

ISSS606

Social Analytics & Applications

Review

(with sample paper solution)

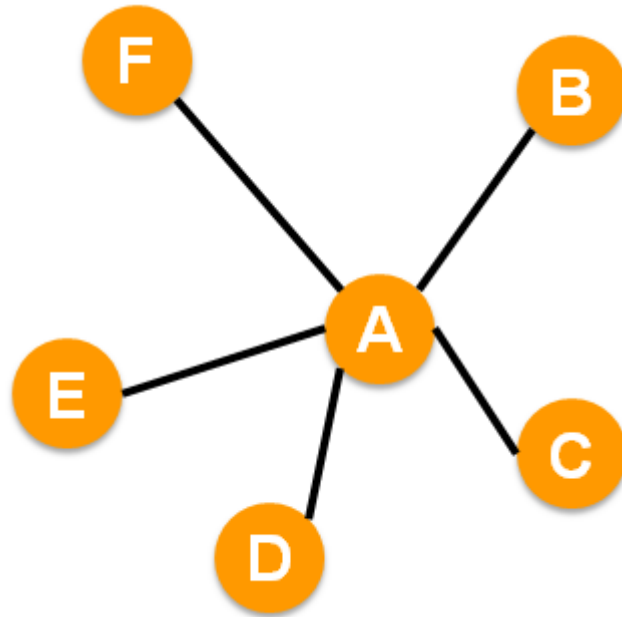
Instructor:
Asst/Prof FANG Yuan

Graphs – basic concepts

- (Un)weighted, (Un)directed graphs
- Complete graphs
- Adjacency matrix/lists
- Bipartite graphs
- Heterogeneous information networks
- Ego-networks
- ...

Graphs – basic concepts

Q1(i)

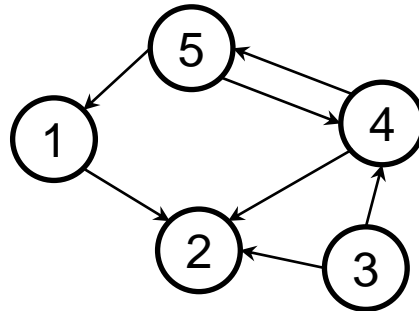


Could this graph be

- a complete graph? False
- an ego-network of node A? True
- a bipartite graph? True

Graphs – basic concepts

Q2(i). Consider the following directed graph. In the adjacency matrix of this graph, how many elements are zeros?



Total number of elements: $5 \times 5 = 25$

Number of non-zeros (edges): 7

Number of zeros: $25 - 7 = 18$

Graph algorithms

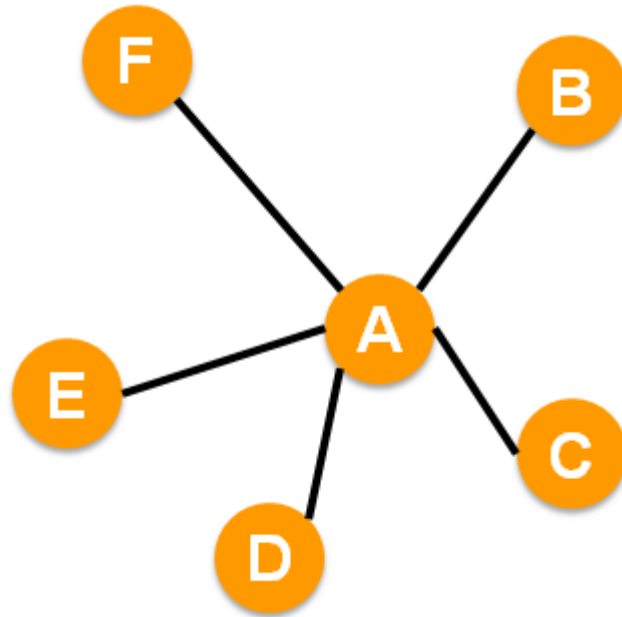
- Graph traversal
 - DFS
 - BFS
 - Key difference?
- Shortest path

Key difference is BFS transverse out to the nearest neighbours first. while DFS can travel to any target node first via a chosen route, when a there is no longer a node that can be visited, DFS backtracks to a non-visited node.

BFS also cannot work on weighted graph while DFS can. This is because when using BFS, the first transversal to the target node may not guarantee it is the shortest path.

Graph algorithms

Q1(ii)



On this graph, is the node sequence "A B C D E F" a valid

- DFS? True
- BFS? True

Graph algorithms

Q4(i). John met Jay at a party, but they forgot to exchange any contact detail before they parted. Now John wishes to find Jay through a social network platform where both of them are active. Describe a method that John can systematically locate Jay on this social network.

