COM 530500
Network Science Final Project

December 24, 2020

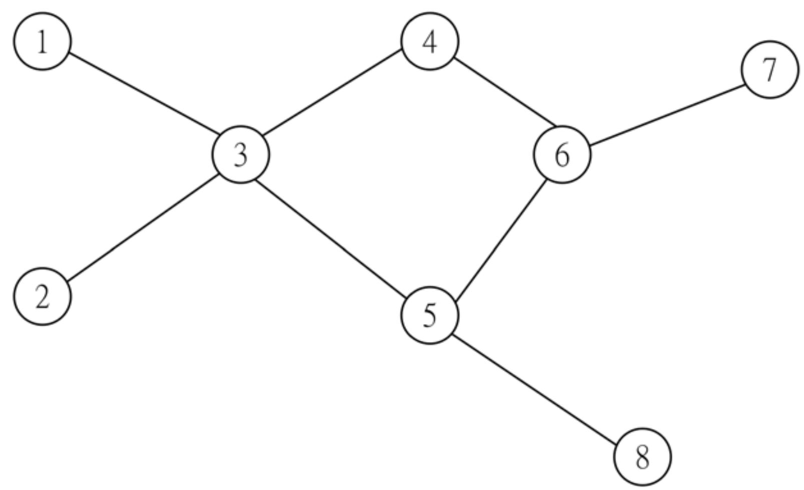
Institute of Communications Engineering
National Tsing Hua University

Important Dates (1/1)

- Final Exam: Jan. 14, 2021
- Project Due: Jan. 21, 2021

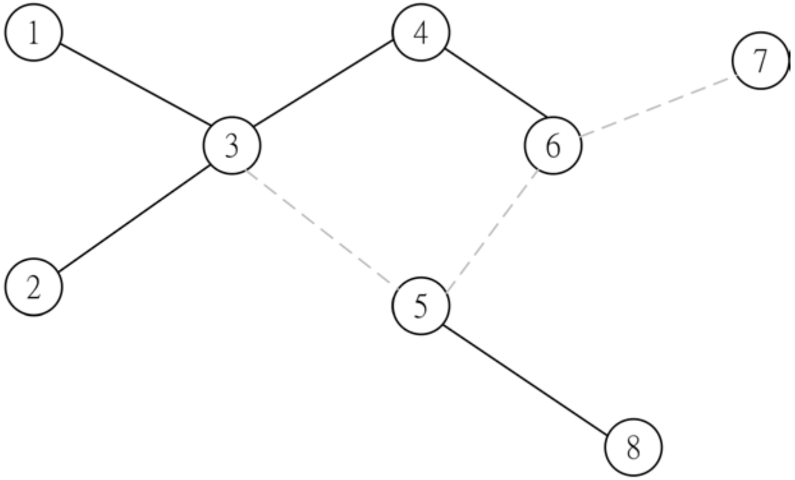
Independent Cascade Model (1/7)

Given: adjacency matrix A, transition probability ϕ, a set of seed nodes S, and distance D.



Independent Cascade Model (2/7)

Step 1: Obtain $\tilde{A}$ by removing edges of A, where each edge is removed with probability $1 - \phi$.

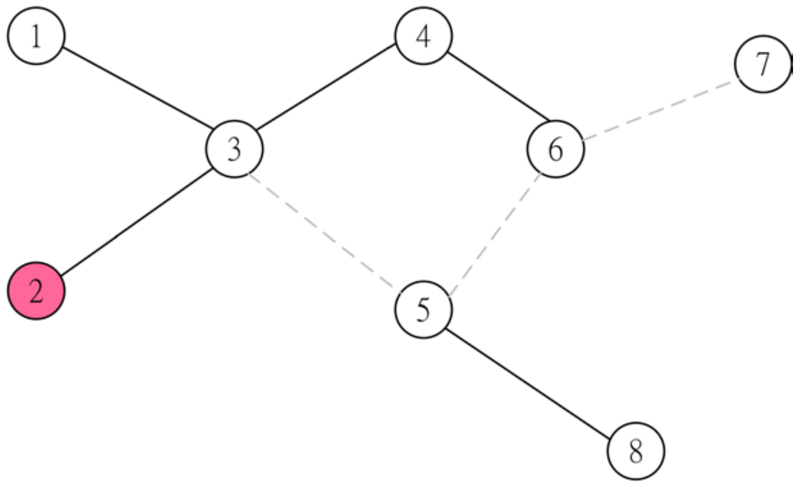


Independent Cascade Model (3/7)

