EPL646 – Advanced Topics in Databases
Advanced Spark

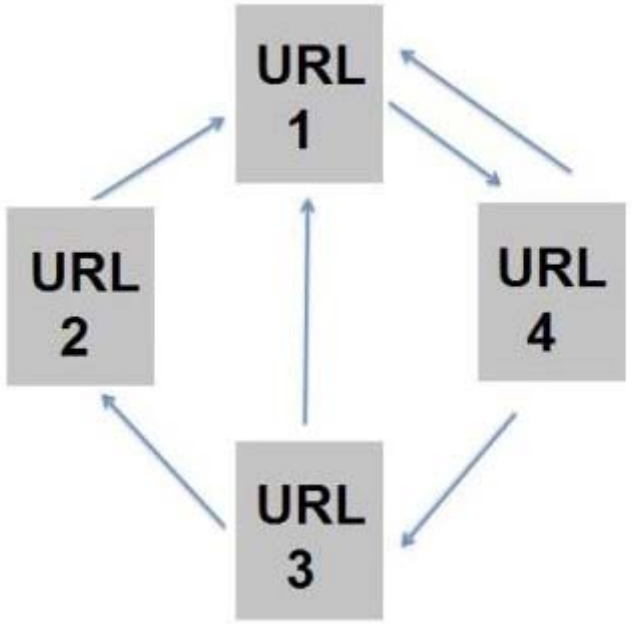
http://www.cs.ucy.ac.cy/~dzeina/courses/epl646/labs/lab.html

Task1: PageRank

- Basic Idea:

Give pages ranks (scores) based on links to them

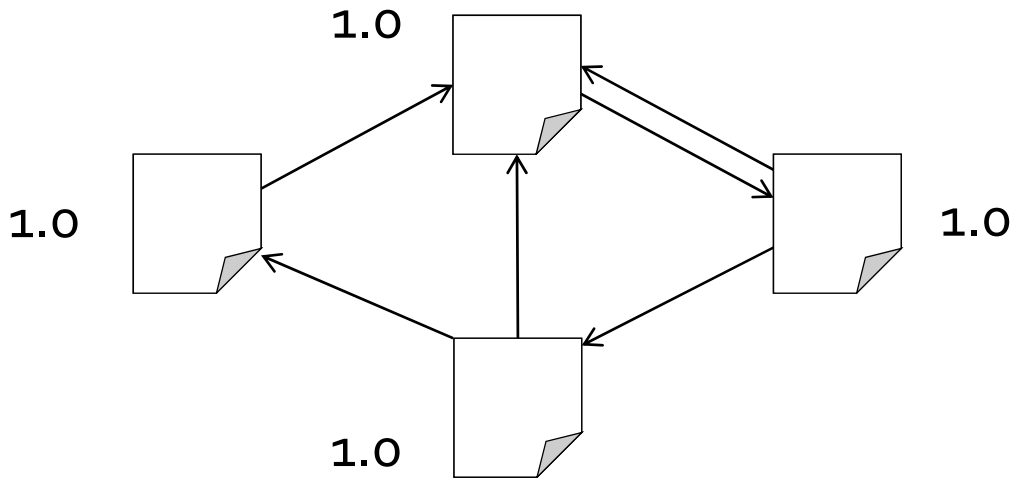
- Links from many pages ➔ high rank
- Link from a high-rank page ➔ high rank



Task1: PageRank

- Algorithm:

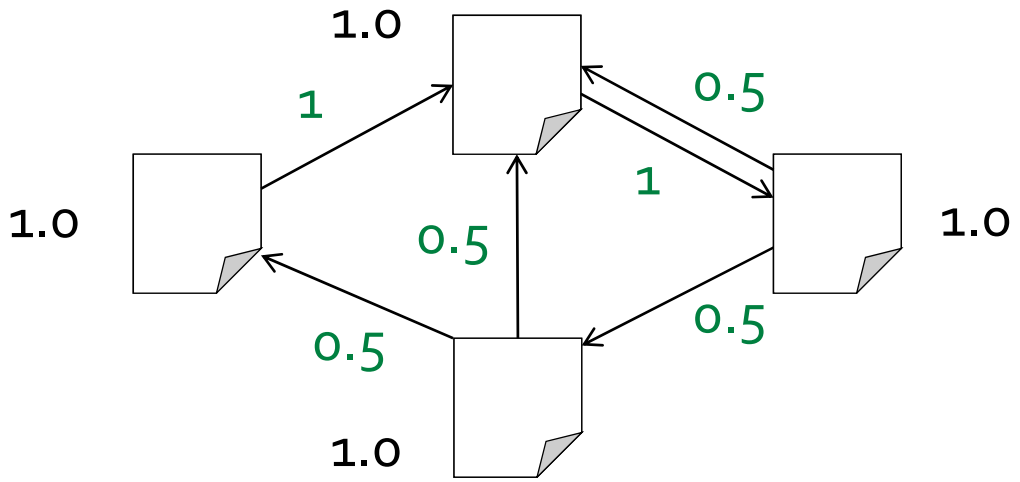
1. Start each page at a rank of 1
2. On each iteration, have page p contribute $rank_p$ / $|neighbors_p|$ to its neighbors
3. Set each page's rank to 0.15 + 0.85 × contribs



