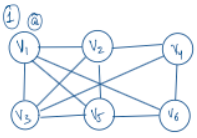
HW #5 - Solution

1. **Given the following adjacency matrix**

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

   a. **Provide the maximal 2-cliques for this graph.**
   b. **Provide the maximal 2-plexes for this graph.**
   c. **Are there any structually equivalent nodes in this graph? If yes, list them.**

(1) (a)



All Nodes are within distance 2 to any other node in the graph.
So, Maximal 2-Clique is $\{V_1, V_2, V_3, V_4, V_5, V_6\}$

(b) For a graph to be k-plex every vertex should have a degree of atleast $|V| - K$, where $|V|$ is number of vertices in the graph.
for entire graph the minimum degree is 3 (for V_5, V_6)
So,

$\Rightarrow |V| - K \leq 3$
$\Rightarrow 6 - K \leq 3$
$\Rightarrow K \geq 3$ (3-plex graph)

So, the entire graph cannot be a 2-plex

Considering for $|V| = 5$)
removing any node results in minimum degree of 2 (again for V_5, V_6)

$\Rightarrow 5 - K \leq 2$
$\Rightarrow K \geq 3$ (again 3-plex)

So, node size of 5 cannot be a 2-plex either.

for Node size = 4,
for all the below subgraphs the minimum degree = 2,

1. V_1, V_2, V_3, V_5
2. V_2, V_4, V_5, V_6
3. V_2, V_3, V_5, V_4
4. V_1, V_2, V_5, V_6
5. V_1, V_3, V_4, V_6
6. V_1, V_3, V_5, V_6
7. V_1, V_2, V_4, V_6
8. V_3, V_4, V_5, V_6
9. V_1, V_2, V_3, V_4.

$\Rightarrow 4 - K \leq 2$
$\Rightarrow K \geq 2$ (2-plex)

So, all the above subgraphs are 2-plex.

Any connected subgraph of nodesize 3, 2, 1 will also be 2-plex, However only the listed subgraphs of nodesize 4 are considered "maximal" 2-plex.

(c) Yes, $\{V_1, V_2\}$ & $\{V_2, V_3\}$ have same neighbours,
So they are structurally equivalent.

OR 1(c)

Two nodes are structurally equivalent when they have the exact same neighbors in the graph

- Node 1 neighbors are 2 3 5 6
- Node 2 neighbors are 1 3 4 5
- Node 3 neighbors are 1 2 4 5
- Node 4 neighbors are 2 3 6
- Node 5 neighbors are 1 2 3 6
- Node 6 neighbors are 1 4 5

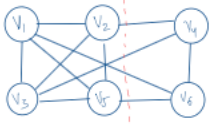
No two nodes share the same neighbor set

➔ **Therefore no pair of nodes is structurally equivalent**

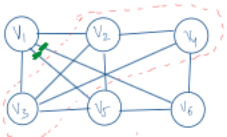
2. **Using the graph from problem 1,**
   a. **Find a minimum cut that creates partitions P_1 and P_2 where $|P_1|=4$ and $|P_2|=2$.**
   b. **Find a minimum cut that creates partitions P_1 and P_2 where $|P_1| = |P_2|= 3$.**
   c. **For each of the cuts in a and b above, calculate**
      i. **Ratio Cut (P)**
      ii. **Normalized Cut (P)**
   d. **Which cut is preferable based on the above metrics?**

(2)

(a) for Partitions $|P_1|=4$ and $|P_2|=2$
Minimum Cut = 4
partitions ⇒ P_1: V_1, V_2, V_3, V_5
P_2: V_4, V_6

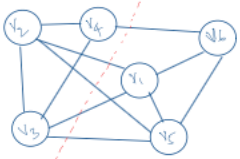


(b) for Partitions $|P_1|=|P_2|=3$
Minimum Cut = 5
partitions ⇒ P_1: V_2, V_3, V_4
P_2: V_1, V_5, V_6