Use BFS. John starts with his own account, first examine his friends. If none of his friends is Jay, he goes to the next level, and examines the friends of each of his friends sequentially. He does this level by level until he found Jay.

Alternatively, use DFS, if you have some reasonable guess who of your friends are also friends with Jay.

Network centrality

- Various centrality measures
 - Degree # of edge of node / # of node in network. Pros: able to be computed quickly, can find nodes that can directly influence most other nodes Cons: only considers direct influence on the ego-network
 - Betweenness Formula and table of shortest path. Pros: able to identify which nodes controls the flow of information. Cons: in a small-world network, most nodes can be reached from every other node in small number of steps, closeness all very similar
 - PageRank
 - ...
- Know their pros and cons
- Know their computation (if covered)
 - Check both slides and excel tool

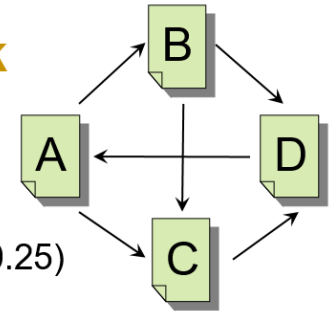
Network centrality

- Computation
 - Starting values?
 - Iterations?
 - Convergence?

Starting value can be anything value for all nodes

Iterations about 30 to convergence


Computation of PageRank



- Iterative updating
- Start with
 $(x_A = 0.25, x_B = 0.25, x_C = 0.25, x_D = 0.25)$
- Next iteration:
 - $x_A = 0.15 / 4 + 0.85 * 0.25 = 0.250$
 - $x_B = 0.15 / 4 + 0.85 * (0.25 / 2) = 0.144$
 - $x_C = 0.15 / 4 + 0.85 * (0.25 / 2 + 0.25 / 2) = 0.250$
 - $x_D = 0.15 / 4 + 0.85 * (0.25 + 0.25 / 2) = 0.356$
- Repeat above until convergence
 $(x_A = 0.297, x_B = 0.164, x_C = 0.233, x_D = 0.306)$

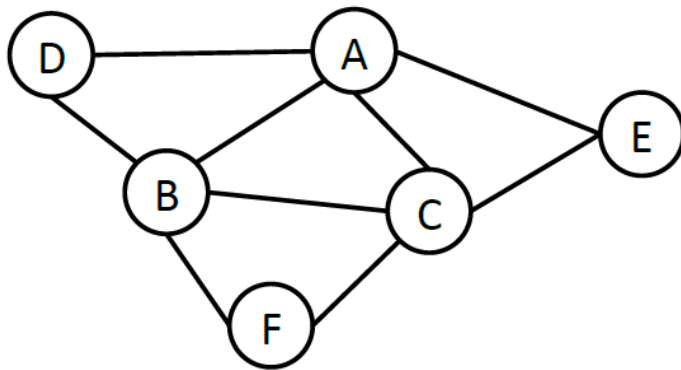
Network centrality

Q1(iv). SimRank vs. PageRank: which is faster to compute on the same graph, assuming the same number of iterations until convergence?

- A. SimRank
-  B. PageRank
- C. Similar
- D. It depends

Network centrality

Q2(iv). In the following graph, suppose the PageRank of node A is q , where $0 < q < 1$. What is the PageRank of node F? Your answer should be expressed in terms of q .



A, B, C are structurally equivalent, so they should have the same PageRank score, all being q .

D, E, F are also structurally equivalent, and should have the same score. Let's say the score is x . Then, $3x + 3q = 1$, since all PageRank scores add up to 1.

$$\therefore x = \frac{1-3q}{3}$$

Community detection

- Concepts
 - Clustering co-efficient Always between 0 and 1
- Hierarchical algorithms
 - Two key questions
 - Modularity-based Initial state is singletons, merge to form communities
 - Centrality-based Choose edge that has high betweenness and remove it

Key question 1 : how to merge or split, each merge or split must be aligned with objectives of community detection, increase density within community edges, and decrease density of between community edges

Key question 2: How to stop? if split too much, all become singletons, if merge too much, it becomes a big community.

Community detection

Q1(viii). What is the difference between clustering co-efficient and modularity?

- A. Clustering co-efficient is more effective in most scenarios
- B. Modularity is more effective in most scenarios
- ✓ C. Clustering co-efficient is an inherent property of a graph
- D. Modularity is an inherent property of a graph.