Step 2: Define the $n \times 1$ seed vector x by

$$x_i = \begin{cases} 1, & \text{for } i \in S \\ 0, & \text{for } i \notin S \end{cases}$$

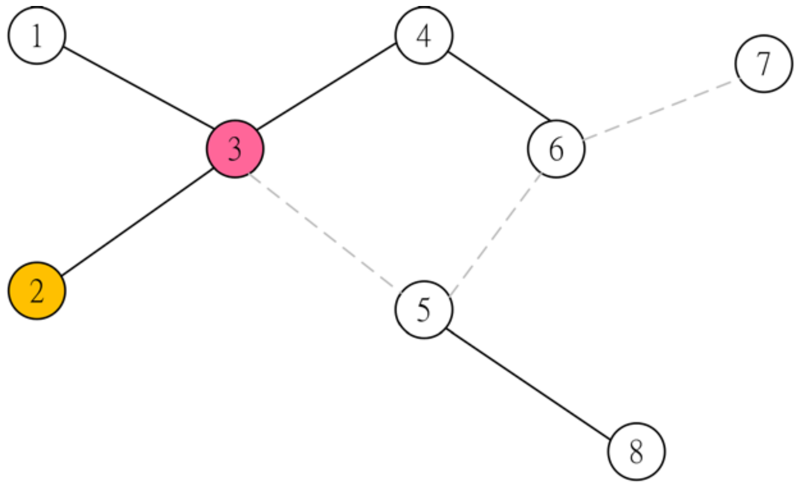
$$S = \{2\}$$



$$x = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Independent Cascade Model (4/7)

Step 3: Calculate $y = (\tilde{A} + I)^D x$ with the AND and the OR operations.

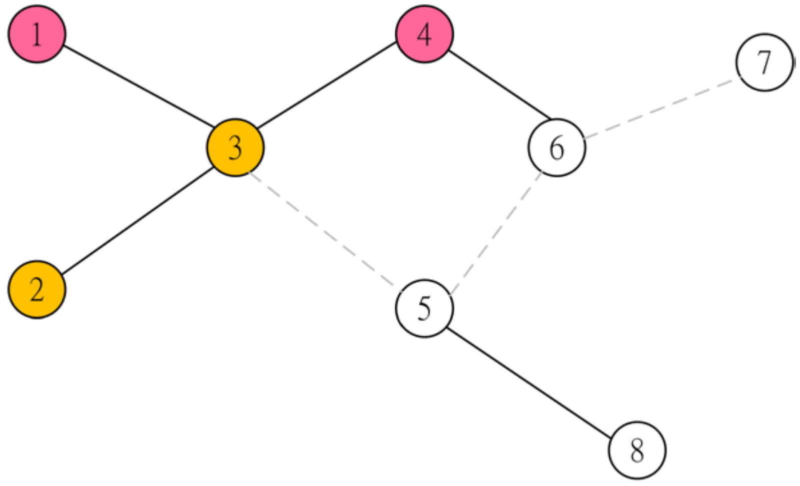


$D = 1$

$$y = (\tilde{A} + I)x = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Independent Cascade Model (5/7)

Step 3: Calculate $y = (\tilde{A} + I)^D x$ with the AND and the OR operations.

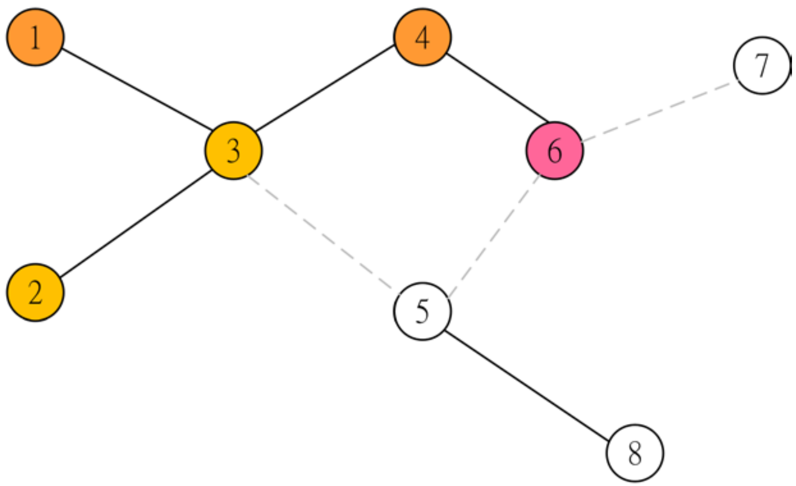


$$D = 2$$

$$y = (\tilde{A} + I)^2 x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Independent Cascade Model (6/7)

Step 3: Calculate $y = (\tilde{A} + I)^D x$ with the AND and the OR operations.