Task1: PageRank

- Algorithm:

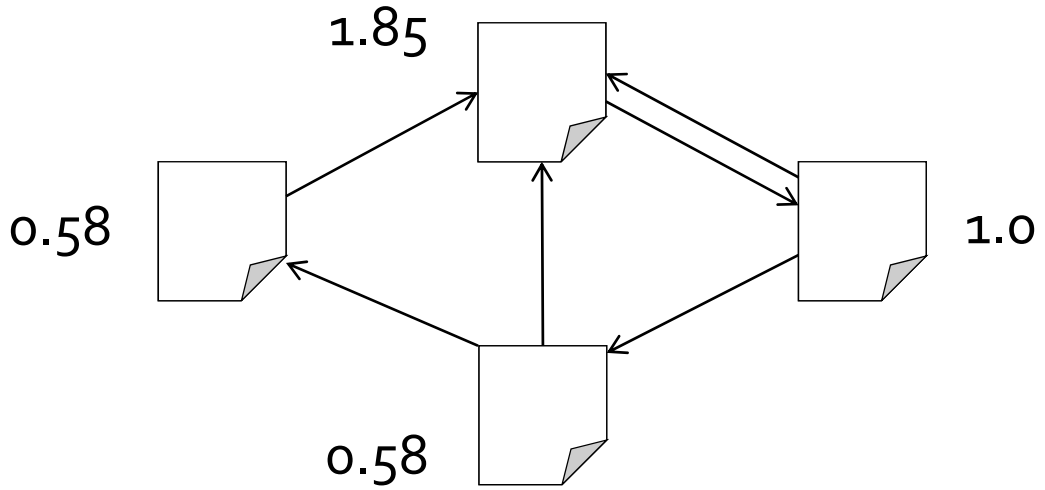
1. Start each page at a rank of 1
2. On each iteration, have page p contribute $rank_p$ / $|neighbors_p|$ to its neighbors
3. Set each page's rank to 0.15 + 0.85 × contribs



Task1: PageRank

- Algorithm:

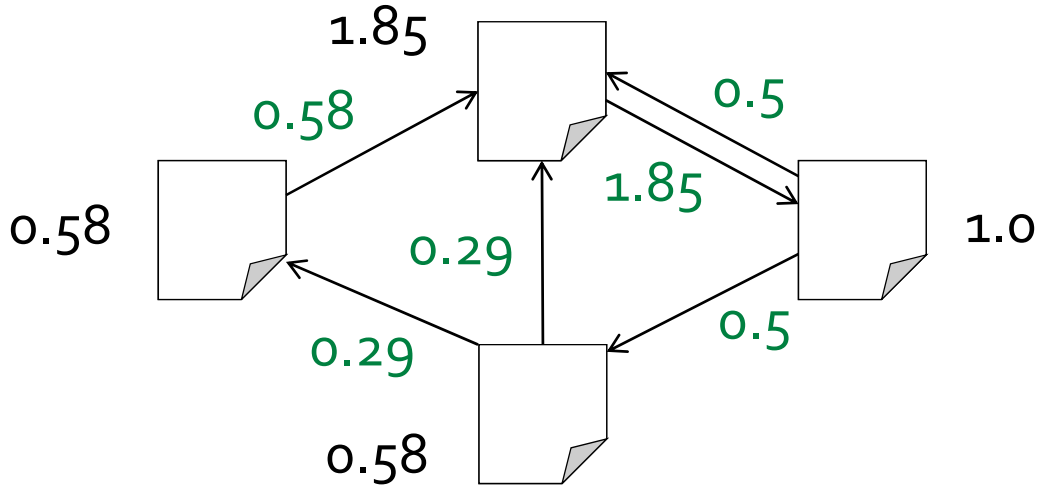
1. Start each page at a rank of 1
2. On each iteration, have page p contribute $rank_p$ / $|neighbors_p|$ to its neighbors
3. Set each page's rank to 0.15 + 0.85 × contribs



Task1: PageRank

- Algorithm:

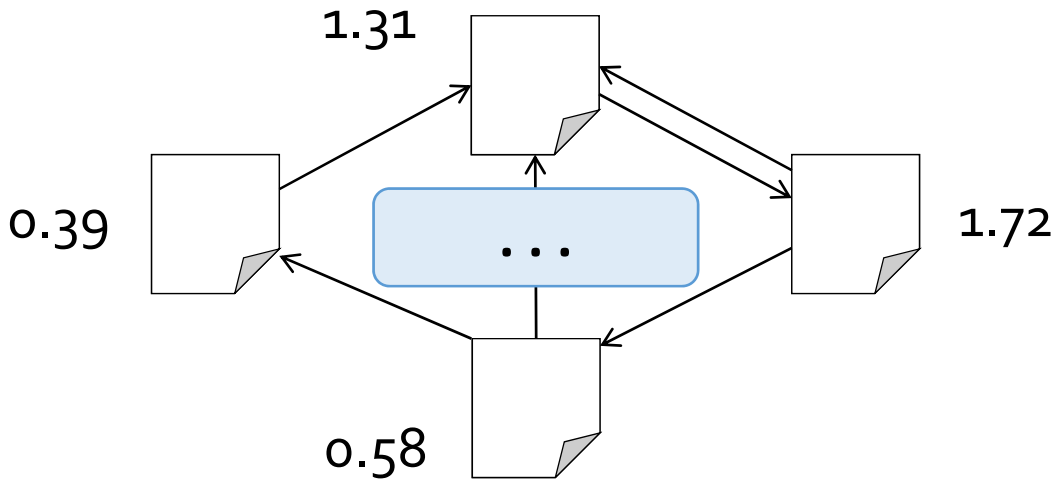
1. Start each page at a rank of 1
2. On each iteration, have page p contribute $rank_p$ / $|neighbors_p|$ to its neighbors
3. Set each page's rank to 0.15 + 0.85 × contribs



Task1: PageRank

- Algorithm:

1. Start each page at a rank of 1
2. On each iteration, have page p contribute $rank_p$ / $|neighbors_p|$ to its neighbors
3. Set each page's rank to 0.15 + 0.85 × contribs

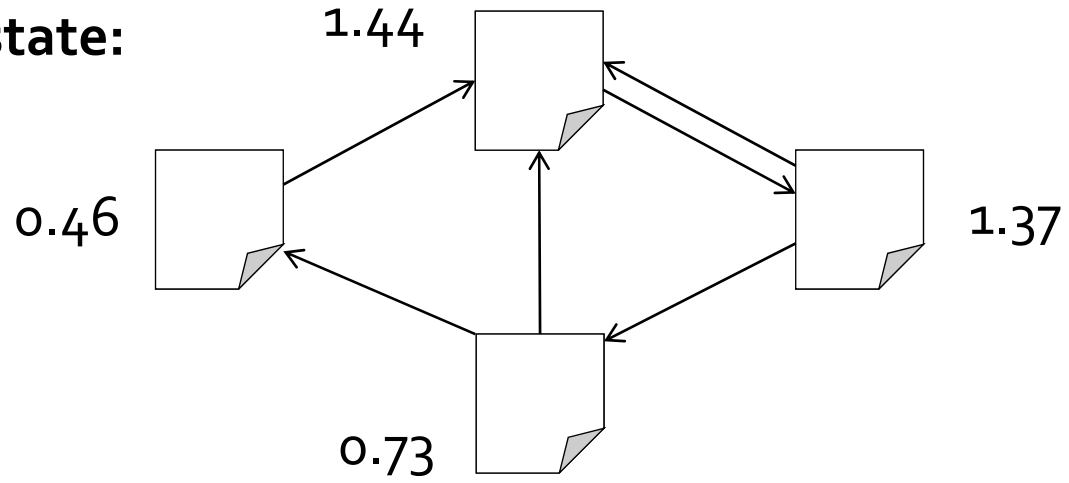


Task1: PageRank

- Algorithm:

1. Start each page at a rank of 1
2. On each iteration, have page p contribute $rank_p$ / $|neighbors_p|$ to its neighbors
3. Set each page's rank to 0.15 + 0.85 × contribs

Final state:



Task1: PageRank

```
val links = // load RDD of (url, neighbors) pairs
var ranks = // load RDD of (url, rank) pairs

for (i <- 1 to ITERATIONS) {
  val contribs = links.join(ranks).flatMap {
    case (url, (links, rank)) =>
      links.map(dest => (dest, rank/links.size))
  }
  ranks = contribs.reduceByKey(_ + _)
                  .mapValues(0.15 + 0.85 * _)
}
ranks.saveAsTextFile(...)
```