∵ Edges marked in green will not be cut

(or) for Better Visualization & Understanding

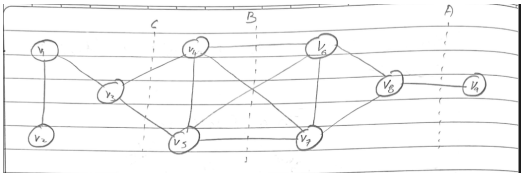


(c) (a) ratio cut & Normalized cut for $|P_1|=4$, $|P_2|=2$
⇒ ratio_cut = $\frac{1}{2}\left(\frac{4}{4}+\frac{4}{2}\right)=\frac{1}{2}(1+2)=\frac{3}{2}$
⇒ Normalized_cut = $\frac{1}{2}\left(\frac{4}{16}+\frac{4}{2}\right)=\frac{1}{2}\left(\frac{1}{4}+\frac{2}{5}\right)=\frac{9}{24}$

(b) ratio cut & Normalized cut for $|P_1|=|P_2|=3$
⇒ ratio_cut = $\frac{1}{2}\left(\frac{5}{2}+\frac{5}{3}\right)=\frac{5}{3}$
⇒ Normalized_cut = $\frac{1}{2}\left(\frac{5}{9}+\frac{5}{9}\right)=\frac{5}{11}$

(d) ⇒ for ratio_cut 4-2 split should be preferred since it is lesser
⇒ for Normalized cut 3-3 split, since it is approx lesser than 4-2.

3. Social Media Mining (SMM) Ch 6, problem 7



Now here we have to compute Jaccard and cosine similarity between nodes V_4 and V_8.

We have to assuming that neighbour of a node excludes the node itself.

The node excludes itself :-

$\therefore N(V_4) = \{V_3, V_5, V_6, V_7\}$

$\therefore N(V_8) = \{V_6, V_7, V_9\}$

We will check for Jaccard

$$J(V_i, V_j) = \frac{|N(V_i) \cap N(V_j)|}{|N(V_i) \cup N(V_j)|}$$

$$J(V_4, V_8) = \frac{|N(V_4) \cap N(V_8)|}{|N(V_4) \cup N(V_8)|}$$

$$= \frac{|\{V_3, V_5, V_6, V_7\} \cap \{V_6, V_7, V_9\}|}{|\{V_3, V_5, V_6, V_7\} \cup \{V_6, V_7, V_9\}|}$$

$$= \frac{|\{V_6, V_7\}|}{|\{V_3, V_5, V_6, V_7, V_9\}|}$$

$$= \frac{2}{5} \quad \boxed{= 0.4}$$

Now lets find the cosine

$$C(V_i, V_j) = \frac{|N(V_i) \cap N(V_8)|}{\sqrt{|N(V_i)| \cdot N(V_j)|}}$$

$$= \frac{|\{V_3, V_5, V_6, V_7\} \cap \{V_6, V_7, V_9\}|}{\sqrt{|\{V_3, V_5, V_6, V_7\}| \cdot |\{V_6, V_7, V_9\}|}}$$

$$= \frac{|\{V_6, V_7\}|}{\sqrt{12}} = \frac{2}{\sqrt{12}} \quad \boxed{= 0.577}$$

⇒ The node is including itself.

$N(V_4) = \{V_3, V_4, V_5, V_6, V_7\}$
$N(V_8) = \{V_6, V_7, V_8, V_9\}$

We will check for Jaccard

$$J(V_4, V_8) = \frac{|N(V_4) \cap N(V_8)|}{|N(V_4) \cup N(V_8)|}$$

$$= \frac{|\{V_3, V_4, V_5, V_6, V_7\} \cap \{V_6, V_7, V_8, V_9\}|}{|\{V_3, V_4, V_5, V_6, V_7\} \cup \{V_6, V_7, V_8, V_9\}|}$$

$$= \frac{|\{V_6, V_7\}|}{|\{V_3, V_4, V_5, V_6, V_7, V_8, V_9\}|} = \frac{2}{7} \quad \boxed{= 0.286}$$

Now lets calculate for cosine

$$C(V_4, V_8) = \frac{|N(V_4) \cap N(V_8)|}{\sqrt{|N(V_4)| \cdot |N(V_8)|}}$$

$$= \frac{|\{V_3, V_4, V_5, V_6, V_7\} \cap \{V_6, V_7, V_8, V_9\}|}{\sqrt{|\{V_3, V_4, V_5, V_6, V_7\}| \cdot |\{V_6, V_7, V_8, V_9\}|}}$$

$$= \frac{|\{V_6, V_7\}|}{\sqrt{5 \cdot 4}}$$

$$= \frac{2}{\sqrt{20}} \quad \boxed{= 0.447}$$

4. **SMM Ch 6, problem 9**

- **What are the maximum and minimum values for NMI? Provide details**.

