Homework 2 Node Similarity and Community Detection

10 points, Due: March 04 2016 [Solutions]

Question 1 [1.5 pts] Please use your own language to briefly explain the following concepts:

PageRank score:
PageRank is an algorithm for calculating the importance of a node (such as a website) in a network (such as Internet). The principle of the PageRank algorithm is to rank a page according to terms appeared at the pages linking to it. The PageRank score of a node *u* is the linear combination of the PageRank scores of all nodes which directly point to node *u*.

Rooted PageRank:
Rooted pagerank denote a special random walk process where a random surfer starts from a node x, and has a constant probability value to return back to x in each step. During rooted Pagerank, a random surfer starts from node x, and follows out-links to move the next node in a random manner. At each step, assume the random surfer is currently located at node y, it will has a probability 1-α to return back to node x, otherwise, it will randomly pick an out-link of y, and walks to the next node.

Network community:
A subset of network where nodes inside the community have a higher degree of interactions between each other than the interactions with nodes outside the community.

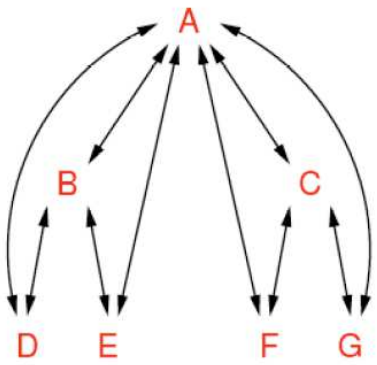
Clique:
A network (or a graph) or a subset of network where all nodes are connected to each other.

k-Clique:
A maximal subgraph in which the largest geodesic distance between any nodes is less or equal to k.

Low-rank approximation:
Rank of a matrix is the number of linearly independent columns (or row) of the matrix. In low-rank approximation, one intends to produce an approximation of a given matrix (say A), by reducing the rank of the matrix A. Low-rank approximation can be used to assess the similarity between nodes in the network.

Question 2 [2 pts]: Given seven web pages with the following link structure,

1. Please use "Power Iteration" (a.k.a simple iteration) to calculate the PageRank scores for each website. (You only need to show the first and the second iterations results, with the initial PageRank scores for each node being set as 1/n=0.15) [1 pt].
2. Please also use Eigenvector based approach to calculate PageRank scores for each web page [1 pt] (please show your solutions.)



Solution

Adjacency Matrix: M=

	A	B	C	D	E	F	G
A	0	1	1	1	1	1	1
B	1	0	0	1	1	0	0
C	1	0	0	0	0	1	1
D	1	1	0	0	0	0	0
E	1	1	0	0	0	0	0
F	1	0	1	0	0	0	0
G	1	0	1	0	0	0	0

After Normalization: M'=

	A	B	C	D	E	F	G
A	0	1/3	1/3	1/2	1/2	1/2	1/2
B	1/6	0	0	1/2	1/2	0	0
C	1/6	0	0	0	0	1/2	1/2
D	1/6	1/3	0	0	0	0	0
E	1/6	1/3	0	0	0	0	0
F	1/6	0	1/3	0	0	0	0
G	1/6	0	1/3	0	0	0	0

The initial importance score = [0.15 0.15 0.15 0.15 0.15 0.15 0.15]

After first iteration = [0.4 0.175 0.175 0.075 0.075 0.075 0.075]

After 2nd iteration = [0.2667 0.1417 0.1417 0.1250 0.1250 0.1250 0.1250]

Use PageRank algorithm: Calculate eigenvalue of M'

Input matrix:

```
0.000 0.333 0.333 0.500 0.500 0.500 0.500
0.167 0.000 0.000 0.500 0.500 0.000 0.000
0.167 0.000 0.000 0.000 0.000 0.500 0.500
0.167 0.333 0.000 0.000 0.000 0.000 0.000
0.167 0.333 0.000 0.000 0.000 0.000 0.000
0.167 0.000 0.333 0.000 0.000 0.000 0.000
0.167 0.000 0.333 0.000 0.000 0.000 0.000
```

Eigenvalues Eigenvectors:

