CS5300 Project2 Report

Xi He(xh243)
Yanjia Li(yl2493)
Long Ma(lm675)

Code Functionality & Structure

1. Filter file

The followings are the parameters in terms selecting unique dataset for the node graph

NetID: xh243
rejectMin: 0.3078
rejectLimit: 0.3178
total edges: 7600595
rejected edges: 76370
Number of edges actually selected: 7524225

The input file format: Each line→ NodeID   PageRank_Value   OutLinks1 OutLink2 ...

2. Simple Computation

(1) Overview

1. Jobs: MapReduce passes. 5 passes in total so it does not converge.

2. Map(read line by line from file)
   a. In: <LineNumofFile(longWritable), LineText(Text)>
   b. out: <NodeID(longWritable), Data(Text)>

3. Reduce(write line by line to file)
   a. in: <NodeID(longWritable), List<Data>(Text)>
   b. out: <NodeID(longWritable), NodeInfoLine(Text)>
   c.

(2) Mapper

1. Mapper reads the data line by line
2. For each line of data/each node, Mapper sends two types of data (1) splitted PR value to each outlink node (2) and also send its own node info line(marked with flag so reducer can distinguish).

(3) Reducer

1. Reducer gets PR value flowing from other nodes for one node and its own node info line (distinguish the two type of data by reading flag in the string)

2. Reducer calculate new PR value for its own node and generate data input for next round. The output is exactly same as the input file in every round except the pagerank value is updated.

3. Increment the residual counter so the job can calculate the average residual.

Note: running the code requires to start from PageRank.java and it requires a argument which is a part of the input file location. In our case:
/Users/BboyKellen/Documents/workspace/project2/input/iter

3. Blocked Computation

(1) Overview

1. Jobs: MapReduce passes. AvgResidual<0.001 to break

2. Map(read line by line from file)
   c. In: <LineNumofFile(longWritable), LineText(Text)>
   d. out: <BlockID(longWritable), Data(Text)>

3. Reduce(write line by line to file)
   d. in: <BlockID(longWritable), List<Data>(Text)>
   e. out: <NodeID(longWritable), NodeInfoLine(Text)>

(2) Mapper

1. Read text file line by line.
2. Each line/node contains {NodeID, PR, Outlinks}

   ex: “123 0.234 45 86 90”

3. For each line / each node
   a. send entire line<flag, line> ex: “NodeInfo 123 0.234 45 86 90”

      send to block u

   b. send BE(edges in same block): {flag, source node id, destination node id, source node degree} ex: “BE 32 43 3”

u→ v (send to block of v)

c. send BC(boundaries) {flag, destination node id, pageRank shared with it}. ex: “BC 34 0.02 ”

u→ v (send to block of v)

(3) Reducer

1. Sort List<Data> for one block, put in the following HashMaps
   a. **degree_map**<nodeID, degree>
   b. **pageRank_map && new_pageRank_map** <nodeID, PR>
   c. **boundary_map** <nodeID, List<PR_shared>>
   d. **edge_map** <nodeID, List<inLinks> >
   e. **Others**: *forNextPass* map<nodeID, NodeInfoText> && *startMap* (same as pageRank_map

2. Run iterations (max 20 times)
   a. **edge_map**: add PR shared to each desti node to new_pageRank_map
   b. **boundary_map**: add PR shared to each desti node from other blocks
   c. **new_pageRank_map**: sum*0.85 +0.15/Num_OF_NODES
   d. Update **pageRank_map**; calculate avgResidual<0.01?break;keep iterate

3. Write To file
   a. Update Node line Info with new PR in forNextPass map. Write to file on the fly. The output is exactly same as the input file in every round except the pagerank value is updated.
   b. Sum the residuals of all nodes in block (start-end)/end, add to counter. Job driver will calculate average residual after one map-reduce pass job to determine if it converges or not.

Note: running the code requires to start from PageRank_block.java and it requires a argument which is a part of the input file location. In our case: /Users/BboyKellen/Documents/workspace/project2/input/iter

Results of Simple computation

The followings are the result from running in local computer and AWS EMR. The results from local and AWS are the same.

