Name:	Mark Gameng
IIT ID Number:	A20419026

CS 579: Online Social Network Analysis

Homework III - Information Diffusion, Community Analysis, Influence and Homophily, Recommendation

Prof. Kai Shu
Due at April 11 2022, 11:59 PM

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

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

ICM does converge because each node has one chance to activate its neighbors.
The algorithm goes like this, node activated at time t has one chance at time t+1 to activate its neighbors. If a node gets activated at time t, then for any of its neighbors, theres a probability that they get activated at time t+1. As such, theres only a finite number of time steps. Either all the nodes eventually becomes activated, or there is no longer a chance of activation due to failing to activate their neighbors.
When ICM stops running, the algorithm has converged as it has run until either all the nodes were activated, or there is no longer a chance for a new node to get activated. The result is either, all the nodes are activated, or only some of them are.

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

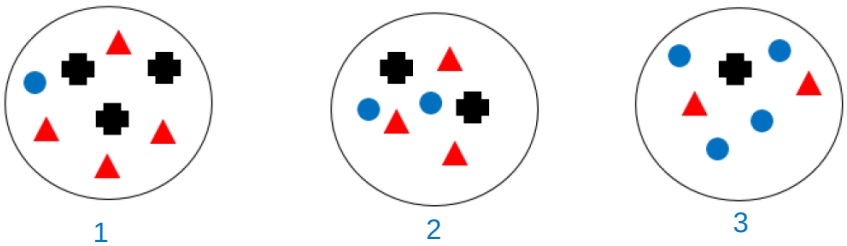


Figure 1: The communities.

Name:	Mark Gameng
IIT ID Number:	A20419026

TP_red = 4c2 + 3c2 + 2c2 = 6 + 3 + 1 = 10
TP_black = 3c2 + 2c2 = 3 + 1 = 4
TP_blue = 2c2 + 4c2 = 1 + 6 = 7

FP_red = (4x4) + (3x4) + (2x5) = 38
FP_black = (3x5) + (2x5) + (1x6) = 31
FP_blue = (1x7) + (2x5) + (4x3) = 29

FN_red = (4x3) + (4x2) + (3x2) = 26
FN_black = (3x2) + (3x1) + (2x1) = 11
FN_blue = (1x2) + (1x4) + (2x4) = 14

TN_red = (4x2) + (4x2) + (4x4) + (4x1) + (3x4) + (3x1) = 51
TN_black = (3x2) + (3x3) + (3x4) + (3x2) + (2x4) + (2x2) = 45
TN_blue = (1x3) + (1x2) + (1x2) + (1x1) + (2x2) + (2x1) = 14

- Precision and recall

5 choose 2 = 5c2

TP = TP_red + TP_black + TP_blue = 10 + 4 + 7 = 21
FP = FP_red + FP_black + FP_blue = 38 + 31 + 29 = 98
FN = FN_red + FN_black + FN_blue = 51
TN = TN_red + TN_black + TN_blue = 110

Precision = TP / (TP + FP) = 21 / (21 + 98) = 21 / 119 = 0.176
Recall = TP / (TP + FN) = 21 / (21 + 51) = 21 / 72 = 0.292

- F-measure

F-measure = 2 * ((P * R) / (P + R))
= 2 * ((0.176 * 0.292) / (0.176 + 0.292))
= 2 * (0.051 / 0.468)
= 0.218

- NMI

P_red = 9 / 22, P_black = 6 / 22, P_blue = 7 / 22
H(L) = -(9/22)log2(1/22) - (6/22)log2(6/22) - (7/22)log2(7/22) = 2.86
P_1 = 8 / 22, P_2 = 7 / 22, P_3 = 7 / 22
H(H) = -(8/22)log2(8/22) - (7/22)log2(7/22) - (7/22)log2(7/22) = 1.58
I(L;H) = H(L) - H(L|H) = 2.86 - (0.46+0.51+0.45) = 1.44