Eigenvalues:

```
( 1.000, 0.000i)
(-0.333, 0.000i)
(-0.667, 0.000i)
( 0.577, 0.000i)
(-0.577, 0.000i)
( 0.000, 0.000i)
( 0.000, 0.000i)
```

Eigenvectors:

```
( 0.717, 0.000i) ( 0.816, 0.000i) (-0.633, 0.000i)
( 0.359, 0.000i) (-0.409, 0.000i) (-0.316, 0.000i)
( 0.359, 0.000i) (-0.409, 0.000i) (-0.316, 0.000i)
( 0.239, 0.000i) ( 0.000, 0.000i) ( 0.316, 0.000i)
( 0.239, 0.000i) ( 0.000, 0.000i) ( 0.316, 0.000i)
( 0.239, 0.000i) ( 0.000, 0.000i) ( 0.316, 0.000i)
( 0.239, 0.000i) ( 0.000, 0.000i) ( 0.316, 0.000i)
```

The final PageRank Score:

[0.717 0.359 0.359 0.239 0.239 0.239 0.239]

Question 3 [1 pt]: In Quesiton 2, please use rooted PageRank to calculate similarity between each pair of nodes. Each time, the random walker has a probability 1-α (where α=*0.2*) to return back to an original node. (Please show your solutions).

Solution

Adjacency Matrix: A=

	A	B	C	D	E	F	G
A	0	1	1	1	1	1	1
B	1	0	0	1	1	0	0
C	1	0	0	0	0	1	1
D	1	1	0	0	0	0	0
E	1	1	0	0	0	0	0
F	1	0	1	0	0	0	0
G	1	0	1	0	0	0	0

D Matrix:

	A	B	C	D	E	F	G
A	6	0	0	0	0	0	0
B	0	3	0	0	0	0	0
C	0	0	3	0	0	0	0
D	0	0	0	2	0	0	0
E	0	0	0	0	2	0	0
F	0	0	0	0	0	2	0
G	0	0	0	0	0	0	2

D^{-1} Matrix:

	A	B	C	D	E	F	G
A	0.167	0	0	0	0	0	0
B	0	0.333	0	0	0	0	0
C	0	0	0.333	0	0	0	0
D	0	0	0	0.5	0	0	0
E	0	0	0	0	0.5	0	0
F	0	0	0	0	0	0.5	0
G	0	0	0	0	0	0	0.5

D^{-1}×A Matrix:

```
0.000 0.167 0.167 0.167 0.167 0.167 0.167
0.333 0.000 0.000 0.333 0.333 0.000 0.000
0.333 0.000 0.000 0.000 0.000 0.333 0.333
0.500 0.500 0.000 0.000 0.000 0.000 0.000
0.500 0.500 0.000 0.000 0.000 0.000 0.000
0.500 0.000 0.500 0.000 0.000 0.000 0.000
```
0.500 0.000 0.500 0.000 0.000 0.000 0.000

$\alpha \times D^{-1} \times A$ Matrix:

0	0.03	0.03	0.033	0.033	0.033	0.033
0.07	0	0	0.067	0.067	0	0
0.07	0	0	0	0	0.067	0.067
0.1	0.1	0	0	0	0	0
0.1	0.1	0	0	0	0	0
0.1	0	0.1	0	0	0	0
0.1	0	0.1	0	0	0	0

$I - \alpha \times D^{-1} \times A$ Matrix:

```
1.000 -0.030 -0.030 -0.033 -0.033 -0.033 -0.033
-0.070  1.000  0.000 -0.067 -0.067  0.000  0.000
-0.070  0.000  1.000  0.000  0.000 -0.067 -0.067
-0.100 -0.100  0.000  1.000  0.000  0.000  0.000
-0.100 -0.100  0.000  0.000  1.000  0.000  0.000
-0.100  0.000 -0.100  0.000  0.000  1.000  0.000
```
-0.100 0.000 -0.100 0.000 0.000 0.000 1.000

Inverse $(I - \alpha \times D^{-1} \times A)^{-1}$:

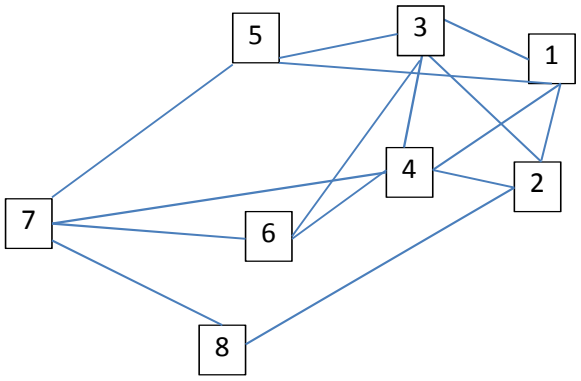
```
1.020 0.038 0.038 0.036 0.036 0.036 0.036
0.086 1.017 0.003 0.071 0.071 0.003 0.003
0.086 0.003 1.017 0.003 0.003 0.071 0.071
0.111 0.105 0.004 1.011 0.011 0.004 0.004
0.111 0.105 0.004 0.011 1.011 0.004 0.004
0.111 0.004 0.105 0.004 0.004 1.011 0.011
```
0.111 0.004 0.105 0.004 0.004 0.011 1.011

$(1-\alpha)\times (\mathbf{I} - \alpha \times \mathbf{D}^{-1} \times \mathbf{A})^{-1}$ (Pair-wise node similarity:)

0.816	0.03	0.03	0.029	0.029	0.029	0.029
0.069	0.814	0.002	0.057	0.057	0.002	0.002
0.069	0.002	0.814	0.002	0.002	0.057	0.057
0.089	0.084	0.003	0.809	0.009	0.003	0.003
0.089	0.084	0.003	0.009	0.809	0.003	0.003
0.089	0.003	0.084	0.003	0.003	0.809	0.009
0.089	0.003	0.084	0.003	0.003	0.009	0.809

Question 4 [1.5 pts]: The following networks show connections between 8 individuals in a small community. For node pairs (1, 7) and (1, 6), please use following measures to calculate their similarity (or distance) value and conclude which pair is more likely to form a link.

a. Jacarrd's Coefficient (0.25 pt)
b. Adamic/Adar (0.25 pt)
c. Preferential attachment (0.25 pt)
d. Katz (with β=0.05) (0.25 pt)
e. SimRank score with C=1 (please show the SimRank score after the 1st iteration). (0.5 pt)



Solution

Jacarrd's Coefficient:

(1, 7) = |4,5|/|2,3,4,5,6,8|=2/6=0.333

(1,6) =|3,4|/|2,3,4,5,7|=2/5=0.4

Adamic/Adar:

(1,6) = 1/log(5)+1/log(5)=1.2426

(1,7)=1/log(5)+1/log(3)=1.531

Preferential Attachment:

(1,6) = 4*3=12

(1,7)=4*4=16

Katz: (β=0.05)

E.g Numb of length-3 path between (1,6) is 7

(1 3, 4, 6); (1, 2, 3, 6); (1, 2, 4, 6); (1, 4, 3, 6); (1, 5, 7, 6); (1, 4, 7, 6); (1, 5, 3, 6)

Weighted:

(1,6)= β^1×0 +β^2×2+β^3×5+β^4×6+β^5×7+β^6×9+β^7×3=0.00595

(1,7)= β^1×0 +β^2×2+β^3×6+β^4×10+β^5×9+β^6×2+β^7×1=0.0058

Unweighted:

(1,6)= β^1×0 +β^2×1+β^3×1+β^4×1+β^5×1+β^6×1+β^7×1=0.00251

(1,7)= β^1×0 +β^2×1+β^3×1+β^4×1+β^5×1+β^6×1+β^7×1=0.00251

SimRank score with C=1

Neighborhood:

Nod 1: {2, 3, 4, 5}; Node 2: {1, 3, 4, 8}; Node 3: {1, 2, 3, 5, 6}; Node 4: {1, 2, 3, 6, 7}

Node 5: {1, 2, 7}; Node 6: {3, 4, 7}; Node 7: {4, 5, 6, 8}; Node 8: {2, 7}

After First iteration:

Sim(1,6)=1/(4×3)*2=1/6

Sim(1,7)=1/(4×4)*2=1/8

After Second iteration:

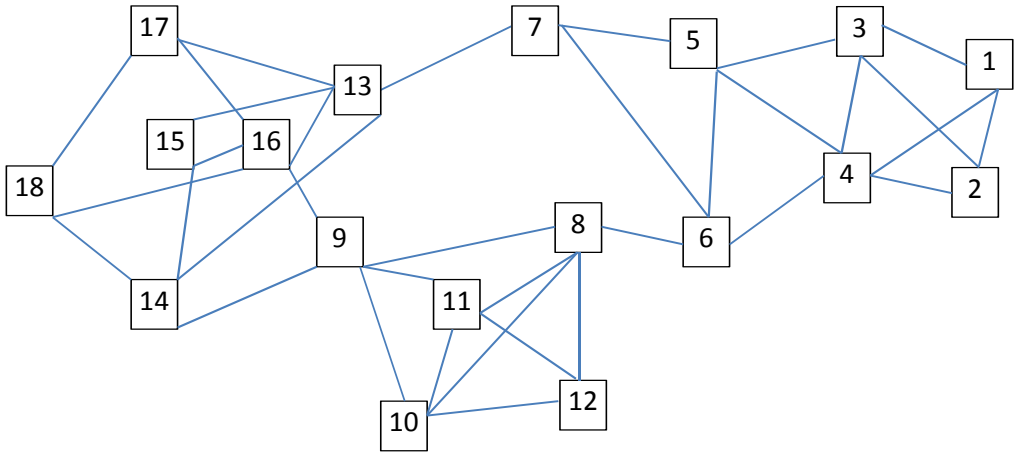
Sim(1, 6)=0.1193

Sim(1,7)=0.1526

Question 5 [4 pts]: In the following network,

1. Please find the complete set of communities by using 3-clquie [0.25pt], 3-club [0.25pt], and 3-core [0.25pt], respectively (If there are multiple sets, please just report the top three sets with the maximum number of nodes).
2. Please calculate the Geodesic distance between each pair of nodes, and use Multidimensional Scaling (MDS) to convert the network into a two dimensional space. Please report the values of all nodes in the two dimensional space and draw all nodes in the two dimensional space [1.25 pt].
3. Implement a k-means clustering algorithm (selecting k=2 and using node 18 and node 1 as the initial centers), and report the community structures after 10 iterations (You may use any other third party tools for k-means clustering. Or you can follow the k-means Excel implementation in the following URL to calculate the results) [2 pts]

k-means: http://www.csse.monash.edu.au/courseware/cse5230/2004/assets/clustering.pdf



Solution

k-clique: a maximal subgraph in which the largest geodesic distance between any nodes <= k

3-Clique: {5,6,7,8,9,10,11,13,14,15,16,17} (the largest)

3-Clique: {5,6,7,8,9,11,13,14,15,16,17}

3-Clique: {5,6,7,8,9,10,13,14,15,16,17}

The list of all 3-cliques is attached to the BB.

k-club: a substructure of diameter <= k

3-Club: {5,6,7,8,9,10,11,13,14,15,16,17} (the largest)

3-Club: {5,6,7,8,9,11,13,14,15,16,17}

3-Club: {5,6,7,8,9,10,13,14,15,16,17}

k-core: a substracture that each node connects to at least k members within the group

3-core: {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18}

Geodesic distance:

G =

0	1	1	1	2	2	3	3	4	4	4	4	4	5	5	5	5	6
1	0	1	1	2	2	3	3	4	4	4	4	4	5	5	5	5	6
1	1	0	1	1	2	2	3	4	4	4	4	3	4	4	4	4	5
1	1	1	0	1	1	2	2	3	3	3	3	3	4	4	4	4	5
2	2	1	1	0	1	1	2	3	3	3	3	2	3	3	3	3	4
2	2	2	1	1	0	1	1	2	2	2	2	2	3	3	3	3	4
3	3	2	2	1	1	0	2	3	3	3	3	1	2	2	2	2	3
3	3	3	2	2	1	2	0	1	1	1	1	3	2	3	2	3	3
4	4	4	3	3	2	3	1	0	1	1	2	2	1	2	1	2	2
4	4	4	3	3	2	3	1	1	0	1	1	3	2	3	2	3	3
4	4	4	3	3	2	3	1	1	1	0	1	3	2	3	2	3	3
4	4	4	3	3	2	3	1	2	1	1	0	4	3	4	3	4	4
4	4	3	3	2	2	1	3	2	3	3	4	0	1	1	1	1	2
5	5	4	4	3	3	2	2	1	2	2	3	1	0	1	2	2	1
5	5	4	4	3	3	2	3	2	3	3	4	1	1	0	1	2	2
5	5	4	4	3	3	2	2	1	2	2	3	1	2	1	0	1	1
5	5	4	4	3	3	2	3	2	3	3	4	1	2	2	1	0	1
6	6	5	5	4	4	3	3	2	3	3	4	2	1	2	1	1	0

P-matrix

P =

9.5154	9.0154	7.2932	6.2932	3.7099	3.1543	0.821	0.6543	-2.4012	-1.7623	-1.7623	-0.6235	-2.1235	-5.9846	-5.2901	-6.0679	-5.2901	-9.1512
9.0154	9.5154	7.2932	6.2932	3.7099	3.1543	0.821	0.6543	-2.4012	-1.7623	-1.7623	-0.6235	-2.1235	-5.9846	-5.2901	-6.0679	-5.2901	-9.1512
7.2932	7.2932	6.071	4.571	3.4877	1.4321	1.5988	-1.0679	-4.1235	-3.4846	-3.4846	-2.3457	-0.3457	-3.2068	-2.5123	-3.2901	-2.5123	-5.3735
6.2932	6.2932	4.571	4.071	2.4877	1.9321	0.5988	0.4321	-1.6235	-0.9846	-0.9846	0.1543	-1.3457	-4.2068	-3.5123	-4.2901	-3.5123	-6.3735
3.7099	3.7099	3.4877	2.4877	1.9043	0.8488	1.0154	-0.6512	-2.7068	-2.0679	-2.0679	-0.929	0.071	-1.7901	-1.0957	-1.8735	-1.0957	-2.9568
3.1543	3.1543	1.4321	1.9321	0.8488	0.7932	0.4599	0.2932	-0.7623	-0.1235	-0.1235	1.0154	-0.4846	-2.3457	-1.6512	-2.429	-1.6512	-3.5123
0.821	0.821	1.5988	0.5988	1.0154	0.4599	1.1265	-1.0401	-3.0957	-2.4568	-2.4568	-1.3179	1.1821	0.321	1.0154	0.2377	1.0154	0.1543
0.6543	0.6543	-1.0679	0.4321	-0.6512	0.2932	-1.0401	0.7932	0.7377	1.3765	1.3765	2.5154	-2.9846	0.1543	-1.6512	0.071	-1.6512	-0.0123
-2.4012	-2.4012	-4.1235	-1.6235	-2.7068	-0.7623	-3.0957	0.7377	1.6821	1.821	1.821	1.4599	-0.0401	2.0988	1.2932	2.0154	1.2932	2.9321
-1.7623	-1.7623	-3.4846	-0.9846	-2.0679	-0.1235	-2.4568	1.3765	1.821	2.9599	2.4599	3.5988	-1.9012	1.2377	-0.5679	1.1543	-0.5679	1.071
-1.7623	-1.7623	-3.4846	-0.9846	-2.0679	-0.1235	-2.4568	1.3765	1.821	2.4599	2.9599	3.5988	-1.9012	1.2377	-0.5679	1.1543	-0.5679	1.071
-0.6235	-0.6235	-2.3457	0.1543	-0.929	1.0154	-1.3179	2.5154	1.4599	3.5988	3.5988	5.2377	-4.2623	-0.1235	-2.929	-0.2068	-2.929	-1.2901
-2.1235	-2.1235	-0.3457	-1.3457	0.071	-0.4846	1.1821	-2.9846	-0.0401	-1.9012	-1.9012	-4.2623	2.2377	2.3765	3.071	2.2932	3.071	3.2099
-5.9846	-5.9846	-3.2068	-4.2068	-1.7901	-2.3457	0.321	0.1543	2.0988	1.2377	1.2377	-0.1235	2.3765	3.5154	3.7099	1.4321	2.2099	5.3488
-5.2901	-5.2901	-2.5123	-3.5123	-1.0957	-1.6512	1.0154	-1.6512	1.2932	-0.5679	-0.5679	-2.929	3.071	3.7099	4.9043	3.6265	2.9043	4.5432
-6.0679	-6.0679	-3.2901	-4.2901	-1.8735	-2.429	0.2377	0.071	2.0154	1.1543	1.1543	-0.2068	2.2932	1.4321	3.6265	3.3488	3.6265	5.2654
-5.2901	-5.2901	-2.5123	-3.5123	-1.0957	-1.6512	1.0154	-1.6512	1.2932	-0.5679	-0.5679	-2.929	3.071	2.2099	2.9043	3.6265	4.9043	6.0432
-9.1512	-9.1512	-5.3735	-6.3735	-2.9568	-3.5123	0.1543	-0.0123	2.9321	1.071	1.071	-1.2901	3.2099	5.3488	4.5432	5.2654	6.0432	8.1821