Clustering co-efficient is inherent property because it depends on how many triangles the graph has

Modularity score depends on the community assignment, which is an input variable. It is not an inherent property of the graph

Community detection

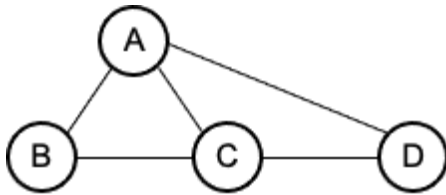
Q2(iii). Consider an undirected and unweighted graph with 250 edges. Suppose X and Y are two nodes from the graph. What is the expected number of edges formed between X and Y, given that node X has a degree of 5 and node Y has a degree of 30?

$$m = 250 \text{ (number of edges)}$$

$$\text{Expected number of edges} = \frac{\deg X}{2m} \times \frac{\deg Y}{2m} \times 2m = \frac{5 \times 30}{500} = 0.3$$

Community detection

Q3. You are given the following graph. In one step of the modularity-based agglomerative algorithm, let us assume $\{A,C\}$ and $\{B,D\}$ are detected as two communities.

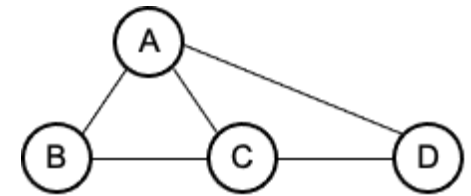


- (i) Calculate the value of modularity Q of the graph under this situation.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(s_i, s_j)$$

- (ii) Is the community structure given possible in any step of the algorithm?

Community detection



$$m = 5$$

Q3(i)

A matrix:

	A	B	C	D
A	0	1	1	1
B	1	0	1	0
C	1	1	0	1
D	1	0	1	0

E matrix:

	A	B	C	D
A	9/10	6/10	9/10	6/10
B	6/10	4/10	6/10	4/10
C	9/10	6/10	9/10	6/10
D	6/10	4/10	6/10	4/10

B matrix:

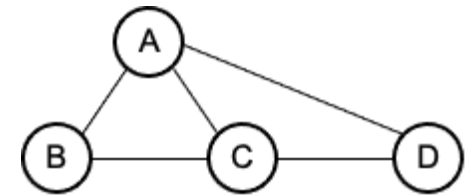
	A	B	C	D
A	-9/10	4/10	1/10	4/10
B	4/10	-4/10	4/10	-4/10
C	1/10	4/10	-9/10	4/10
D	4/10	-4/10	4/10	-4/10

Given communities {A, C}, {B, D}, delta matrix:

	A	B	C	D
A	1	0	1	0
B	0	1	0	1
C	1	0	1	0
D	0	1	0	1

$$Q = \left(-\frac{9}{10} + \frac{1}{10} - \frac{4}{10} - \frac{4}{10} + \frac{1}{10} - \frac{9}{10} - \frac{4}{10} - \frac{4}{10} \right) / 10 = -\frac{32}{100}$$

Community detection



$$m = 5$$

Q3(ii)

A matrix:

	A	B	C	D
A	0	1	1	1
B	1	0	1	0
C	1	1	0	1
D	1	0	1	0

E matrix:

	A	B	C	D
A	9/10	6/10	9/10	6/10
B	6/10	4/10	6/10	4/10
C	9/10	6/10	9/10	6/10
D	6/10	4/10	6/10	4/10

B matrix:

	A	B	C	D
A	-9/10	4/10	1/10	4/10
B	4/10	-4/10	4/10	-4/10
C	1/10	4/10	-9/10	4/10
D	4/10	-4/10	4/10	-4/10

Given initial singletons, delta matrix:

	A	B	C	D
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

$$Q = \left(-\frac{9}{10} - \frac{4}{10} - \frac{9}{10} - \frac{4}{10} \right) / 10 = -\frac{26}{100}$$


It's not possible since the modularity of {A, C}, {B, D} is smaller than the initial state where all communities are singletons.

Network-based similarity

- Methods
 - Jaccard Similarity
 - Adamic-Adar
 - SimRank
- Understand their intuition
- Understand their computation
 - Check slides, lab solution

Network-based similarity

Q1(vi). Suppose a graph contains n nodes v_1, v_2, \dots, v_n . Which of the following could be a valid SimRank score if the damping factor is set to $C = 0.85$?

