Internet Analytics (COM-308): Midterm Exam

April 24, 2017

Duration: **1h45**.

Total points: **100**.

Number of pages: **11**.

Allowed documents: **class notes, lab handouts, homeworks, your own code**.

There should in general be enough room below every question for intermediate calculations and your answer. However, you are allowed to use additional sheets of paper; please **write your name on every sheet**, number them, and staple them to this document before handing in.

The use of **mobile phones, tablets, laptop computers**, and other communication devices is **prohibited**.

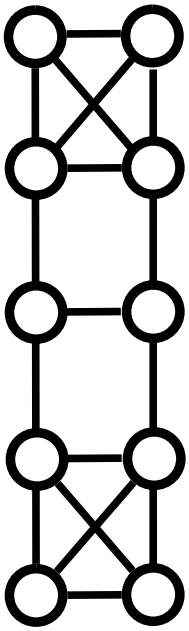
Name:
First name:
SCIPER number:
Signature:

Please leave blank.

1	2	3	4	5	Total
15	20	25	20	20	100

Question 1: Strong Triadic Closure (15 points)

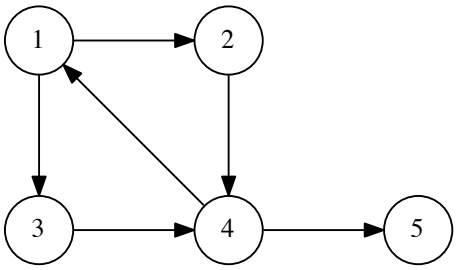
- (15 pts) Find the largest possible set of strong links such that the strong triadic closure (STC) property holds for all nodes.



Question 2: Page-Rank (20 points)

Notations: Let n denote the number of nodes, and o_u denote the out-degree of a node u.

1. (10 pts) **Linear algebra**



Consider the graph shown above.

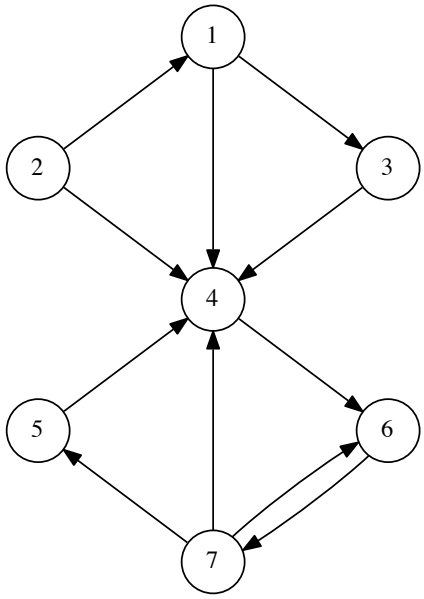
(a) Give the score-flow matrix H, defined as

$$H_{u,v} = \begin{cases} 1/o_u, & (u,v) \in E \\ 0, & \text{otherwise.} \end{cases}$$

(b) Give the Google matrix $G = \theta\hat{H} + (1-\theta)\frac{e^T e}{n}$ obtained using $\theta = 2/3$.

(c) Starting with $\pi_0 = [1/5, 1/5, 1/5, 1/5, 1/5]$, give the PageRank estimate π_1 obtained after one power iteration.

2. (5 pts) **Rank reversal**



Consider the graph shown on the left.

(a) Order the nodes by decreasing PageRank (ties are allowed), using

(i) $\theta = 0.999$.

(ii) $\theta = 0.001$.

Justify your answers in a few words (computations are not necessary).

(b) For $\theta = 1$, what is the exact PageRank of node ②?

3. (5 pts) Consider the same graph. You know that Google uses $\theta = 0.999$ and you want to maximize the PageRank of node ⑤ by rewiring a single edge. Which edge would you rewire if you are only allowed to

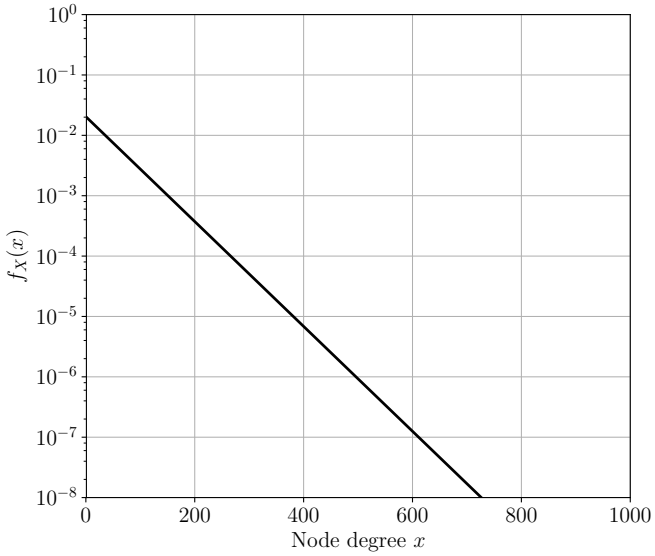
(a) change the destination of the edge.