The fist and the second eigen values:

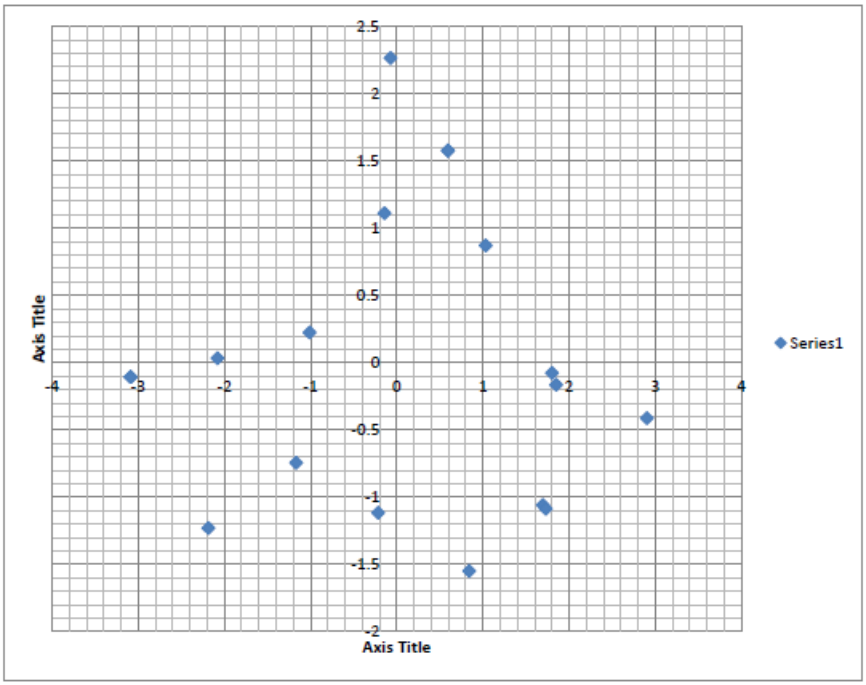
λ_1=54.0732, λ_2=20.3791.

The final MDS values:

MDS =

-3.0892	-0.107
-3.0892	-0.107
-2.1847	-1.2306
-2.0788	0.033
-1.1662	-0.744
-1.0089	0.2235
-0.2155	-1.1164
-0.1404	1.1092
1.0339	0.8708
0.5956	1.5764
0.5956	1.5764
-0.0721	2.2644
0.8412	-1.5493
1.8008	-0.0767
1.6979	-1.0586
1.8501	-0.1657
1.7295	-1.0861
2.9009	-0.4126

Projection on the two-dimension space:



K-mean clustering

We use the data from matrix MDS.