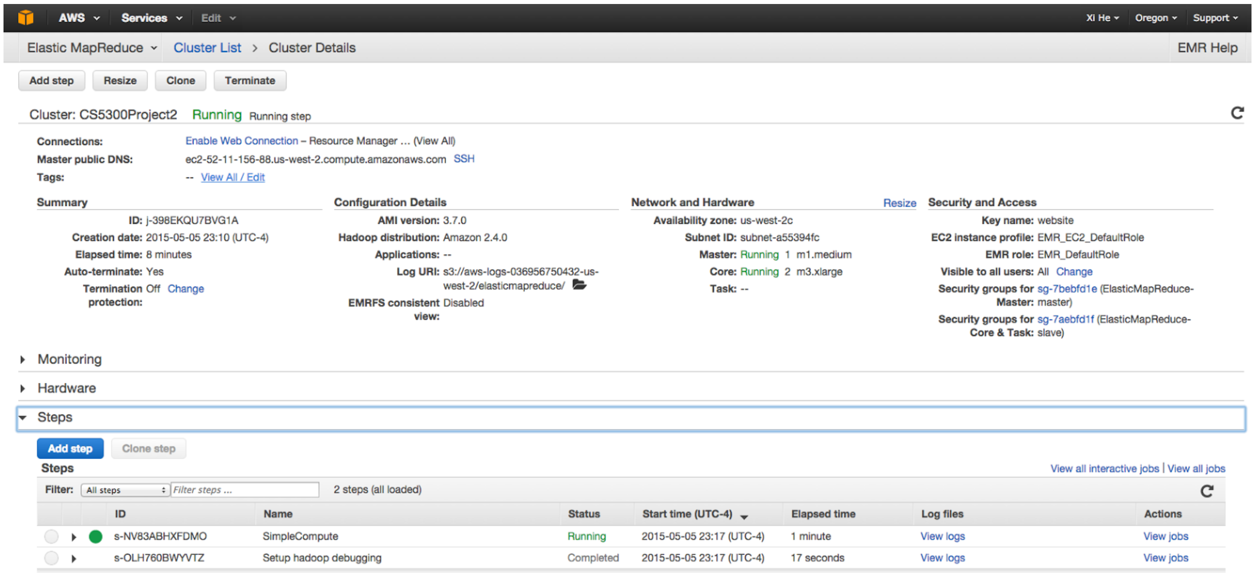
1. Run at local computer

Output from local computer:

```
The avergae residual of iter0 is 2.338917429928637
The avergae residual of iter1 is 0.32286126213096333
The avergae residual of iter2 is 0.19209467914422892
The avergae residual of iter3 is 0.09412065993899858
The avergae residual of iter4 is 0.06281029263167112
```

2. Run from Amazon AWS EMR

Using one master node and two slave nodes



Cluster: CS5300Project2 Running Running step

Connections: Enable Web Connection – Resource Manager ... (View All)
Master public DNS: ec2-52-11-156-88.us-west-2.compute.amazonaws.com SSH
Tags: -- View All / Edit

Summary
ID: j-398EKQU7BVG1A
Creation date: 2015-05-05 23:10 (UTC-4)
Elapsed time: 8 minutes
Auto-terminate: Yes
Termination protection: Off Change

Configuration Details
AMI version: 3.7.0
Hadoop distribution: Amazon 2.4.0
Applications: --
Log URI: s3://aws-logs-036956750432-us-west-2/elasticmapreduce/
EMRFS consistent view: Disabled