NMI = I(L;H) / ((1/2)(H(L)+H(H))
= 1.44 / ((1/2)(2.86 + 1.58))
= 0.65

P(red|1) = 4/8, P(black|1) = 3/8, P(blue|1) = 1/8
H(L|1) = -(4/8)log2(4/8) - (3/8)log2(3/8) - (1/8)log2(1/8) = 1.40, 1.40/3 = 0.46
P(red|2) = 3/7, P(black|2) = 2/7, P(blue|2) = 2/7
H(L|2) = -(3/7)log2(3/7) - (2/7)log2(2/7) - (2/7)log2(2/7) = 1.55, 1.55/3 = 0.51
P(red|3) = 2/7, P(black|3) = 1/7, P(blue|3) = 4/7
H(L|3) = -(2/7)log2(2/7) - (1/7)log2(1/7) - (4/7)log2(4/7) = 1.37, 1.37/3 = 0.45

- Purity

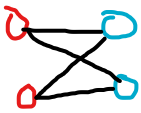
Purity = (4 + 3 + 4) / 22
= 11 / 22
= 0.5

$$P_U = \frac{1}{N} \sum_{i=1}^{k} \max_j |C_i \cap L_j|$$

3. [**Influence and Homophily**] What is the range $[\alpha_1, \alpha_2]$ for modularity Q values? Provide examples for both extreme values of the range, as well as cases where Modularity becomes zero. Modularity is defined as,

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{d_i d_j}{2m}] \delta(c_i, c_j) \quad (1)$$

Name:	Mark Gameng
IIT ID Number:	A20419026



Modularity ranges from [-1/2, 1].

Q = -1/2, Negative means the number of edges between nodes of the same color is less than the expected number of edges between them. Max negative value means max amount of edges between nodes of different colors, and no edges between nodes of same color

Q = 0, exactly as many edges within same colors as we would expect in a null model, or a randomly generated graph

Q = 1, Positive means the number of edges between nodes of the same color is more than the expected number of edges between them. The extreme value of this means that max amount of edges between nodes of same color, and no edges between nodes of different colors

4. [**Recommendation**] Consider the user-item matrix in table 1 and answer the following questions.

	God	Le Cercle Rouge	Cidade de Deu	Rashomon	La vita e bella	$\overline{r_u}$
Newton	3	0	3	2	4	2.4
Einstein	5	4	0	2	3	2.8
Gauss	1	2	4	3	1	2.2
Aristotle	3	? [1]	4	2	2	2.75
Euclid	2	2	0	1	5	2

Table 1: **User-Item Matrix**

(a) Compute the missing rating (Aristole-Le Cercle Rouge) in this table using *user-based* collaborative filtering (CF). Use cosine similarity to find the two nearest neighbors.

sim(ari, newton) = (3*3+4*3+2*2+2*4)/(sqrt(9+16+4+4)*sqrt(9+9+4+16)) = 0.93
sim(ari, einstein) = (3*5+4*0+2*2+2*3)/(sqrt(33)*sqrt(25+4+9)) = 0.70
sim(ari, gauss) = (3*1+4*4+2*3+2*1)/(sqrt(33)*sqrt(1+16+9+1)) = 0.90
sim(ari, euclid) = (3*2+4*0+2*1+2*5)/(sqrt(33)*sqrt(4+1+25)) = 0.572
Newton and Gauss are the two nearest neighbors

r_aristotle-lecerclerogue = 2.75 + ((0.93(0-2.4) + 0.90(2-2.2)) / (0.93 + 0.90))
= 1.43 and then rounded down to 1

(b) Consider group $G = \{Newton, Einstein, Gauss\}$, compute the aggregated ratings for all products using average satisfactory, least misery, and most pleasure. What is the firs product recommended to the group using each strategy.

Average Satisfaction	Least Misery	Most pleasure
r_god = (3+5+1)/3 = 3	r_god = min{3,5,1} = 1	r_god = max{3,5,1} = 5
r_rogue = (0+4+2)/3 = 2	r_rogue = min{0,4,2} = 0	r_rogue = max{0,4,2} = 4
r_deu = (3+0+4)/3 = 2.3	r_deu = min{3,0,4} = 0	r_deu = max{3,0,4} = 4
r_rash = (2+2+3)/3 = 2.3	r_rash = min{2,2,3} = 2	r_rash = max{2,2,3} = 3
r_bella = (4+3+1)/3 = 2.6	r_bella = min{4,3,1} = 1	r_bella = max{4,3,1} = 4
Recommend God	Recommend Rashomon	Recommend God