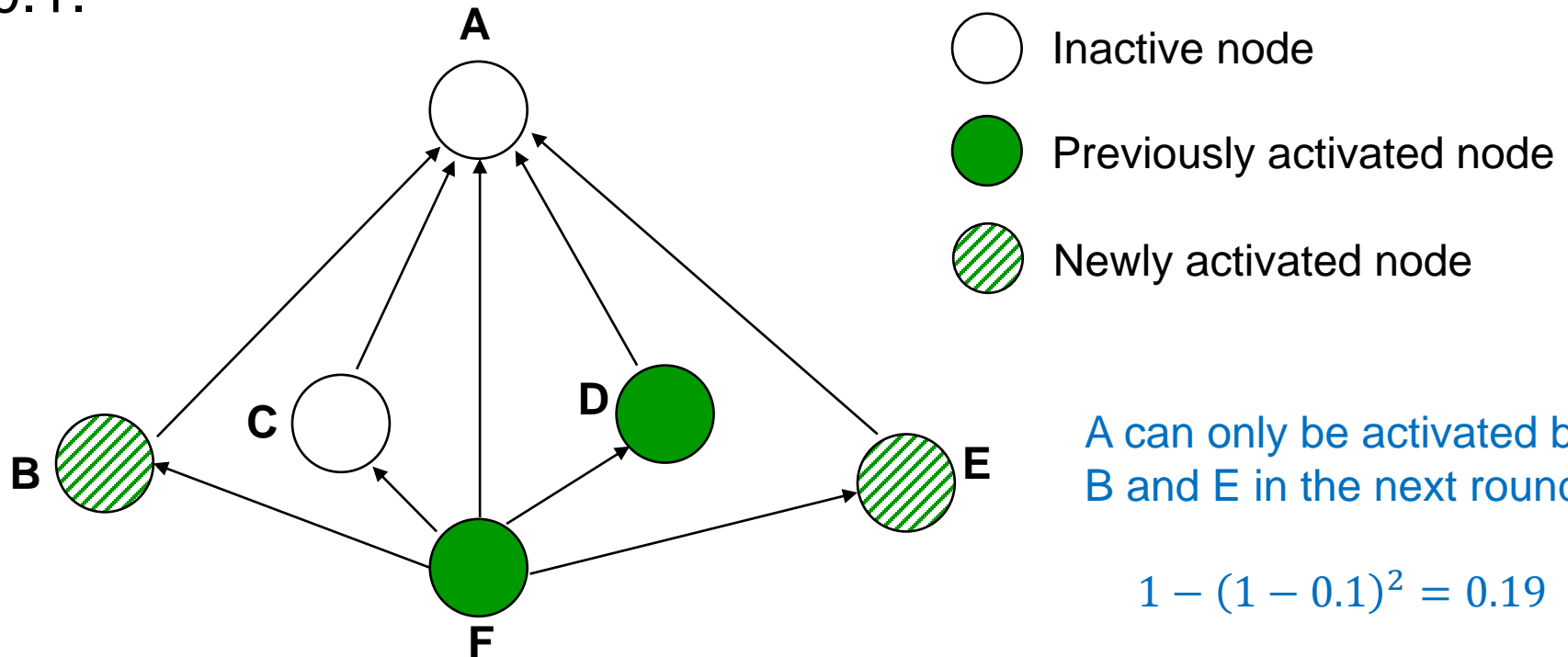
- A. $\text{SimRank}(v_2, v_2) = 0.99$
- B. $\text{SimRank}(v_1, v_2) = 0.9$
- C. $\text{SimRank}(v_2, v_3) = 1$
-  D. $\text{SimRank}(v_2, v_5) = 0.8$

Influence diffusion

- Seeds
- Activation path
- Model: Independent cascade

Influence diffusion


Q2(ii). Compute the probability of A being activated in the next round of independent cascade, as shown in the diagram below. Suppose the probability of activation on each edge is 0.1.



$$1 - (1 - 0.1)^2 = 0.19$$

Influence diffusion

Q1(v). You are tasked with promoting a product on a social network. Which one of the following techniques is relevant?

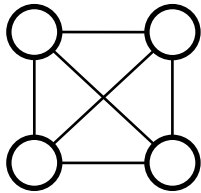
- A. Network centrality
- B. Information diffusion
- C. Sentiment analysis
-  D. All of the above

Network evolution

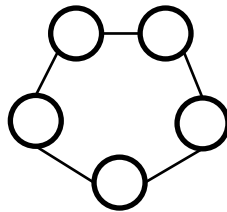
- Homophily
 - Existing nodes to form new links
- Preferential attachment
 - For new nodes
 - Computation

Network evolution

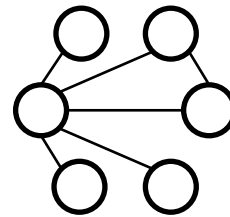
Q1(vii). Which one of the following networks is the most likely outcome of the preferential attachment model?



