikipedia administration

Filename	:	wiki-Vote.txt
Damping	:	0.500000000
Epsilon	:	0.000000100
RESULT	NODE	PageRank
1	3	0.000139558
2	4	0.000139558
3	5	0.000139558
4	6	0.000139558

Table 8: PageRanks of internal voting results for Wikipedia administration

Filename	:	wiki-Vote.txt
Damping	:	0.500000000
Epsilon	:	0.000000100
RESULT	NODE	PageRank
5	7	0.000139558
6	8	0.000139558
7	9	0.000139558
8	10	0.000139558
9	11	0.000139558
10	12	0.000139558
921	1000	0.000139558
1000	1087	0.000139558
1843	2000	0.000139558
2000	2177	0.000139558
2747	3000	0.000139558
3000	3274	0.000139558
3622	4000	0.000139558
4000	4446	0.000139558
4470	5001	0.000139558
5001	5640	0.000139558
5297	6000	0.000139558
6000	6870	0.000139558
6102	7000	0.000139558
7000	8148	0.000139558
7112	8294	0.000070274
7113	8295	0.000070274
7114	8296	0.000070274
7115	8297	0.000070274

2.2.1.3 Observations In this incredibly large dataset, it appears to have a lot of very closely related PageRank values. Table 8 contains select data points from the calculated PageRank for the WIKI-VOTE dataset. While the table displays a damping value of 0.500, we experimented with a lower value and found that it finished far faster – about 11 seconds instead of 44. However, the results were even less useful, with nearly every value being identical. When applying a dampening value of 0.95, the runtime increased to 465 seconds and 41 iterations, about a 10x increase in time and iterations. These lead me to believe that there is a bug in the code, or that data is lost in the computation.

2.2.2 P2P-GNUTELLA05

P2P-GNUTELLA05 represents all the peer-to-peer connections that occurred on the peer-to-peer file sharing network Gnutella on August 5, 2002. This is a directed graph in which an outgoing edge indicates that the user requested data from the other user/server.

2.2.2.1 Settings
With $d = 0.50$ and $\epsilon = 0.0000001$, this dataset did not complete the PageRank computation.

2.2.3 SLASHDOT-ZOO-NOV6-2008

SLASHDOT-ZOO-NOV6-2008 represents all the friend/foe links between user accounts on the website Slashdot as of November 2008. This is a directed graph, where friend links are links from that user to the other. At this time, the graph does not account for foe links.

2.2.3.1 Settings
When running this computation, we used $d = 0.50$ and $\epsilon = 0.0000001$, this dataset did not complete the PageRank computation.

2.2.4 AMAZON-MAY03

AMAZON-MAY03 represents all products that were purchased together on Amazon.com during May 2003. This is represented as a directed graph in which an edge points from product A to product B if people who purchased product A also tended to purchase product B. The interpretation of this graph remains directed, since if someone who buys product A tends to buy product B, the converse isn't necessarily true.

2.2.4.1 Settings
With $d = 0.50$ and $\epsilon = 0.0000001$, this dataset did not complete the PageRank computation.

2.2.5 LIVEJOURNAL1

LIVEJOURNAL1 shows all following relationships between user accounts on Livejournal.com at a specific, unspecified time. This is a directed graph from follower to the followed blog.

2.2.5.1 Settings
With $d = 0.50$ and $\epsilon = 0.0000001$, this dataset did not complete the PageRank computation.

3 Performance Evaluation

3.0.5.2 Memory Usage
Memory usage was the most limiting factor of this project. An initial implementation using dictionaries was impossible to run with the SPAN datasets. Running everything was mostly fine, and times scaled intuitively, but parsing the LiveJournal dataset was impossible on only 8GB of ram. Only after reimplementing the graph to avoid a lagging garbage collector were we able to parse the largest dataset. Nevertheless, nothing larger than the SLASHDOT dataset finished in any reasonable amount of time (3 hours).

3.0.5.3 Timing
Table 9 contains the breakdown of each dataset tested. For the non-SPAN datasets, parse, and PageRank calculation times are reasonable, taking a fraction of a second. Of course, the larger SPAN datasets take much longer. For example, the Slashdot dataset took approximately 45 minutes to compute the PageRank with 1 iteration.

Table 9: Performance Overview of All Datasets

Filename	d	e	$\|N\|$	$\|E\|$	ParseTime(sec)	PageRankTime(s)	Iters
stateborders.csv	0.5	1e-07	49	214	0.0011491775	0.013484954	13
NCAA_football.csv	0.5	1e-07	324	1537	0.0085580348	0.038016080	13
karate.csv	0.5	1e-07	34	156	0.0008490085	0.015735387	15
dolphins.csv	0.5	1e-07	62	318	0.0030930042	0.020778417	12
lesmis.csv	0.5	1e-07	77	508	0.0026361942	0.022474527	17
polblogs.csv	0.5	1e-07	1224	19090	0.0959103107	0.304592370	14
chicken.csv	0.5	1e-07	8	13	0.0001974105	0.020763397	21
wiki-Vote.txt	0.5	1e-07	7115	103689	0.3580873012	44.538764715	4
soc-sign-Slashdot081106.txt	0.5	1e-07	70491	396378	1.77220392	2836.84788	1
wiki-Vote.txt	0.5	1e-07	7115	103689	0.328359365	None	None
p2p-Gnutella05.txt	0.5	1e-07	8846	31839	0.103373050	None	None
soc-LiveJournal1.txt	non	1e-07	4847571	68993773	239.2832429	None	None