Frist, we select two centers; C_1 = (2.9009, -0.4126), C_2 = (-3.0892, -0.107)

Frist	K1_X	K1_Y	K2_X	K2_Y
	2.9009	-0.4126	-3.0892	-0.107

X	Y	D1	D2	cluster	K1_X	K1_Y	K2_X	K2_Y
-3.0892	-0.107	5.99789	0	K2			-3.0892	-0.107
-3.0892	-0.107	5.99789	0	K2			-3.0892	-0.107
-2.1847	-1.2306	5.150966	1.442428	K2			-2.1847	-1.2306
-2.0788	0.033	4.999597	1.020053	K2			-2.0788	0.033
-1.1662	-0.744	4.080579	2.025759	K2			-1.1662	-0.744
-1.0089	0.2235	3.961207	2.10639	K2			-1.0089	0.2235
-0.2155	-1.1164	3.194884	3.045823	K2			-0.2155	-1.1164
-0.1404	1.1092	3.400791	3.189759	K2			-0.1404	1.1092
1.0339	0.8708	2.265569	4.237458	K1	1.0339	0.8708		
0.5956	1.5764	3.044754	4.051122	K1	0.5956	1.5764		
0.5956	1.5764	3.044754	4.051122	K1	0.5956	1.5764		
-0.0721	2.2644	4.000632	3.837503	K2			-0.0721	2.2644
0.8412	-1.5493	2.352541	4.186678	K1	0.8412	-1.5493		
1.8008	-0.0767	1.150239	4.890094	K1	1.8008	-0.0767		
1.6979	-1.0586	1.365476	4.880765	K1	1.6979	-1.0586		
1.8501	-0.1657	1.079417	4.939649	K1	1.8501	-0.1657		
1.7295	-1.0861	1.351214	4.917164	K1	1.7295	-1.0861		
2.9009	-0.4126	0	5.99789	K1	2.9009	-0.4126		

At this time, we can get the new centers = >

New	K1_X (AVG.)	K1_Y(AVG.)	K2_X(AVG.)	K2_(AVG.)
	1.4495	-0.03615556	-1.44944444	0.036122222

Next, we use the new centers; C_1 = (1.4495, -0.036155556), C_2 = (-1.449444444, 0.036122222)

X	Y	D1	D2	cluster	K1_X	K1_Y	K2_X	K2_Y	Node
-3.0892	-0.107	4.539253	1.64599	K2			-3.0892	-0.107	1
-3.0892	-0.107	4.539253	1.64599	K2			-3.0892	-0.107	2
-2.1847	-1.2306	3.825455	1.464645	K2			-2.1847	-1.2306	3
-2.0788	0.033	3.528978	0.629363	K2			-2.0788	0.033	4
-1.1662	-0.744	2.709784	0.829951	K2			-1.1662	-0.744	5
-1.0089	0.2235	2.472074	0.478738	K2			-1.0089	0.2235	6
-0.2155	-1.1164	1.98473	1.688469	K2			-0.2155	-1.1164	7
-0.1404	1.1092	1.959495	1.692659	K2			-0.1404	1.1092	8
1.0339	0.8708	0.997643	2.619864	K1	1.0339	0.8708			9
0.5956	1.5764	1.824686	2.560207	K1	0.5956	1.5764			10
0.5956	1.5764	1.824686	2.560207	K1	0.5956	1.5764			11
-0.0721	2.2644	2.758228	2.619599	K2			-0.0721	2.2644	12
0.8412	-1.5493	1.630839	2.785788	K1	0.8412	-1.5493			13
1.8008	-0.0767	0.353632	3.252202	K1	1.8008	-0.0767			14
1.6979	-1.0586	1.052186	3.332296	K1	1.6979	-1.0586			15
1.8501	-0.1657	0.421025	3.305711	K1	1.8501	-0.1657			16
1.7295	-1.0861	1.086639	3.371212	K1	1.7295	-1.0861			17
2.9009	-0.4126	1.499424	4.373425	K1	2.9009	-0.4126			18

At this time, we can get the new centers =

New	K1_X (AVG.)	K1_Y(AVG.)	K2_X(AVG.)	K2_(AVG.)
	1.4495	-0.03615556	-1.44944444	0.036122222

We see that the new centers = the old centers. That means, we get the final result. So, we can separate all of nodes to two groups. First group is {9,10,11,13,14,15,16,17,18} and Second is {1,2,3,4,5,6,7,8,12}.