The NMI is always constrained by the values of 0 and 1.

When the minimum value of 0 occurs it means that the detected communities and the true labels have no relationship whatsoever. In such cases, knowing the community to which a given node belongs does not help you guess its true label at all.

When the maximum value of 1 occurs, there is a perfect match between the communities and true labels (although it may only use different names). In this case, once you know the community, you know the label for sure so it is a perfect match.

Thus, NMI $\in [0, 1]$. Values close to 1 indicate very strong similarity between communities found and communities corresponding to the true labels and values close to 0 indicate there is essentially no relationship.

- **Explain how NMI works (describe the intuition behind it)**

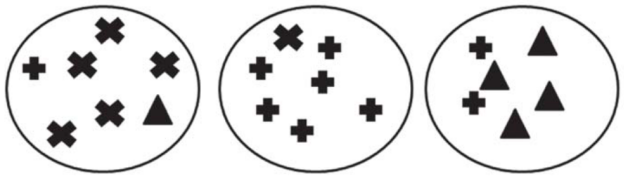
NMI takes the idea of mutual information and transforms it into a score from 0 to 1. Mutual information tells us how much knowing one of the thing gives you information to use in learning about the other thing. Here, the two things are:

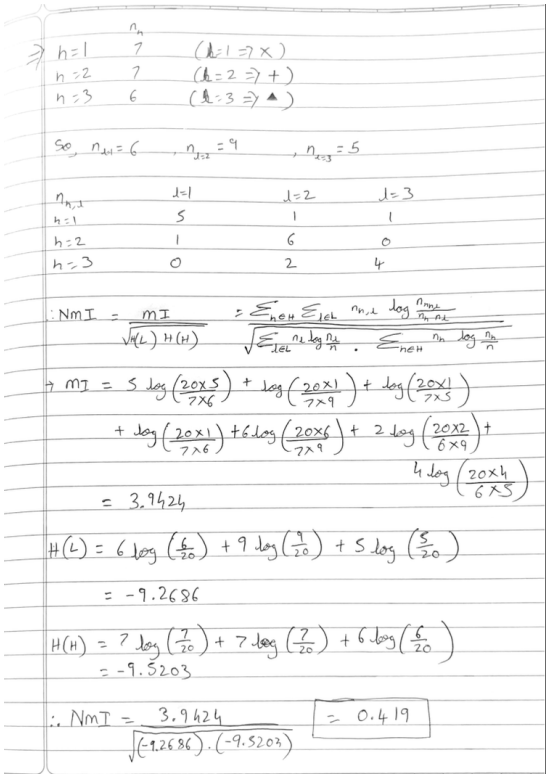
- the community assigned to a given node, and
- the true label of that node.

When clustering is good, knowing the community provides us with a lot of information about the true label. In this case, NMI is high. When clustering is bad, knowing the community does not help us guess the label of the node and NMI is low.

Since we normalize this information we are assured NMI will stay between 0 and 1 allowing us to compare different clusterings fairly even if they have different numbers of communities or different label distributions.

5. **SMM Ch 6, problem 10**





6. SMM Ch 6, problem 11

High Precision alone is insufficient for community evaluation because it ignores completeness. Precision focuses only on the percentage of correctly identified elements in clusters.

A model with high precision may still fail to identify many true positives, leading to low recall and poor overall performances.

(1) Precision :- Emphasizes the accuracy of the detected clusters.

(2) Recall :- Highlights the completeness of the detected clusters.

Combining precision and recall ensures a balanced and comprehensive evaluation of clustering performances.

Let us consider an example of assessing community detection in a network containing both discovered clusters and real communities.

