Name:	PRAVEENAA KULANBHAVEL
ASU ID Number:	1225814683

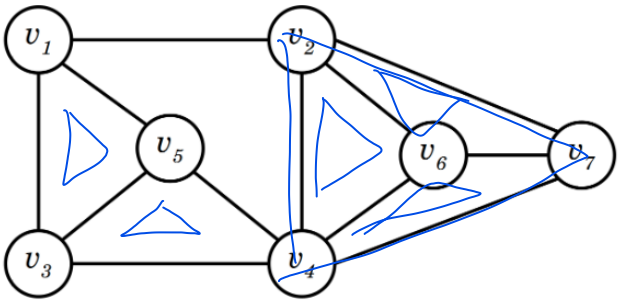
CSE 472: Social Media Mining

Homework III - Information Diffusion, Community Analysis

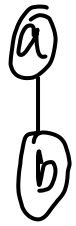
Prof. Huan Liu
Due at 2022 Oct. 27, 11:59 PM

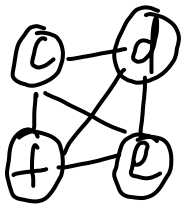
This is an ***individual*** homework assignment. Please submit a digital copy of this homework to **Gradescope**. For your solutions, even when not explicitly asked you are supposed to concisely justify your answers.

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



a = (V1, V3, V5)
b = (V4, V3, V5)
c = (V2, V6, V4)
d = (V7, V6, V4)
e = (V7, V6, V2)
f = (V7, V4, V2)





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

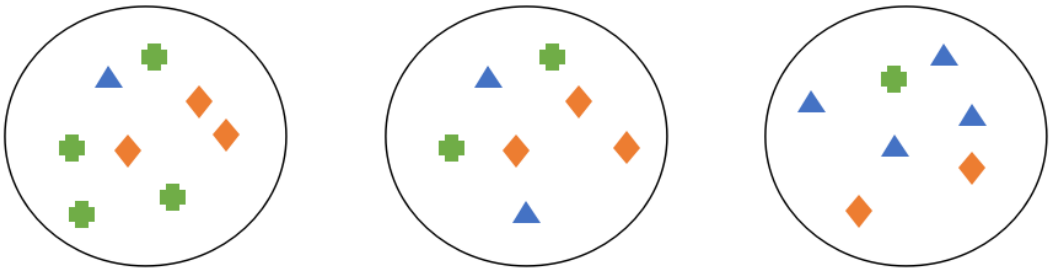


Figure 1: The communities.

- Precision and recall

Total pairs = $\frac{22+21}{2} = 231$

Precision = $\frac{TP}{TP+FP} = \frac{21}{21+49} = 0.3$

Recall = $\frac{TP}{TP+FN} = \frac{21}{21+49} = 0.3$

True Positives + False Positives = $\binom{8}{2}+\binom{7}{2}+\binom{1}{2} = 28+21+21 = 70$

True Positives = $\binom{3}{2}+\binom{4}{2}+\binom{3}{2}+\binom{2}{2}+\binom{2}{2}+\binom{4}{2}+\binom{2}{2} = 3+6+3+1+1+6+1 = 21$

False Positives = $70-21 = 49$

True Negatives + False Negatives = $231-70 = 161$

False Negatives = $4\times3+3\times5+1\times6+3\times2+2\times4+2\times1$
$= 12+15+6+6+8+2 = 34+15 = 49$

True Negatives = $161-49 = 112$

- F-measure

$$\text{F-Measure} = 2\times\left(\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}\right) = \frac{2\times(0.3\cdot0.3)}{0.3+0.3}$$

$$= \frac{2\times0.09}{0.6} = 0.3$$

- NMI

$$H(Y) = -\left(\frac{7}{22}\log\frac{7}{22} + \frac{8}{22}\log\frac{8}{22} + \frac{7}{22}\log\frac{7}{22}\right) = 0.45$$

$$H(Y|C=1) = \boxed{-P(C=1)\sum P(Y=y|C=1)\log(P(Y=y|C=1))}$$

$$= \frac{-1}{3}\times\left(\frac{3}{8}\log\frac{3}{8} + \frac{4}{8}\log\frac{4}{8} + \frac{1}{8}\log\frac{1}{8}\right) = \frac{-1}{3}\times(-0.159 + -0.15 + -0.11) = 0.13$$

$$H(Y|C=2) = \frac{-1}{3}\times\left(\frac{3}{7}\log\frac{3}{7} + \frac{2}{7}\log\frac{2}{7} + \frac{2}{7}\log\frac{2}{7}\right) = \frac{-1}{3}\times(0.157 + -0.155 + -0.15) = 0.155$$

$$H(Y|C=3) = \frac{-1}{3}\times\left(\frac{4}{7}\log\frac{4}{7} + \frac{2}{7}\log\frac{2}{7} + \frac{1}{7}\log\frac{1}{7}\right) = \frac{1}{3}\times(0.138 + 0.155 + 0.12) = 0.137$$