Table 10: Performance Overview of All Datasets, Ordered By PageRank Calculation Time

Filename	d	e	$\|N\|$	$\|E\|$	ParseTime(sec)	PageRankTime(s)	Iters
stateborders.csv	0.5	1e-07	49	214	0.0011491775	0.013484954	13
karate.csv	0.5	1e-07	34	156	0.0008490085	0.015735387	15
chicken.csv	0.5	1e-07	8	13	0.0001974105	0.020763397	21
dolphins.csv	0.5	1e-07	62	318	0.0030930042	0.020778417	12
lesmis.csv	0.5	1e-07	77	508	0.0026361942	0.022474527	17
NCAA_football.csv	0.5	1e-07	324	1537	0.0085580348	0.038016080	13
polblogs.csv	0.5	1e-07	1224	19090	0.0959103107	0.304592370	14
wiki-Vote.txt	0.5	1e-07	7115	103689	0.3580873012	44.538764715	4
soc-sign-Slashdot081106.txt	0.5	1e-07	70491	396378	1.77220392	2836.84788	1
wiki-Vote.txt	0.5	1e-07	7115	103689	0.328359365	None	None
p2p-Gnutella05.txt	0.5	1e-07	8846	31839	0.103373050	None	None
soc-LiveJournal1.txt	non	1e-07	4847571	68993773	239.2832429	None	None

Table 11: Performance Overview of All Datasets, Ordered By Parse Time

Filename	d	e	$\|N\|$	$\|E\|$	ParseTime(sec)	PageRankTime(s)	Iters
karate.csv	0.5	1e-07	34	156	0.0008490085	0.015735387	15
chicken.csv	0.5	1e-07	8	13	0.0001974105	0.020763397	21
stateborders.csv	0.5	1e-07	49	214	0.0011491775	0.013484954	13
lesmis.csv	0.5	1e-07	77	508	0.0026361942	0.022474527	17
dolphins.csv	0.5	1e-07	62	318	0.0030930042	0.020778417	12
NCAA_football.csv	0.5	1e-07	324	1537	0.0085580348	0.038016080	13
polblogs.csv	0.5	1e-07	1224	19090	0.0959103107	0.304592370	14
wiki-Vote.txt	0.5	1e-07	7115	103689	0.3580873012	44.538764715	4
soc-sign-Slashdot081106.txt	0.5	1e-07	70491	396378	1.77220392	2836.84788	1
wiki-Vote.txt	0.5	1e-07	7115	103689	0.328359365	None	None
p2p-Gnutella05.txt	0.5	1e-07	8846	31839	0.103373050	None	None
soc-LiveJournal1.txt	non	1e-07	4847571	68993773	239.2832429	None	None

4 Conclusion

In general PageRank results for undirected graphs are not particularly interesting due to their undirected nature. Since situations such as an outgoing edge with no incoming edge cannot happen, overall the situations reduce to having the nodes with the most number of edges having the highest PageRank and those with the least edges having the lowest PageRanks.

With directed graphs the PageRank results become less naively predictable due to the sharing of prestige. Nodes with no incoming edges predictably have the lowest PageRank. However, having outgoing edges to many nodes means that each of those nodes receives less prestige from the original, which makes the nodes that effectively act as hubs contribute less prestige to each node they link to. The total number of incoming edges no longer directly predicts the PageRank.

Appendix A

A.1 Readme

```
      :'######::'########::'########::::'##:::::::::'#######:::'#######::
      '##... ##: ##.... ##: ##.....::::: ##:::'##::'##.... ##:'##.... ##:
       ##:::..:: ##:::: ##: ##:::::::::: ##::: ##:: ##::::..:: ##::::..::
       ##::::::: ########:: ######:::::: ##::: ##:: ########:: ########::
       ##::::::: ##.....::: ##...::::::: #########: ##.... ##: ##.... ##:
       ##::: ##: ##:::::::: ##::::::::::...... ##:: ##:::: ##: ##:::: ##:
      . ######:: ##:::::::: ########:::::::::: ##::. #######::. #######::
      :......:::..:::::::::........:::::::::::..::::.......::::.......:::
                 '##::::::::::'###::::'########::::::'#######::
                  ##:::::::::'## ##::: ##.... ##::::'##.... ##:
                  ##::::::::'##:. ##:: ##:::: ##::::..::::: ##:
                  ##:::::::'##:::. ##: ########::::::'#######::
                  ##::::::: #########: ##.... ##:::::...... ##:
                  ##::::::: ##.... ##: ##:::: ##::::'##:::: ##:
                  ########: ##:::: ##: ########:::::. #######::
                 ........::..:::::..::........:::::::.......:::
./run.py -f filename
```