Ground truth communities :-

- ➔ Community 1 :- { A, B, C, D, E }
- ➔ Community 2 :- { F, G, H, I, J }

Detected Clusters:-

- Cluster 1 : {A, B}
- Cluster 2 : {F, G, H, I, J}

Calculating Precision and Recall for Cluster:-
(1) True Positive :- {A, B} ⇒ 2
(2) False Positive :- Ø None ⇒ 0
(3) False Negative :- {C, D, E} ⇒ 3

⇒ Precision = TP/(TP + FP)
= 2/(2 + 0)
= 1

⇒ Recall = TP/(TP + FN)
= 2/(2 + 3)
= 0.4

This demonstrates that while precision is perfect for Cluster 1, the recall is low.

In conclusion, a high precision alone can be misleading as a good measure of performance, especially when the data are imbalanced; only by taking into account precision and recall can we meaningfully assess the performance of a community detection method.

7. SMM Ch 6, problem 12

Purity is a metric that measures the homogeneity of clusters with respect to the ground truth labels.

Although it is easy and intuitive to compute, purity is either inappropriate or misleading in some cases as a method of evaluating community detection performances.

Situation where Purity is Misleading :-

(1) **Lack of Balance between Clusters** :-
The value of purity can be driven upwards when the algorithm obtains multiple small "pure" clusters. These will not capture the global community structure well.

Let take an example to solve it.
• Ground Truth communities:-
Community 1 : { A, B, C, D, E }
Community 2 :- { F, G, H, I, J }

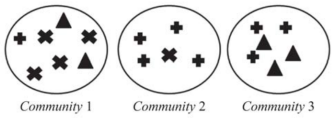
• Detected Cluster:-
Cluster1 – { A }
Cluster2 – { B }
Cluster3– { C }
Cluster4 – { F, G, H, I, J }

If each detected cluster includes nodes exclusively from the same ground truth community, purity will be 1.

(2) **Ignoring Completeness :-**
Purity focuses only on how homogenous a cluster is. It overlooks whether all relevant nodes from a community are assigned to the same cluster. As a result cluster is ignored.

(3) **Imbalanced Data :-**
When ground truth communities are imbalanced purity becomes deceptive. Assigning nodes from smaller communities to larger ones artificially inflates the score, favoring larger community unfairly.

8. SMM Ch 6, problem 13



- So let us solve the Precision and recall

(1) True Positive :- similar members are assigned to the same community.

$$TP = \underbrace{\left(\binom{4}{2}+\binom{2}{2}\right)}_{\text{Community 1}} + \underbrace{\binom{4}{2}}_{\text{Community 2}} + \underbrace{\left(\binom{3}{2}+\binom{3}{2}\right)}_{\text{Community 3}}$$

$$= (6+1) + 6 + (3+3)$$

$$= 7 + 6 + 6$$

$$\boxed{= 19}$$

(2) False Positive :- Disimilar members are assigned to the different community.

$$FP = \underbrace{(4\times1 + 4\times2 + 2\times1)}_{\text{Community 1}} + \underbrace{(4\times1)}_{\text{Community 2}} + \underbrace{(3\times3)}_{\text{Community 3}}$$

$$= (4+8+2) + 4 + 9$$

$$\boxed{= 27}$$

(3) False negative :- Similar members are assigned to different communities.

$$FN = \underset{(\times)}{(4\times1)} + \underset{(+1)}{(1\times4 + 4\times3 + 3\times1)} + \underset{(\Delta)}{(2\times3)}$$

$$= 4 + 19 + 6$$

$$= 29$$

(4) True negative :- dissimilar members are assigned to the different communities.