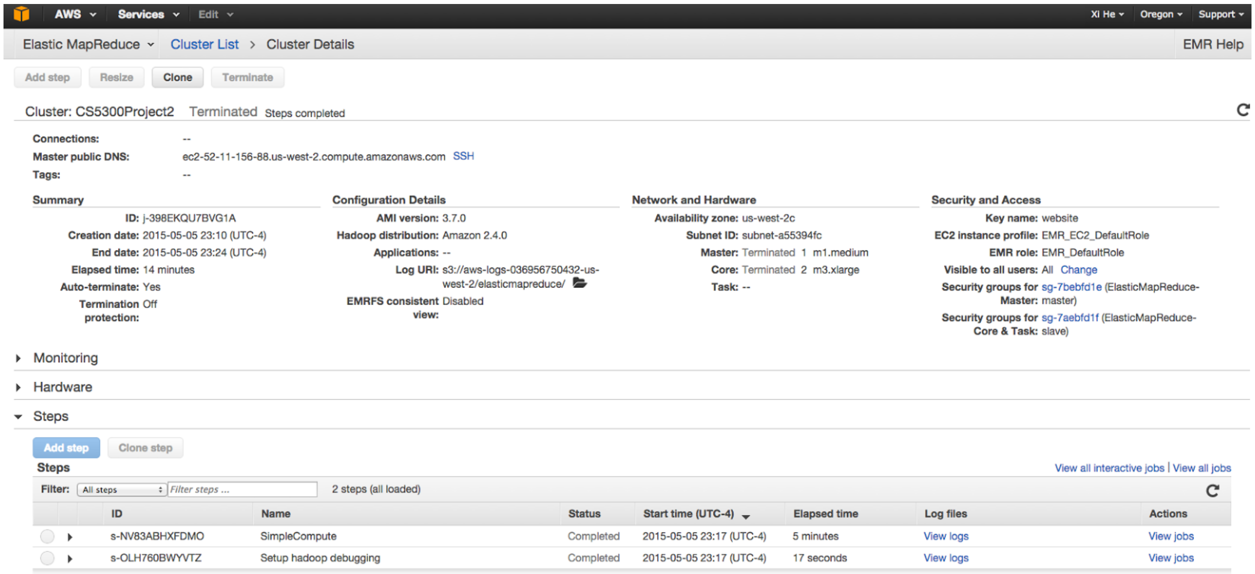
Network and Hardware Resize
Availability zone: us-west-2c
Subnet ID: subnet-a55394fc
Master: Running 1 m1.medium
Core: Running 2 m3.xlarge
Task: --

Security and Access
Key name: website
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-7bebfd1e (ElasticMapReduce-master)
Security groups for Core & Task: sg-7aebfd1f (ElasticMapReduce-slave)

Steps
View all interactive jobs | View all jobs
Filter: All steps Filter steps ... 2 steps (all loaded)

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
s-NV83ABHXFDMO	SimpleCompute	Running	2015-05-05 23:17 (UTC-4)	1 minute	View logs	View jobs
s-OLH760BWYVTZ	Setup hadoop debugging	Completed	2015-05-05 23:17 (UTC-4)	17 seconds	View logs	View jobs

Figure 1. Simple Computation Task Running



Cluster: CS5300Project2 Terminated Steps completed

Connections: --
Master public DNS: ec2-52-11-156-88.us-west-2.compute.amazonaws.com SSH
Tags: --

Summary

ID: j-398EKQU7BVG1A
Creation date: 2015-05-05 23:10 (UTC-4)
End date: 2015-05-05 23:24 (UTC-4)
Elapsed time: 14 minutes
Auto-terminate: Yes
Termination protection: Off

Configuration Details

AMI version: 3.7.0
Hadoop distribution: Amazon 2.4.0
Applications: --
Log URI: s3://aws-logs-036956750432-us-west-2/elasticmapreduce/
EMRFS consistent view: Disabled

Network and Hardware

Availability zone: us-west-2c
Subnet ID: subnet-a55394fc
Master: Terminated 1 m1.medium
Core: Terminated 2 m3.xlarge
Task: --

Security and Access

Key name: website
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-7bebfd1e (ElasticMapReduce-master)
Security groups for Core & Task: sg-7aebfd1f (ElasticMapReduce-slave)

Steps

Filter: All steps | Filter steps ... | 2 steps (all loaded)

		ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
○	▸	s-NV83ABHXFDMO	SimpleCompute	Completed	2015-05-05 23:17 (UTC-4)	5 minutes	View logs	View jobs
○	▸	s-OLH760BWYVTZ	Setup hadoop debugging	Completed	2015-05-05 23:17 (UTC-4)	17 seconds	View logs	View jobs

Figure 2. Simple Computation Task Completed in 5 minutes

Output from EMR:

```
The avergae residual of iter0 is 2.3389174300161999
The avergae residual of iter1 is 0.322861262145557
The avergae residual of iter2 is 0.19209467914422892
The avergae residual of iter3 is 0.09412065993899858
The avergae residual of iter4 is 0.06281029263167112
```

