CS579 - HW3 – Srivasa Pranathy Mutcherla – A20555875

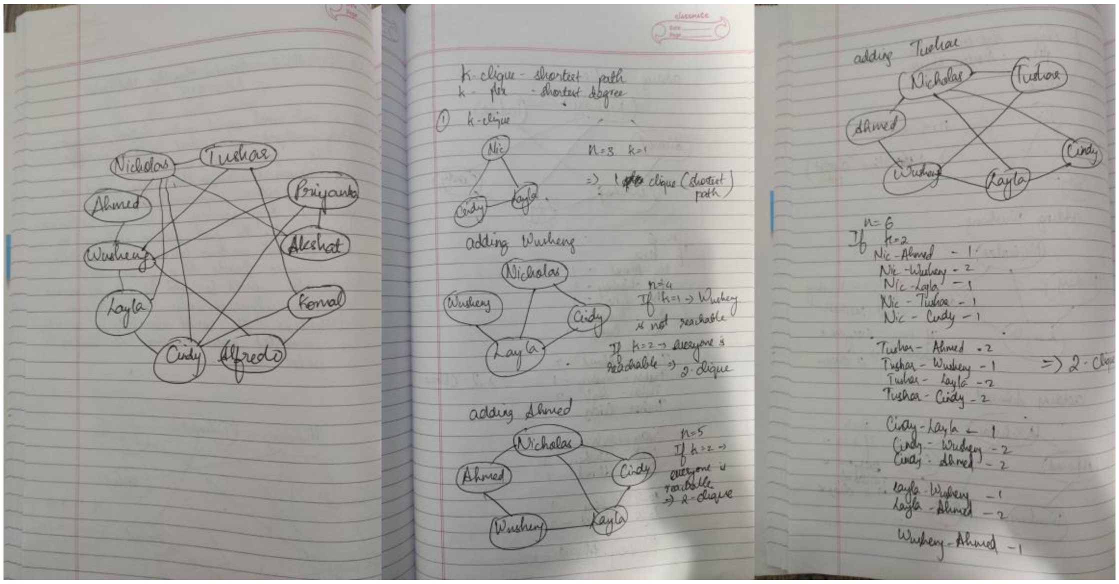
1. In HW#2, you simulated three different network models (random graph, small world, and preferential attachment). Given a real-world social media data set and an assignment to use this dataset to decide where to allocate resources for community building, describe (a) how you would determine which of the ab which of the above network models to use and (b) how you might use the models for your assignment.

(a) In a real-world social media dataset, the most likely network model that could be generated is the 'preferential model.' This works because we usually notice that people with many followers keep growing in this space. Influencers may be more focused on their follower count, which, when increased, leads to a greater reach.

(b) This model could allocate resources to highly influential nodes, leading to more distribution. It could also help connect less connected nodes to these hubs, leading to a broader reach.

2. Given the friendship graph from HW#1, find all (a) k-cliques and (b) k-plexes. (c) Describe the difference between these communities.

k-cliques and k-plexes:



k-clique - shortest path
k - plex - shortest degree

① k-clique

n=3 k=1

=> 1 clique (shortest path)

adding Wusheng

n=4
If k=1 -> Wusheng is not reachable

If k=2 -> everyone is reachable => 2-clique

adding Ahmed

n=5
If k=2 -> everyone is reachable => 2-clique

adding Tushar

n=6
If k=2
Nic-Ahmed - 1
Nic-Wusheng - 2
Nic-Layla - 1
Nic - Tushar - 1
Nic - Cindy - 1

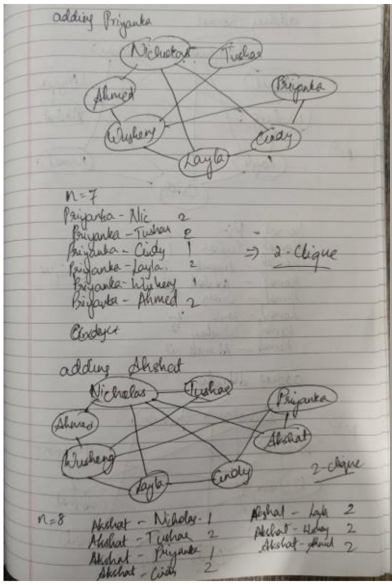
Tushar - Ahmed - 2
Tushar - Wusheng - 1
Tushar - Layla - 2
Tushar - Cindy - 2