Mutual Information $I(Y;C) = H(Y) - H(Y|C)$

$= 0.45 - (0.13 + 0.155 + 0.137) = 0.028$

$$NMI(Y,C) = \frac{2\times I(Y;C)}{[H(Y)+H(C)]} = \frac{2\times 0.028}{[0.45+1]} = \frac{0.056}{1.45} = 0.038$$

- Purity

Community 1 – 4 ✩

Community 2 – 3 ◇

Community 3 – 4 △

$$\text{Purity} = \frac{① + ② + ③}{\text{Total}}$$

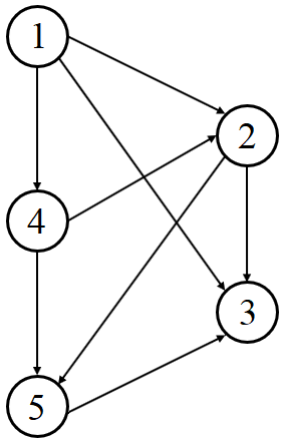
$$= \frac{4+3+4}{22} = \frac{11}{22}$$

3. **[Information Diffusion]** Suppose we want to maximize the spread of cascades using the following greedy algorithm that we learned in class.

Algorithm 1 Maximizing the spread of cascades – Greedy algorithm

Require: Diffusion graph $G(V, E)$, budget k

return Seed set S (set of initially activated nodes)
$i = 0$;
$S = \{\}$;
while $i \neq k$ **do**
  $v = \arg\max_{v \in V \setminus S} f(S \cup \{v\})$;
  or equivalently $\arg\max_{v \in V \setminus S} f(S \cup \{v\}) - f(s)$
  $S = S \cup \{v\}$;
  $i = i + 1$;
end while
Return S;



Assume that node v activates node w when $|v - w^2| \equiv 0 (\text{mod } 3)$ and our budget is $k = 2$. For the given network, between node 1 and node 4, which one should the algorithm activate at time 0? (Show **all** your work and detail **all** your calculations).

Node 1	**Node 4**
①⇒②,④,⑤ ②⇒ $\|1-2^2\| \equiv 0 (\text{mod } 3) \Rightarrow 3 \equiv 0 (\text{mod } 3)$ ✓ ④⇒ $\|1-4^2\| \equiv 0 (\text{mod } 3) \Rightarrow 15 \equiv 0 (\text{mod } 3)$ ✓ ⑤⇒ $\|1-5^2\| \equiv 0 (\text{mod } 3) \Rightarrow 8 \equiv 1 (\text{mod } 3)$ ✗ So, ①⇒②④ Next ②→outgoing ⑤&③ ③⇒ $\|2-3^2\| \Rightarrow \|7\| \equiv \text{mod } 3$ ✗ ⑤⇒ $\|2-5^2\| \Rightarrow \|23\| \equiv \text{mod } 3$ ✗ Next ④→outgoing ②&⑤ ②→ $\|4-2^2\| \Rightarrow \|21\| \equiv \text{mod } 3$ ✓ Next ⑤→outgoing ③ ③⇒ $\|5-3^2\| = \|4\| \equiv \text{mod } 3$ ✗ Nodes activated by ①⇒②,④,⑤	②⇒ $\|4-2^2\| \equiv 0 (\text{mod } 3)$ $0 \equiv 0 (\text{mod } 3)$ ✓ ⑤⇒ $\|4-5^2\| \equiv 0 (\text{mod } 3)$ $\|21\| \equiv 0 (\text{mod } 3)$ ✓ Next ②→⑤,③ ③⇒ $\|2-3^2\| \equiv 0 (\text{mod } 3)$ $\|7\| \equiv 0 (\text{mod } 3)$ ✗ Next ⑤⇒③ ③⇒ $\|5-3^2\| = 0 (\text{mod } 3) \Rightarrow \|4\| \equiv (\text{mod } 3)$ ✗ Nodes activated by ④⇒②,⑤
To maximize the spread, which node should get activated at time 0?	Node ①