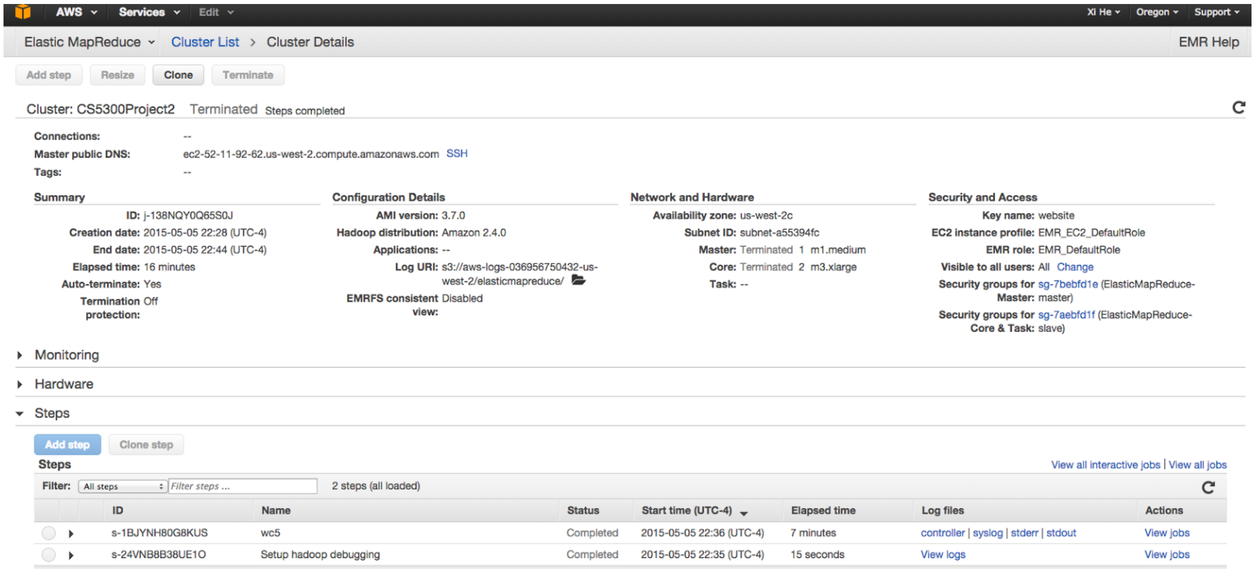
Results of the Blocked computation

It takes 6 map-reduce passes to coverage. The followings are the result from running in local computer and AWS EMR. The results from local and AWS are exactly the same.

1. Run at local computer

Average Residual:
The average residual of iter1 is **2.8138796845584695**
The average reducer iterations of pass 1 is 16.014705882352942
The average residual of iter2 is **0.03858919129343432**
The average reducer iterations of pass 2 is 7.897058823529412
The average residual of iter3 is **0.023914335347255664**
The average reducer iterations of pass 3 is 5.897058823529412
The average residual of iter4 is **0.009912790318579162**
The average reducer iterations of pass 4 is 3.926470588235294
The average residual of iter5 is **0.0038443647388468104**
The average reducer iterations of pass 5 is 2.5
The average residual of iter6 is **9.667192621455569E-4**
The average reducer iterations of pass 6 is 1.338235294117647

2. Run from Amazon AWS EMR



Cluster: CS5300Project2 Terminated Steps completed

Connections: --
Master public DNS: ec2-52-11-92-62.us-west-2.compute.amazonaws.com SSH
Tags: --

Summary
ID: j-138NQY0Q65S0J
Creation date: 2015-05-05 22:28 (UTC-4)
End date: 2015-05-05 22:44 (UTC-4)
Elapsed time: 16 minutes
Auto-terminate: Yes
Termination protection: Off

Configuration Details
AMI version: 3.7.0
Hadoop distribution: Amazon 2.4.0
Applications: --
Log URI: s3://aws-logs-036956750432-us-west-2/elasticmapreduce/
EMRFS consistent view: Disabled