=> 2-Clique

Cindy - Layla - 1
Cindy - Wusheng - 2
Cindy - Ahmed - 2

Layla - Wusheng - 1
Layla - Ahmed - 2

Wusheng - Ahmed - 1


adding Priyanka

$n=7$

Priyanka - Nic	2
Priyanka - Tushar	2
Priyanka - Cindy	1
Priyanka - Layla	2
Priyanka - Wusheng	1
Priyanka - Ahmed	2

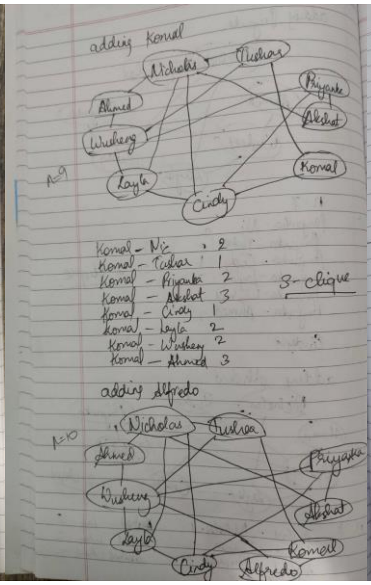
⇒ 2-Clique

adding Akshat

2-clique

$n=8$

Akshat - Nicholas	1
Akshat - Tushar	2
Akshat - Priyanka	1
Akshat - Cindy	2
Akshat - Layla	2
Akshat - Wusheng	2
Akshat - Ahmed	2


adding Komal

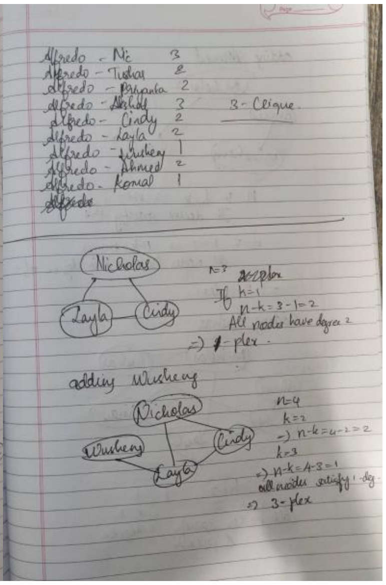
$n=9$

Komal - Nic	2
Komal - Tushar	1
Komal - Priyanka	2
Komal - Akshat	3
Komal - Cindy	1
Komal - Layla	2
Komal - Wusheng	2
Komal - Ahmed	3

3-clique

adding Alfredo

$n=10$



Alfredo - Nic	3
Alfredo - Tushar	2
Alfredo - Priyanka	2
Alfredo - Akshat	3
Alfredo - Cindy	2
Alfredo - Layla	2
Alfredo - Wusheng	1
Alfredo - Ahmed	2
Alfredo - Komal	1

~~Alfredo~~

3-Clique

$n=3$ ~~2-plex~~

If $k=1$

$n-k=3-1=2$

All nodes have degree 2

⇒ 1-plex.

adding Wusheng

$n=4$

$k=2$

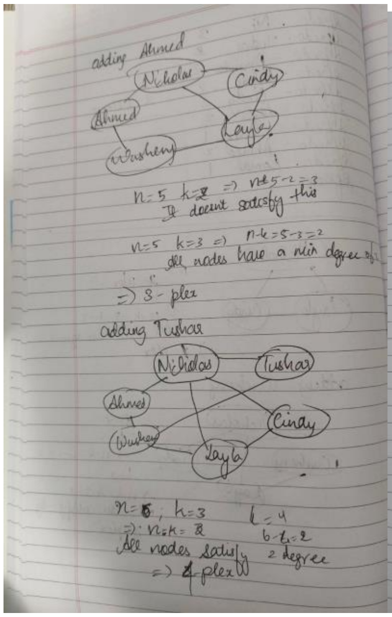
⇒ $n-k=4-2=2$

$k=3$

⇒ $n-k=4-3=1$

all nodes satisfy 1-deg.

