Wikipedia Graphing

Jonathan Chiu and Brian Goodacre

Background

- Wikipedia has nearly 4 million English articles
- All articles have links to other articles
- Goal: Play with these links as a graph

Implementation Details

- Got data already parsed
  - One text file with names
  - One text file with pageID and outgoing links
  - Over 5 million pages
- Divided large data file (1GB) into many smaller text files
- Used binary search then sequential search to get outgoing links

Example of Data

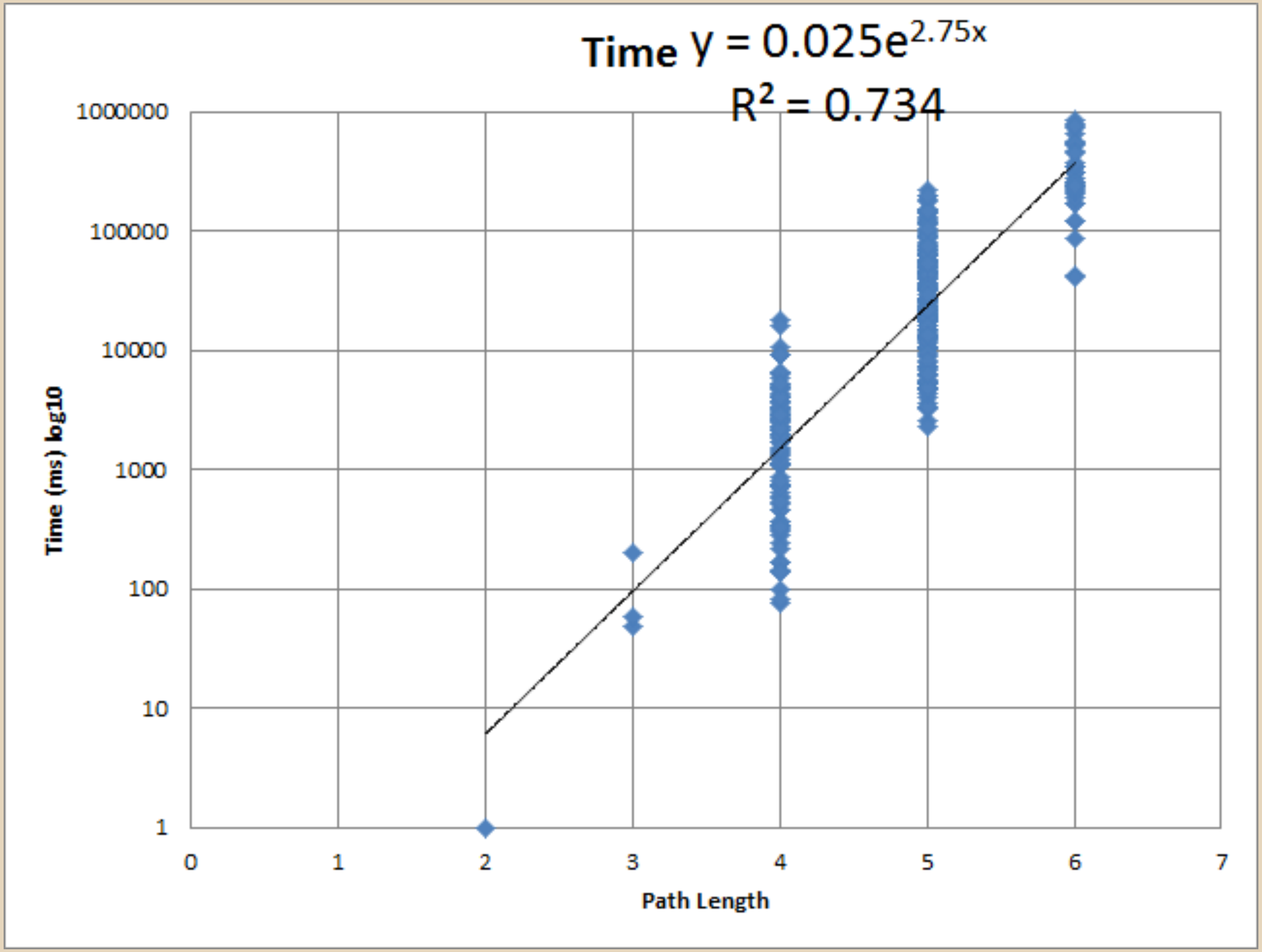
- Ambracians
- Ambrai_Smalltalk
- Ambrakia
- Ambrakikos_Gulf
- Ambras_Castle
- Ambras_syndrome
- Ambrault
- Ambrazevicius
- Ambre_(band)
- Ambre_Anderson