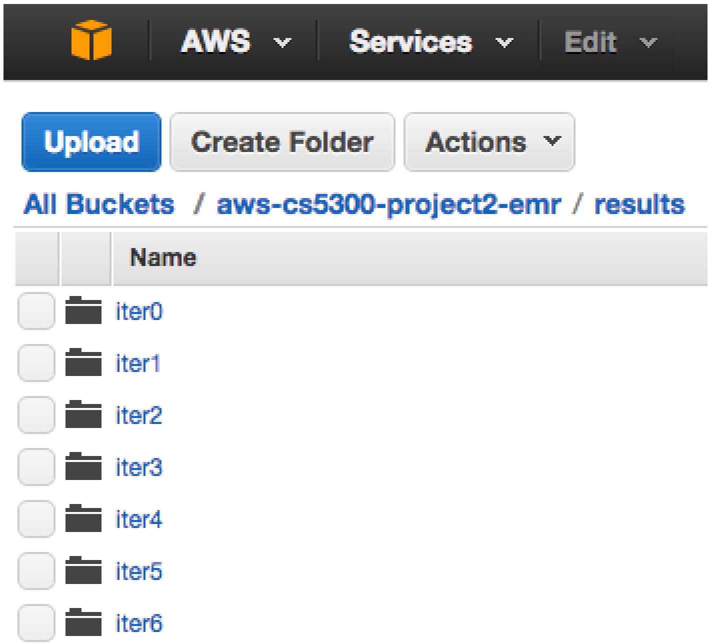
Network and Hardware
Availability zone: us-west-2c
Subnet ID: subnet-a55394fc
Master: Terminated 1 m1.medium
Core: Terminated 2 m3.xlarge
Task: --

Security and Access
Key name: website
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-7bebfd1e (ElasticMapReduce-master)
Security groups for Core & Task: sg-7aebfd1f (ElasticMapReduce-slave)

Steps
Filter: All steps Filter steps ... 2 steps (all loaded)

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
s-1BJYNH80G8KUS	wc5	Completed	2015-05-05 22:36 (UTC-4)	7 minutes	controller | syslog | stderr | stdout	View jobs
s-24VNB8B38UE1O	Setup hadoop debugging	Completed	2015-05-05 22:35 (UTC-4)	15 seconds	View logs	View jobs

Figure 3. Blocked Computation Task Completed in 7 minutes



All Buckets / aws-cs5300-project2-emr / results

Name
iter0
iter1
iter2
iter3
iter4
iter5
iter6

Figure 4. S3 iteration folders (6 pass)

Note*: iter0 is input folder, iter1- iter6 are the output folders*

Average Residual:

```
<Map_Reduce> The average residual of iter1 is 2.81387968455584695
<Map_Reduce> The average residual of iter2 is 0.03858919129343432
<Map_Reduce> The average residual of iter3 is 0.023914335347255664
<Map_Reduce> The average residual of iter4 is 0.009912790318579162
<Map_Reduce> The average residual of iter5 is 0.0038443647388468104
<Map_Reduce> The average residual of iter6 is 9.667192621455569E-4
```