(b) change the source of the edge.

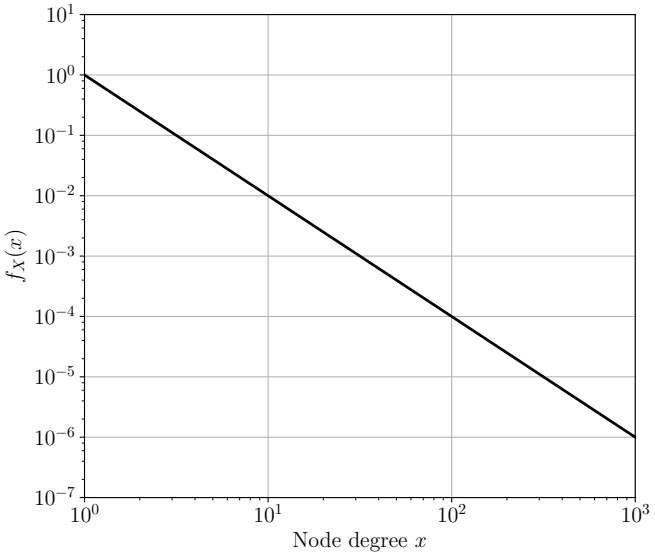
Justify your answers in a few words (computations are not necessary).

Question 3: Social Networks (25 points)

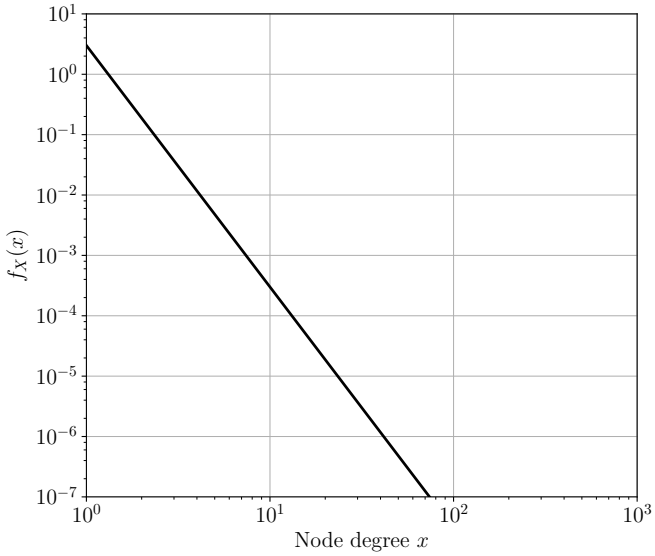
The following plots represent the distribution of node degree in four different social networks[1].



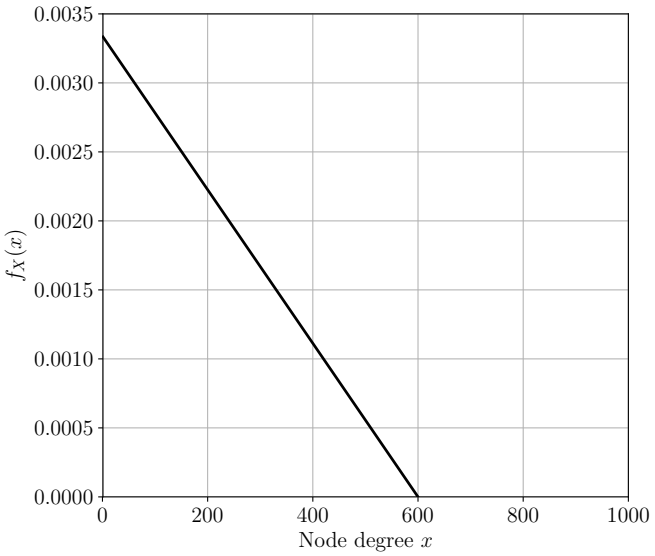
(a)



(b)



(c)



(d)

1. (5 pts) Indicate, for each of the networks, whether the distribution follows a power law or not. For each network, justify your answer in a few words.

[1]For simplicity, we approximate the degree distribution with a continuous distribution, even though the degrees actually take discrete values.

2. (10 pts) Let γ denote the exponent of a power-law distribution, i.e., such that $f_X(x) \propto x^{-(\gamma+1)}$. For each power-law distribution that you identified, estimate the parameter γ.