- 5703678: 5703424
- 5703679: 5703424
- 5703680: 318949 1373708 2158188 2239819 2264388 2328376 2398807 2876433 2876459 2978801 3105487 3827312 4095633 4306159 5368424 5697530 5703692 5703675 5703685 5703645 5703654 5703661 5703708 5709232 5709336 5709338 5709350 5710016 5710020 5710021
- 5703681: 230937 318949 1373708 1603276 1985355 2041550 2239819 2328376 2398807 2411247 2671526 2876433 2876459 2978801 3105640 3420555 3827312 4095633 4306159 5491829 5703795

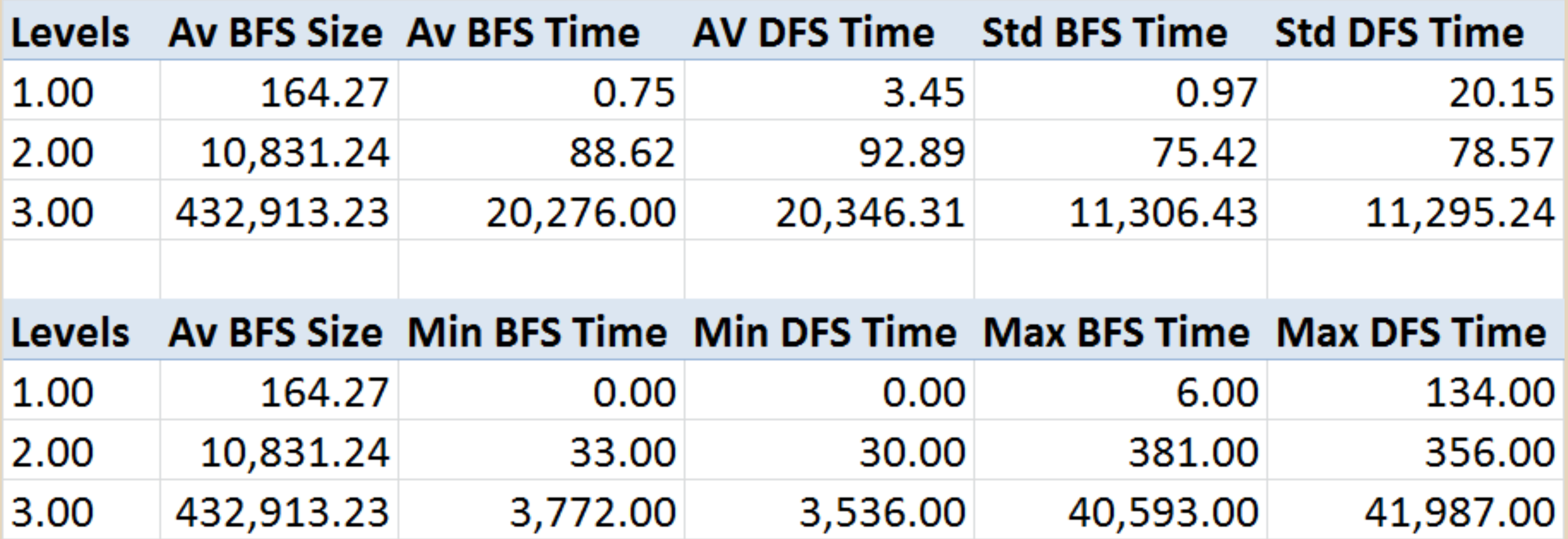
BFS: Page A to Page B

Path Length	Count	Average Time (ms)	Stdev Time (ms)	Min Time (ms)	Max Time (ms)
2	1	1	0	1	1
3	3	103	87	48	203
4	97	2,747	3151	75	17923
5	181	42,548	43,853	2,319	221,251
6	37	343,608	210,075	40,437	842,418