The following in the PageRank values for the two lowest-numbered Nodes in each Block. This is when it converged which is after the 6th iteration.

```
Block 0: nodeID 0 PageRank val: 1.7523507706503166E-5
Block 0: nodeID 1 PageRank val: 5.247302353180255E-5
Block 1: nodeID 10328 PageRank val: 4.320536703455484E-7
Block 1: nodeID 10329 PageRank val: 2.4343234927539884E-7
Block 2: nodeID 20373 PageRank val: 2.1890460137472089E-7
Block 2: nodeID 20374 PageRank val: 2.1890460137472089E-7
Block 3: nodeID 30629 PageRank val: 2.65351916910355E-7
Block 3: nodeID 30630 PageRank val: 2.1890460137472089E-7
Block 4: nodeID 40645 PageRank val: 2.4785173536526804E-5
Block 4: nodeID 40646 PageRank val: 2.4785173536526804E-5
Block 5: nodeID 50462 PageRank val: 0.004977779022854708
Block 5: nodeID 50463 PageRank val: 0.005057745184055878
Block 6: nodeID 60841 PageRank val: 1.2146979118951305E-5
Block 6: nodeID 60842 PageRank val: 3.108408989160212E-6
Block 7: nodeID 70591 PageRank val: 2.4785173536526804E-5
Block 7: nodeID 70592 PageRank val: 2.4785173536526804E-5
Block 8: nodeID 80118 PageRank val: 5.258927811254045E-7
Block 8: nodeID 80119 PageRank val: 9.729016450819983E-7
Block 9: nodeID 90497 PageRank val: 9.128502044150412E-7
Block 9: nodeID 90498 PageRank val: 1.1314892274058739E-5
Block 10: nodeID 100501 PageRank val: 1.0375565648745236E-6
Block 10: nodeID 100502 PageRank val: 1.2190843200423103E-6
Block 11: nodeID 110567 PageRank val: 0.0010554128069618766
Block 11: nodeID 110568 PageRank val: 9.644197267211728E-5
Block 12: nodeID 120945 PageRank val: 8.090945096697029E-7
Block 12: nodeID 120946 PageRank val: 6.691256160272805E-7
Block 13: nodeID 130999 PageRank val: 2.287043002976799E-7
Block 13: nodeID 131000 PageRank val: 2.287043002976799E-7
Block 14: nodeID 140574 PageRank val: 0.0010429502553185982
Block 14: nodeID 140575 PageRank val: 2.287043002976799E-7
Block 15: nodeID 150953 PageRank val: 9.950038995034244E-4
Block 15: nodeID 150954 PageRank val: 0.001796355564433476
Block 16: nodeID 161332 PageRank val: 2.2854548267780417E-7
Block 16: nodeID 161333 PageRank val: 2.2854548267780417E-7
Block 17: nodeID 171154 PageRank val: 2.2937950464802463E-7
Block 17: nodeID 171155 PageRank val: 2.2937950464802463E-7
Block 18: nodeID 181514 PageRank val: 2.245912327480114E-7
Block 18: nodeID 181515 PageRank val: 7.432491902512471E-4
Block 19: nodeID 191625 PageRank val: 2.587950552753508E-4
Block 19: nodeID 191626 PageRank val: 2.1890460137472089E-7
Block 20: nodeID 202004 PageRank val: 9.108084685489672E-7
Block 20: nodeID 202005 PageRank val: 9.659526838870594E-7
Block 21: nodeID 212383 PageRank val: 0.0011315697473620518
Block 21: nodeID 212384 PageRank val: 2.1890460137472089E-7
Block 22: nodeID 222762 PageRank val: 2.287301988613208E-7
Block 22: nodeID 222763 PageRank val: 2.287301988613208E-7
Block 23: nodeID 232593 PageRank val: 2.3668267144597314E-7
Block 23: nodeID 232594 PageRank val: 2.968574935685045E-7
Block 24: nodeID 242878 PageRank val: 1.087991708726062E-6
Block 24: nodeID 242879 PageRank val: 5.809708071903946E-7
Block 25: nodeID 252938 PageRank val: 5.971063617589379E-6
Block 25: nodeID 252939 PageRank val: 5.445771785664395E-7
Block 26: nodeID 263149 PageRank val: 9.764425522370657E-6
Block 26: nodeID 263150 PageRank val: 1.7950395696361192E-6
Block 27: nodeID 273210 PageRank val: 2.7561792566085423E-5
Block 27: nodeID 273211 PageRank val: 2.4476685721546168E-6
Block 28: nodeID 283473 PageRank val: 2.8760046520703846E-7
Block 28: nodeID 283474 PageRank val: 2.8760046520703846E-7
Block 29: nodeID 293255 PageRank val: 2.727146044193484E-6
Block 29: nodeID 293256 PageRank val: 1.1149665154001324E-5
Block 30: nodeID 303043 PageRank val: 9.767189507287338E-7
Block 30: nodeID 303044 PageRank val: 5.16957509354777E-6
Block 31: nodeID 313370 PageRank val: 2.549039249713656E-7
Block 31: nodeID 313371 PageRank val: 2.3423949100377943E-5
Block 32: nodeID 323522 PageRank val: 5.414059849650272E-7
Block 32: nodeID 323523 PageRank val: 1.7597247306889524E-6
Block 33: nodeID 333883 PageRank val: 1.7659301588113937E-5