3. (5 pts) Let A and B be two social networks with power-law degree distributions such that $\gamma_A < \gamma_B$. Suppose that you use the following method to compute the average degree:

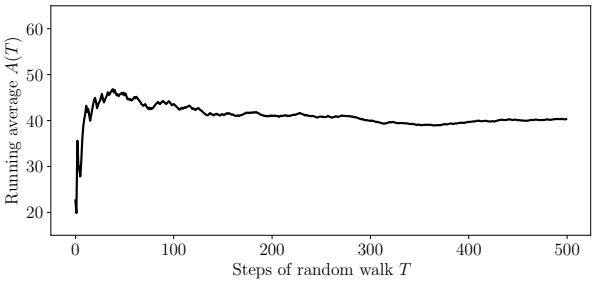
   - Starting from seed node, take a random walk of length T, resulting in a sequence of nodes $v_1, \ldots, v_T$.
   - Estimate the average degree as $M = \frac{1}{T} \sum_{i=1}^{T} \deg(v_i)$

   Which of the two networks, A or B, will result in an estimate that is *most biased*? Explain briefly.

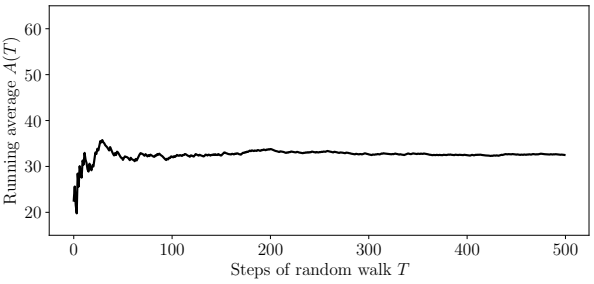
4. (5 pts) Consider two networks that both contain roughly 50% of *old* users, whose average age is around 60, and 50% of *young* users, whose average age is around 20. The two plots below show the running estimate of average age for both networks, obtained from a random walk $v_1, v_2, \ldots$ as follows:

$$A(T) = \frac{1}{T} \sum_{i=1}^{T} \text{age}(v_i)$$

For each plot, explain what you observe and describe possible causes for the behavior (e.g., in terms of properties of the network).



(a)



(b)

Question 4: Dimensionality Reduction (20 points)

We generate a set of points $X_1, \ldots, X_n$ i.i.d. according to a Gaussian distribution $N(0, \Sigma)$, where Σ is the covariance matrix, and we perform PCA on this set of points in order to visualize them in two dimensions.

More precisely, we project every point X_i onto the first and second principal component, i.e., the eigenvectors of the empirical covariance matrix associated with the largest and second-largest eigenvalue.

You are given the following five 2-d plots as possible results.

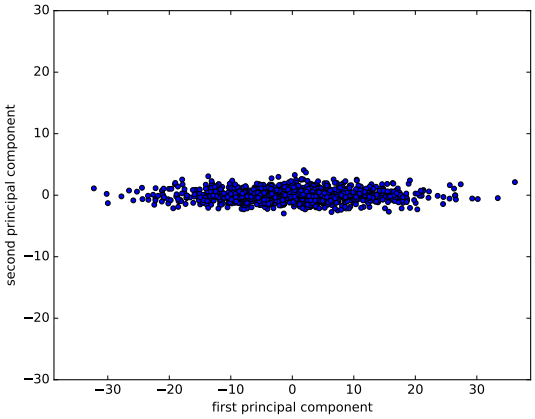
1. (10 pts) Which of the 6 plots are guaranteed to never result from PCA on the above distribution, and which are plausible? Explain your reasoning.

2. (10 pts) For the following covariance matrices, state which plot would result (among the plausible plots you identified above). Briefly justify your answers, without necessarily resorting to any calculations.

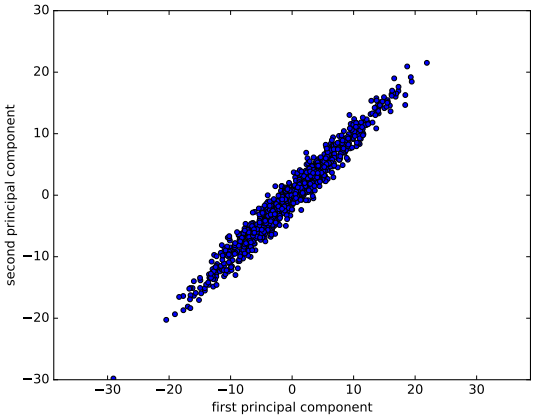
$$\Sigma_1 = \begin{bmatrix} 100 & 0 \\ 0 & 30 \end{bmatrix} \tag{1}$$