$$TN = \underset{\text{community 1 \& 2}}{(4\times4 + 1\times1 + 2\times4 + 2\times1)} + \underset{\text{community 2 \& 3}}{(4\times3 + 1\times3 + 1\times3)}$$

$$+ \underset{\text{community 1 \& 3}}{(4\times3 + 4\times3 + 1\times3 + 2\times3)}$$

$$= (16+1+8+2) + (12+3+3) + (12+12+3+6)$$

$$= 27 + 33 + 18$$

$$= 78$$

So Now Calculate Precision and recall

$$\therefore P = \frac{TP}{TP+FP} = \frac{19}{19+27} = 0.4130$$

$$\therefore R = \frac{TP}{TP+FN} = \frac{19}{19+29} = 0.3958$$

- F-measure :- $2 \times \frac{P \times R}{P+R}$

$$= 2 \times \frac{(0.4130 \times 0.3958)}{(0.4130 + 0.3958)}$$

$$= 2 \times \frac{0.1634}{0.8088}$$

$$= 0.4040$$

- Purity :- $\frac{1}{N} \sum_{i=1}^{K} \max_{j} |C_i \cap L_j|$

Here, $N = 18$

$K = 3$

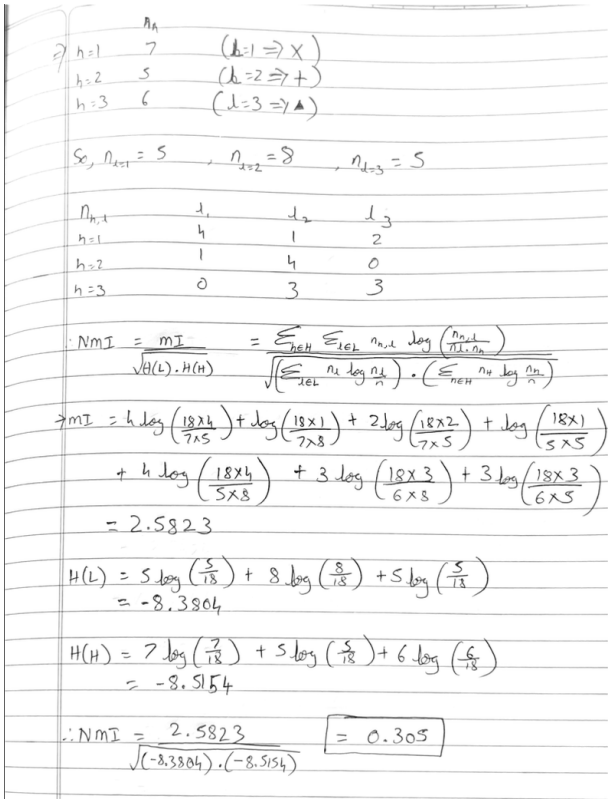
Majority labels in com-1 is X, which has 4 nodes
Majority labels in com-2 is +, which has 4 nodes
Majority labels in com-3 is Δ, which has 3 nodes

$$\therefore \text{Purity} = \frac{1}{18} \sum_{i=1}^{3} \max_{j} |C_i \cap L_j|$$

$$= \frac{1}{18}(4+4+3)$$

$$= \frac{11}{18}$$

$$= 0.6111$$



9. SMM Ch 7, problem 2

We need to establish an experiment that addresses herd behavior, where individuals can make decisions consecutively, each being able to see what the previous individuals have decided, but still have some private information of their own. There will need to be two clear choices for example lets take Community Area X vs Community Area Y. Each individual is assigned a private signal and individuals are instructed not to share this signal with anyone or talk about the signal itself. Individuals are instructed only to view what others have decided. Under this procedure we will be able to see if people continue to use their own information or they start to use the majority behavior they see in all the previous people.

For a concrete design, consider an example concerning a study of the best locations for new housing: Community Area X or Community Area Y. The city knows in secret that one area has slightly better conditions than the other, but the participants do not. The participants come in one at a time. Each participant is given a private brochure that weakly recommends one of the two community areas X or Y that has been simulated on data related to safety, schools, or services. After reading the brochure, the participant must publicly state I would recommend Area X or I would recommend Area Y, and their recommendation is written on a board that all of the subsequent participants will see. The next participant enters the room after all of the previous participants have made their public recommendation. The subsequent participant sees all of the previous recommendations on the board, reads the brochure they have received, and makes their public recommendation thereafter. Over time, if many of the earlier participants pick the same community area to recommend, then the late participants may ignore

their own brochures and simply follow the apparent majority recommendation. This is an example of ensemble behavior in a realistic community area setting.