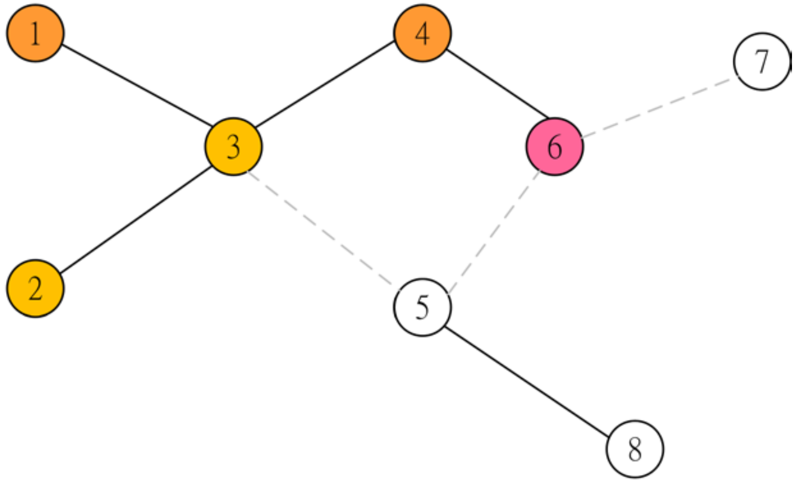


$D = 3$

$$y = (\tilde{A} + I)^3 x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Independent Cascade Model (7/7)

Step 4: Obtain the number of infected nodes by counting the number of 1's in vector y.



$$D = 3$$

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \|y\|_1 = 5$$

Project Requirements (1/3)

Problem 1. **(40%)** Consider the **ego-Facebook** dataset. A node in this dataset represents a user on Facebook, and an edge between two nodes represents the relationship between two users.

(a) (10%) List some statistical information of this dataset, such as the number of nodes, number of edges, average clustering coefficient, diameter, average degree, maximum degree, etc.

(b) (10%) Visualize the dataset by plotting it.

(c) (10%) Plot the degree distribution with log-log scale.

(d) (10%) List the top 10 nodes ranked by the following centrality measures.

- Degree centrality
- Katz centrality
- Eigenvector centrality
- Betweenness centrality
- Closeness centrality

Project Requirements (2/3)

Problem 2. (60%+bonus 10%) In this problem, we want to investigate the disease propagation by the independent cascade (IC) model in **ego-Facebook** [**?**] dataset. Assume the propagation probability is ϕ, and the set of seeds nodes S are randomly selected. Collect the set of infected nodes within the distance D of the seed nodes, and calculate the prevalence rate r_1 (which is defined by the ratio of the number of infected nodes to the total number of nodes). Set $\phi = 0.1, |S| = 5$, and D the diameter of the graph.

Project Requirements (3/3)

(a) (40%) Simulate the disease propagation by IC model after removing the top 0%, 10%, 20%, …, 50% of nodes from the following centrality measures respectively, and calculate the corresponding prevalence rate r_1. Please plot the curves of r_1 vs. the percentage of nodes removed. (*Note: Please run the simulation 100 times and average the results.*)

- Degree centrality
- Katz centrality
- Eigenvector centrality
- Betweenness centrality
- Closeness centrality

(b) (bonus 10%) Could you find a centrality measure that achieves a better performance?

(c) (20%) Write a report to compare and discuss the results of different centrality measures.

Summary (1/1)

Given: adjacency matrix A, transition probability ϕ, a set of seed nodes S, and distance D.

Step 1: Obtain A' by removing edges of A, where each edge is removed with probability $1 - \phi$.

Step *: **Obtain $\tilde{A}$ by removing some specific nodes and all their edges from A'.**

Step 2: Define the $n \times 1$ seed vector x by $x_i = 1$ for $i \in S$ and $x_i = 0$ for $i \notin S$.

Step 3: Calculate $y = (\tilde{A} + I)^D x$ with the AND and the OR operations.

Step 4: Obtain the number of infected nodes by counting the number of 1's in vector y.

Step 5: Repeat **Step 1** to **Step 5** for 100 times, and average the results.

Step 6: Calculate the prevalence rate r_1.