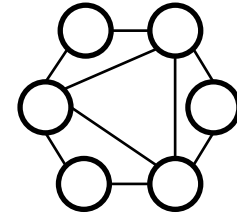
A



B



C



D

Text analytics

- Basic topics
 - Vector space model / bag-of-words
 - Document similarity
 - Sentiment analysis
 - Topic modeling
- Word embedding
 - Word vectors
 - Skip-gram model
 - Applications

Text analytics

Q1(iii). Consider the two documents below.

D1: "The dog is chasing the cat outside."

D2: "It is raining cats and dogs outside."

What is the cosine similarity of D1 and D2 if stemming is performed / not performed? In your calculation, ignore the following four stop words: "the", "is", "it", "and".

Since stop words are ignored, each document has four unique words.

If stemming is performed, there are three common words (dog/cat/outside).

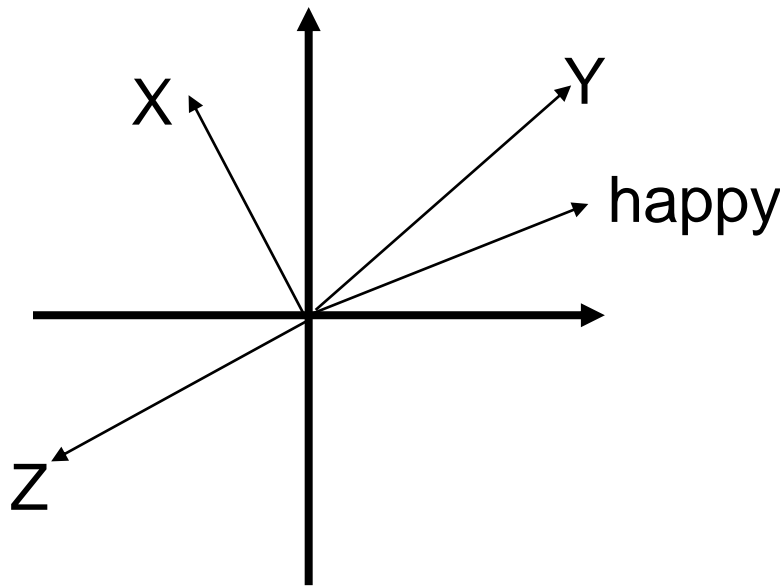
$$\text{Cosine similarity: } \frac{3}{\sqrt{4}\sqrt{4}} = 0.75$$

If stemming is not performed, there is only one common word (outside).

$$\text{Cosine similarity: } \frac{1}{\sqrt{4}\sqrt{4}} = 0.25$$

Text analytics

Q2(v) In the diagram below, there are four word vectors based on word embedding, to represent four words “angry”, “happy”, “pleased” and “neutral.” One of the vectors is known to represent the word “happy” as labelled. What words do the other three vectors labelled X, Y and Z represent?



X: Neutral
Y: Pleased
Z: Angry

Advanced network topics

- Network classification
 - Different with community detection?
 - Basic concepts of supervised learning
 - Evaluation metrics
 - Various versions of label averaging algorithms
- Network embedding
 - Resemblance to word embedding
 - How it can be built upon word embedding?
 - Applications?

Advanced network topics

Q4(ii). Consider an MITB social network where each student may list his/her favourite MITB course. Some users did not list their favourite course. For those missing favourite subjects, suggest *two* methods to predict if they would be “ISSS606 Social Analytics & Applications” or not.

- Network-based classification. Treat ISSS606 as labels; students with ISSS606 as favourite are labelled 1; those with other favourite courses are 0; students without listing their favourite are unlabeled nodes. We can perform either Majority Voting and Label Averaging to predict for the unlabeled nodes.
- Network embedding. After learning the node vectors using DeepWalk, we can find the cosine similarity of two students, and assign the same favourite based on the most similar student. (Or perform a traditional classification using the node vectors as features.)

Advanced network topics

Q4(ii). Consider an MITB social network where each student may list his/her favourite MITB course. Some users did not list their favourite course. For those missing favourite subjects, suggest *two* methods to predict if they would be “ISSS606 Social Analytics & Applications” or not.

- Community detection. Find communities on the network first. Then infer the missing favourite course based on the most popular one in the community.
- Network-based similarity. Find the Jaccard/AA/SimRank between each pair of students. For students with missing favourite course, find their most similar node on the network, and assign the same favourite course to them.

QUESTIONS?!

QUESTIONS?!