10. SMM Ch 7, problem 3

The dissemination of a novel idea or product is modelled by

$dA(t)/dt = i(t)[P-A(t)]$

where A(t) refers to the number of the adopters at time t, P signifies the total number of individuals who could adopt, and i(t) denotes the strength of influence at time t. The phrase [P−A(t)] relays the number of individuals left to adopt the idea or product.

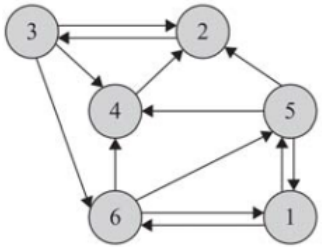
In the **external-influence model**, a person adopts primarily due to external factors, particularly the mass media and advertising, so adoption is instantaneous and grows at an even rate over time. In the **internal-influence model**, a person adopts due to their social network, and we will see the adoption of new ideas follow an S-curve with initial slow adoption, rapid adoption, and finally, a decreasing gain as adoption nears completion. In the **mixed-influence model**, you see both influences, so we will see early adoption and rapid gain in adoption due to external influence but eventually rapid gain, especially when the majority adopts because of internal influence.

11. SMM Ch 7, problem 4

The innovation is a community safety app introduced in an urban neighborhood to address local safety needs. The population is all residents who could potentially use the app. At first, only a few early adopters download it and report issues like broken streetlights or suspicious activity, influencing their neighbors by using and talking about it. At the same time the city promotes the app through ads, emails, signs, adding external influence. Together these internal and external influences drive the diffusion process: adoption is slow in the beginning then speeds up as more people hear about and use the app and finally levels off once most interested residents have adopted it.

An intervention is any deliberate action that changes this diffusion process. To speed up diffusion the city can increase external influence by running a larger awareness campaign or sending direct invitations and increase internal influence by rewarding people for inviting neighbors. To slow down a harmful innovation officials can warn people reduce promotion and discourage sharing so the idea spreads much more slowly.

12. SMM Ch 7, problem 7



The process follows the Independent Cascade Model. Each node gets only one chance, in the time step after it becomes active, to activate each of its outgoing neighbors. In this problem a node i activates a neighbor j only if there is a directed edge i → j and $i - j \equiv 1 \pmod 3$.

Node 5 is given as active at time 0. Let A(i, t) = 1 if node i is active at time t.

Time 0:-
Only node 5 is active.
A(5, 0) = 1.
Active set = {5}.

Time 1:-
Node 5 now tries to activate its out-neighbors 4, 1 and 2.

$5 - 4 = 1 \equiv 1 \pmod 3$ ⇒ node 4 becomes active, A(4, 1) = 1.
$5 - 1 = 4 \equiv 1 \pmod 3$ ⇒ node 1 becomes active, A(1, 1) = 1.
$5 - 2 = 3 \equiv 0 \pmod 3$ ⇒ node 2 is not activated.

New active nodes at time 1: {4, 1}.
Cumulative active set after time 1: {5, 4, 1}.

Time 2:-
Only the nodes that became active at time 1 (nodes 4 and 1) get a chance now.

Node 4 has out-neighbor 2:
$4 - 2 = 2 \equiv 2 \pmod 3$ ⇒ node 2 is not activated.

Node 1 has out-neighbors 5 and 6:
$1 - 5 = -4 \equiv 2 \pmod 3$ ⇒ node 5 is not newly activated.
$1 - 6 = -5 \equiv 1 \pmod 3$ ⇒ node 6 becomes active, so A(6, 2) = 1.

New active node at time 2: {6}.
Cumulative active set after time 2: {5, 4, 1, 6}.

Time 3:-
Only node 6 is newly active, so it now gets its one chance to activate its neighbors 1, 4 and 5.

$6 - 1 = 5 \equiv 2 \pmod 3 \Rightarrow$ node 1 is not newly activated.
$6 - 4 = 2 \equiv 2 \pmod 3 \Rightarrow$ node 4 is not newly activated.
$6 - 5 = 1 \equiv 1 \pmod 3$ but node 5 is already active, so there will be no change.

No new nodes are activated after time 3, so the process has converged.

The Independent Cascade Model converges with the active nodes {1, 4, 5, 6}.