Task1: PageRank

- Start HDFS:

```
$start-all.sh
```

- Upload the data:

```
$hadoop fs -put urldata.txt  /user/epl-646/
```

- Go to SPARK home directory:

```
$cd $SPARK_HOME
```

- Start the spark-shell

```
$bin/spark-shell
```

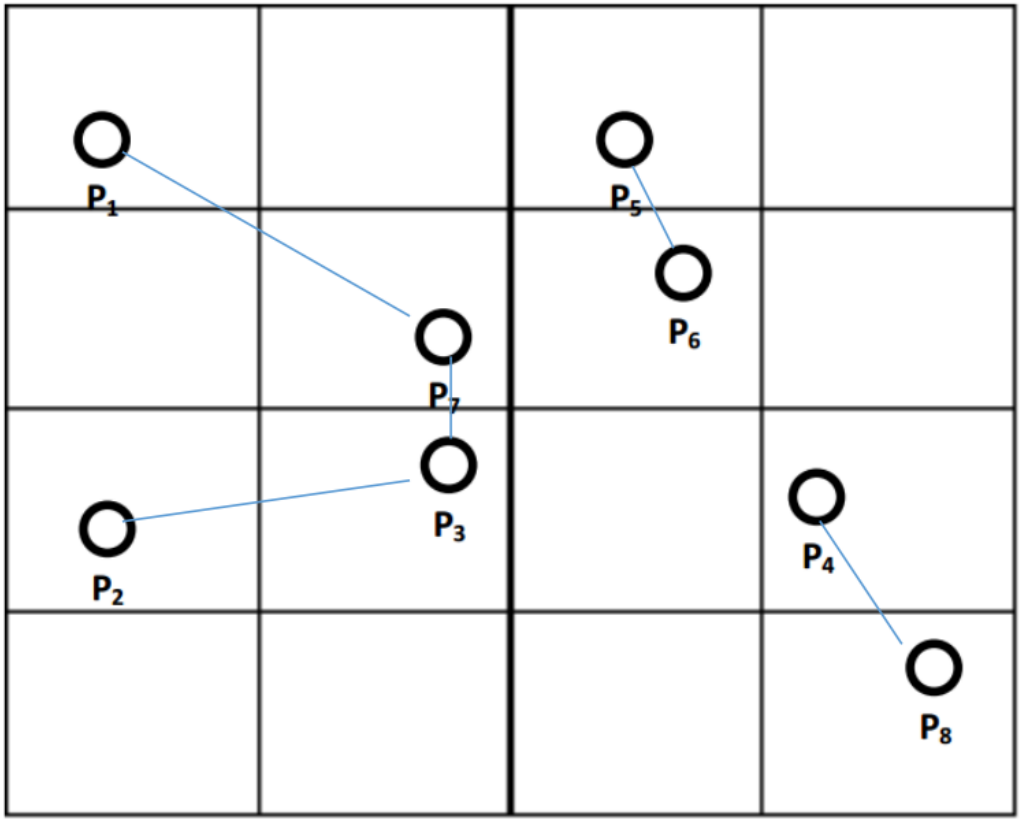
- Run interactively the code!!!

Task1: PageRank

- Results:

```
url_4 has rank: 1.3758228705372555.
url_2 has rank: 0.4633039012638519.
url_3 has rank: 0.7294952436130331.
url_1 has rank: 1.4313779845858583.
```

Task2: AkNN



A1NN(O)={(P1,P7), (P2,P3), (P7,P3), (P3,P7), (P5,P6), (P6,P5), (P4,P8), (P8,P4)}.

Task2: AkNN

- Start HDFS:
  $start-all.sh
- Upload the data:
- $hadoop fs -put random.10.data.txt /user/epl-646/
- Go to SPARK home directory:
- $cd $SPARK_HOME
- Start the spark-shell
- $bin/spark-shell
- Run interactively the code!!!

Task2: AkNN

- Results:

```
8->9,25.731 6,51.103 10,89.064 5,134.352 7,162.421 3,198.194 2,210.988
4->2,51.681 1,59.638 7,107.288 5,126.365 10,160.264 3,179.134 9,240.160
5->10,51.963 7,92.499 2,106.475 4,126.365 6,129.909 9,133.897 8,134.352
9->8,25.731 6,75.518 10,83.877 5,133.897 7,146.746 3,174.487 2,198.606
1->2,43.857 4,59.638 7,88.925 3,132.340 5,150.290 10,166.872 9,235.480
6->8,51.103 9,75.518 10,100.631 5,129.909 7,185.250 2,225.019 3,237.672
2->1,43.857 4,51.681 7,57.181 5,106.475 10,125.238 3,129.525 9,198.606
7->2,57.181 10,85.817 1,88.925 3,92.174 5,92.499 4,107.288 9,146.746
3->7,92.174 2,129.525 1,132.340 10,154.676 9,174.487 4,179.134 5,180.035
10->5,51.963 9,83.877 7,85.817 8,89.064 6,100.631 2,125.238 3,154.676
```

Questions?

http://www.cs.ucy.ac.cy/~dzeina/courses/epl646/labs/lab.html