Name:	Jackson Wakefield
ASU ID Number:	1216414420

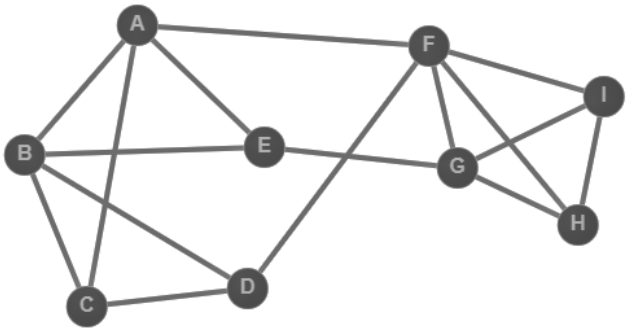
CSE 472: Social Media Mining

Homework III - Information Diffusion, Community Analysis

Prof. Huan Liu
Due at 2021 Nov 02, 11:59 PM

This is an ***individual*** homework assignment. Please submit a digital copy of this homework to **Gradescope**. For your solutions, even when not explicitly asked you are supposed to concisely justify your answers.

1. [**Community Analysis**] For the given Graph and k=3, Using Clique Percolation Method (CPM) what are the detected communities?



Cliques:
ABC, ABE, BCD, FGH, FHI, FGI, GHI

Communities:
ABC --- ABE
|
BCD

FGH --- FHI
| X |
FGI --- GHI

(In the second community, the X represents the connections between FGH->GHI and FHI ->FGI)

2. [**Community Analysis**] Compute the following metrics for the given figure:

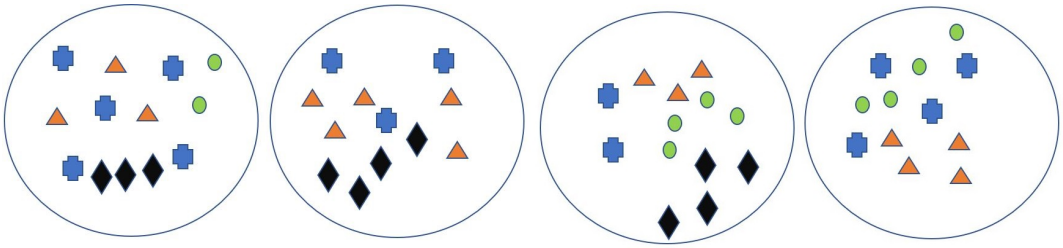


Figure 1: The communities.

- Precision and recall

TP = (2 c 2) + (3 c 2) + (3 c 2) + (5 c 2) + (5 c 2) + (4 c 2) + (3 c 2) + (4 c 2) + (3 c 2) + (4 c 2) + (2 c 2) + (4 c 2) + (4 c 2) + (4 c 2) = 70

FP = (2 * 3 * 3 * 5) + (4 * 5 * 3) + (4 * 3 * 4 * 2) + (4 * 4 * 4) = 310

FN = (2 * 4 * 4) + (3 * 5 * 3 * 4) + (3 * 4 * 4) + (5 * 3 * 2 * 4) = 380

Precision = 70/(70+310) = 0.1842

Recall = 70/(70+380) = 0.1556

- F-measure

F = 2 * (P * R)/(P + R) = 2 * (0.1842 * 0.1556)/(0.1842 + 0.1556) = 0.1687

- NMI

GroundTruth ->
Community(A) = Cross
Community(B) = Triangle
Community(C) = Diamond
Community(D) = Circle

N = 50

N(h): A-13, B-12, C-13,D-12
N(l): A-14, B-15, C-11, D-10

N(h, l):

	A	B	C	D
A	5	3	3	2
B	3	5	4	0
C	3	3	4	4
D	4	4	0	4

(This is gonna be a doozy)

Numerator: (5*log(50*5)/(13*14) + 3*log(50*3)/(13*15) + 3*log(50*3)/(13*11) + 2*log(50*2)/(13*10) + 3*log(50*3)/(12*14) + 5*log(50*5)/(12*15) + 4*log(50*4)/(12*11) + 2*log(50*2)/(13*14) + 3*log(50*3)/(13*15) + 4*log(50*4)/(13*11) + 4*log(50*4)/(13*10) + 4*log(50*4)/(12*14) + 4*log(50*4)/(12*15) + 4*log(50*4)/(12*10)) = 0.7242

(took out the two log(0/X)'s ^^^)

Denominator: sqrt((13 * log(13/50) + 12 * log(12/50) + 13 * log(13/50) + 12 * log(12/50)) * (14 * log(14/50) + 15 * log(15/50) + 11 * log(11/50) + 10 * log(10/50))) = 29.9455

NMI = 0.0241845016

- Purity

P = (Maj(C1) + Maj(C2) + ... Maj(Cn))/(total nodes) = (5 + 5 + 4 + 4) / 50 = 18/50 = 0.36

3. [**Information Diffusion**] Given the following graph, the weight of the edges denote the $p_{v,w}$ which is the probability of v activating w and vice-versa (for an undirected graph). Follow the ICM procedure until it converges for the following graph such that node A is activated at time 0. Mention all the nodes that are activated by the end. Consider the threshold value to be 0.3.

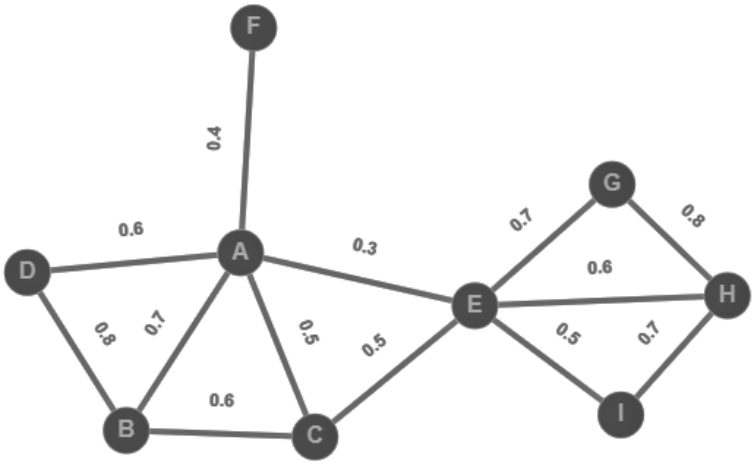


Figure 2: Independent Cascade Model Graph

Timestamp	Activated Nodes
0	A
1	A, F, D, B, C
2	A, F, D, B, C, E
3	A, F, D, B, C, E, G, H, I

4. [**Information Diffusion**] Does Independent Cascade Model (ICM) converge? Why? When the ICM stops running the algorithm has converged? Please justify your answer with details.

The convergence of the Independent Cascade Model will depend on the values of the probabilities. Thinking logically, there should be relative convergence with values that are greater than or less than 0.5. The convergence of the algorithm will be much more defined with values close to 0, and much moreso with values close to 1 (meaning almost no one or everyone recieves the information). In a graph where every probability is 0.5, there will be absolutely no convergence.

When the average probability of activation dips below 0.5, It seems like convergence would be based on each individual graph (maybe a certain part of the graph is more likely to get activated than another, proving some kind of convergence)

When the average probability of activation gets closer to 1.0, the ICM algorithm will almost certainly converge with an exponentially larger (depending on how close the average probability is to 1) set of activated nodes.