Block 33: nodeID 333884 PageRank val: 4.067090383612633E-6
Block 34: nodeID 343663 PageRank val: 2.443580369096522E-6
Block 34: nodeID 343664 PageRank val: 3.664019253241096E-7
Block 35: nodeID 353645 PageRank val: 1.994186166918186E-6
Block 35: nodeID 353646 PageRank val: 9.25275723270379E-7
Block 36: nodeID 363929 PageRank val: 1.5068473807499254E-5
Block 36: nodeID 363930 PageRank val: 4.6063642198925323E-7
Block 37: nodeID 374236 PageRank val: 2.39331041910047E-7
Block 37: nodeID 374237 PageRank val: 9.282129689120674E-6
Block 38: nodeID 384554 PageRank val: 6.582568430548817E-5
Block 38: nodeID 384555 PageRank val: 5.294252279764473E-6
Block 39: nodeID 394929 PageRank val: 3.95704187266802E-6
Block 39: nodeID 394930 PageRank val: 2.5961696861641865E-6
Block 40: nodeID 404712 PageRank val: 6.892586413844494E-7
Block 40: nodeID 404713 PageRank val: 8.030218198680796E-7
Block 41: nodeID 414617 PageRank val: 4.883586314017715E-7
Block 41: nodeID 414618 PageRank val: 3.7004729544201926E-7
Block 42: nodeID 424747 PageRank val: 2.530613162138082E-6
Block 42: nodeID 424748 PageRank val: 3.5026288102322745E-6
Block 43: nodeID 434707 PageRank val: 5.281027023093166E-7
Block 43: nodeID 434708 PageRank val: 3.2808462434874405E-6
Block 44: nodeID 444489 PageRank val: 5.069018773975154E-7
Block 44: nodeID 444490 PageRank val: 2.4030404467089753E-7
Block 45: nodeID 454285 PageRank val: 3.5183009441460406E-7
Block 45: nodeID 454286 PageRank val: 3.186396811408006E-7
Block 46: nodeID 464398 PageRank val: 5.294432868591476E-7
Block 46: nodeID 464399 PageRank val: 5.294432868591476E-7
Block 47: nodeID 474196 PageRank val: 5.249324563858699E-7
Block 47: nodeID 474197 PageRank val: 8.997104298039208E-7
Block 48: nodeID 484050 PageRank val: 1.6334529172060845E-5
Block 48: nodeID 484051 PageRank val: 1.2353797504037491E-5
Block 49: nodeID 493968 PageRank val: 8.264413346754528E-7
Block 49: nodeID 493969 PageRank val: 1.3861962650324838E-6
Block 50: nodeID 503752 PageRank val: 7.909762200078046E-7
Block 50: nodeID 503753 PageRank val: 7.909762200078046E-7
Block 51: nodeID 514131 PageRank val: 0.0011086330010622722
Block 51: nodeID 514132 PageRank val: 2.6297314402703114E-5
Block 52: nodeID 524510 PageRank val: 8.902046034406036E-4
Block 52: nodeID 524511 PageRank val: 6.581354509556912E-5
Block 53: nodeID 534709 PageRank val: 0.009100999807326645
Block 53: nodeID 534710 PageRank val: 4.136521163106482E-6
Block 54: nodeID 545088 PageRank val: 0.0017604983452414321
Block 54: nodeID 545089 PageRank val: 3.256691020403494E-5
Block 55: nodeID 555467 PageRank val: 0.0018000826033106473
Block 55: nodeID 555468 PageRank val: 7.477788539966055E-7
Block 56: nodeID 565846 PageRank val: 2.1890460137472089E-7
Block 56: nodeID 565847 PageRank val: 2.1890460137472089E-7
Block 57: nodeID 576225 PageRank val: 1.1623800019976774E-5
Block 57: nodeID 576226 PageRank val: 2.1890460137472089E-7
Block 58: nodeID 586604 PageRank val: 3.905993033379477E-4
Block 58: nodeID 586605 PageRank val: 4.1579489769031345E-7
Block 59: nodeID 596585 PageRank val: 1.3328991029953962E-6
Block 59: nodeID 596586 PageRank val: 1.3861606038888671E-6
Block 60: nodeID 606367 PageRank val: 3.1553275902627255E-7
Block 60: nodeID 606368 PageRank val: 2.2475424731411227E-7
Block 61: nodeID 616148 PageRank val: 9.839328759630283E-7
Block 61: nodeID 616149 PageRank val: 4.282200854154952E-7
Block 62: nodeID 626448 PageRank val: 2.1890460137472089E-7
Block 62: nodeID 626449 PageRank val: 2.5808132575326E-7
Block 63: nodeID 636240 PageRank val: 2.2894929629334115E-7
Block 63: nodeID 636241 PageRank val: 2.287043002976799E-7
Block 64: nodeID 646022 PageRank val: 5.68202306226931E-7
Block 64: nodeID 646023 PageRank val: 4.109379868053116E-7
Block 65: nodeID 655804 PageRank val: 2.2894929629334115E-7
Block 65: nodeID 655805 PageRank val: 2.1890460137472089E-7
Block 66: nodeID 665666 PageRank val: 2.5611838360842345E-7
Block 66: nodeID 665667 PageRank val: 4.0855816343692305E-7
Block 67: nodeID 675448 PageRank val: 2.1890460137472089E-7
Block 67: nodeID 675449 PageRank val: 4.897094180747991E-6
```