BFS: Page A to Page B



BFS and DFS: N Levels

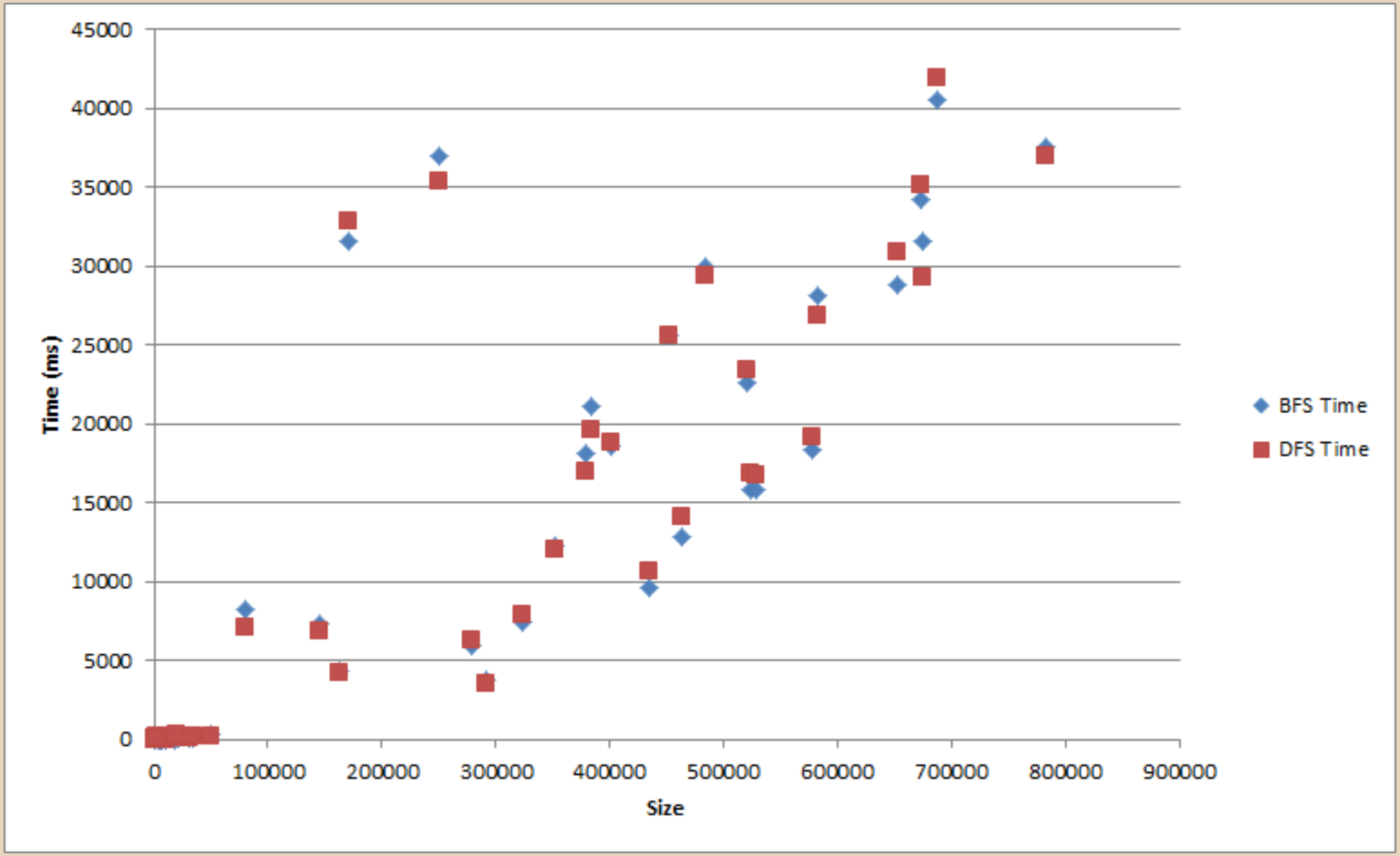


Levels	Av BFS Size	Av BFS Time	AV DFS Time	Std BFS Time	Std DFS Time
1.00	164.27	0.75	3.45	0.97	20.15
2.00	10,831.24	88.62	92.89	75.42	78.57
3.00	432,913.23	20,276.00	20,346.31	11,306.43	11,295.24

Levels	Av BFS Size	Min BFS Time	Min DFS Time	Max BFS Time	Max DFS Time
1.00	164.27	0.00	0.00	6.00	134.00
2.00	10,831.24	33.00	30.00	381.00	356.00
3.00	432,913.23	3,772.00	3,536.00	40,593.00	41,987.00

*Times are in milliseconds

BFS and DFS: N Levels



Difficulties Faced

- Data set is enormous
  - lack of main memory
- Old data set
- Many pages with no incoming links
  - redirects
- Cannot tell difference between outgoing links
  - 1st link, GPS, years

Demo

- Please give us a starting page title and an ending page title
- Live Results: http://j.mp/I9FZPK