$$\Sigma_2 = \begin{bmatrix} 50 & 49 \\ 49 & 50 \end{bmatrix} \tag{2}$$

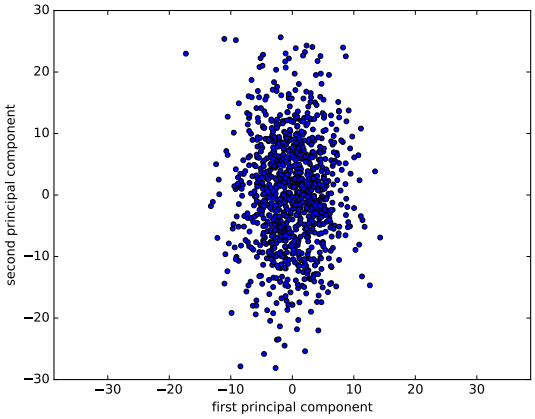
$$\Sigma_3 = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 100 \end{bmatrix} \tag{3}$$


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

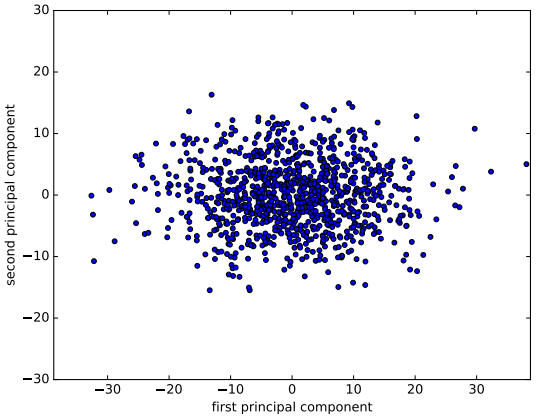
(a)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

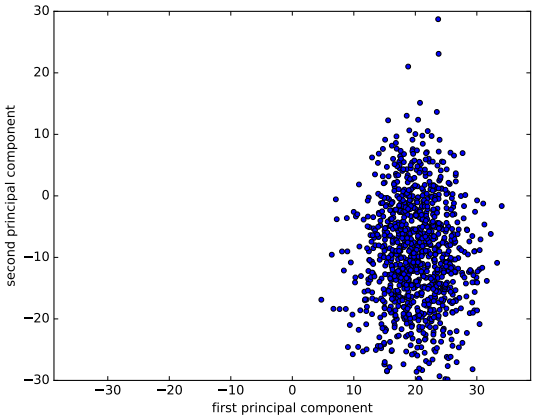
(b)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

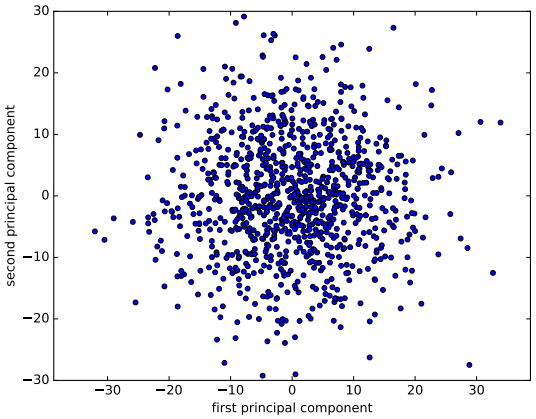
(c)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

(d)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

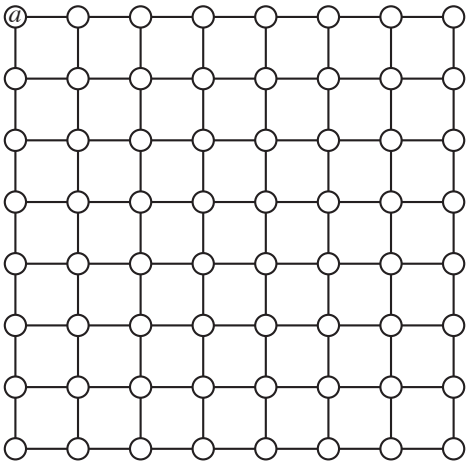
(e)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

(f)

Question 5: Conductance and Mixing (20 points)

Consider the chessboard graph G, with 8×8 vertices in a square configuration, and every node is connected to its north, south, east and west immediate neighbor (if it exists).



1. (5 pts) Determine the stationary distribution π of a random walk on G.

2. (10 pts) Compute the conductance of this graph.

3. (5 pts) Let $p_{ai}(t)$ be the probability that a random walk starting at a reaches i after t steps. What is the smallest t such that $|p_{ai}(t) - \pi_i| \leq 10^{-6}$ for all i? Circle the correct answer and show your calculations.

   - 1000
   - 5000
   - 10000
   - 50000

   Hint: you might find the following inequality useful $(1-x)^n \leq \frac{1}{1+nx}$.