⇒ 3-plex


adding Ahmed

$n=5$ $k=2$ ⇒ $n-k=5-2=3$

It doesnt satisfy this

$n=5$ $k=3$ ⇒ $n-k=5-3=2$

All nodes have a min degree of 2

⇒ 3-plex

adding Tushar

$n=6$, $k=3$

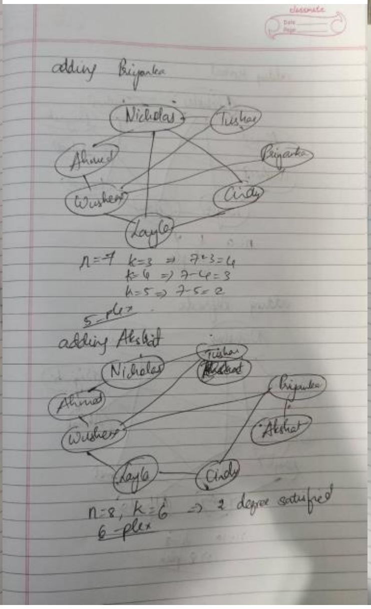
⇒ $n-k=3$

$k=4$

$6-4=2$

All nodes satisfy 2 degree

⇒ 4-plex


adding Priyanka

$n=7$ $k=3$ ⇒ $7-3=4$

$k=4$ ⇒ $7-4=3$

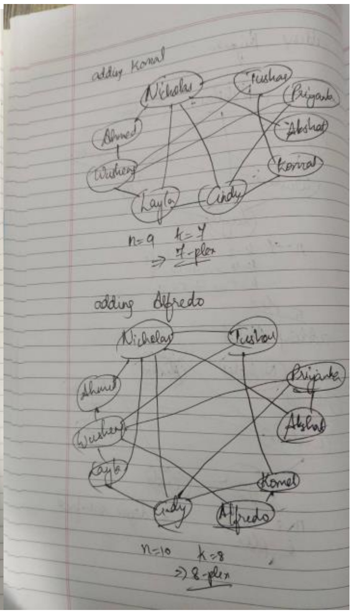
$k=5$ ⇒ $7-5=2$

5-plex

adding Akshat

$n=8$, $k=6$ ⇒ 2 degree satisfied

6-plex


adding Komal

$n=9$ $k=7$

⇒ 7-plex

adding Alfredo

$n=10$ $k=8$

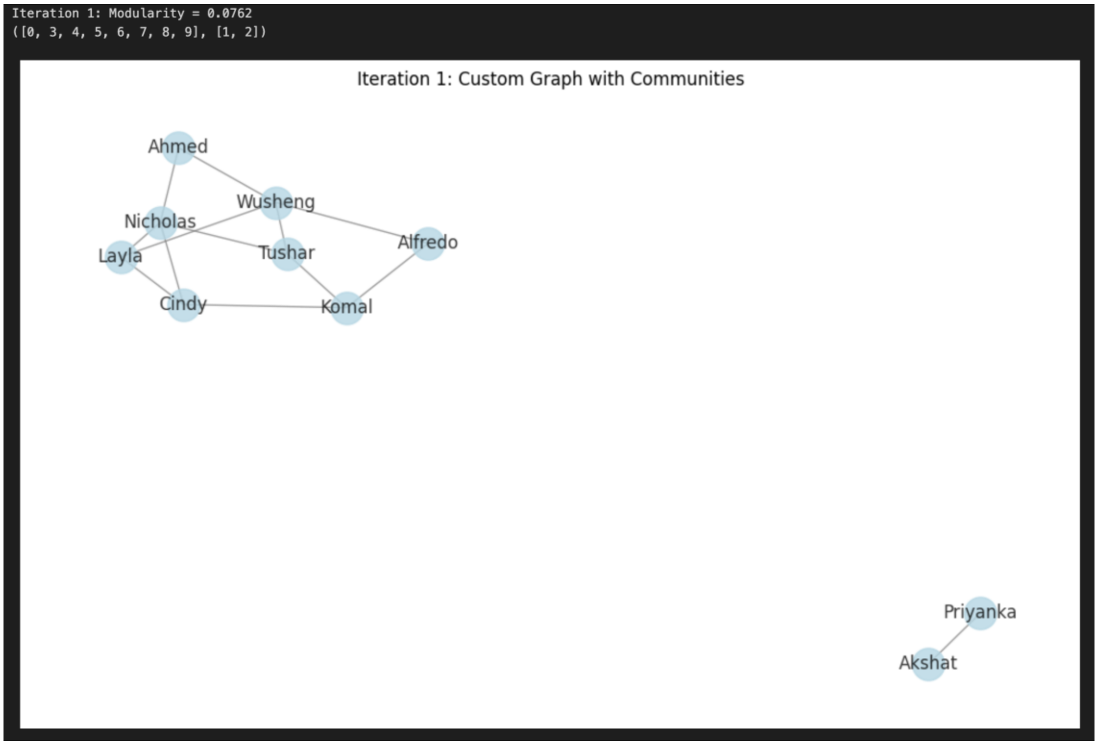
⇒ 8-plex

Apart from this, I’ve submitted the code file that generates all possible combinations. I’ve listed the maximal combinations by hand.

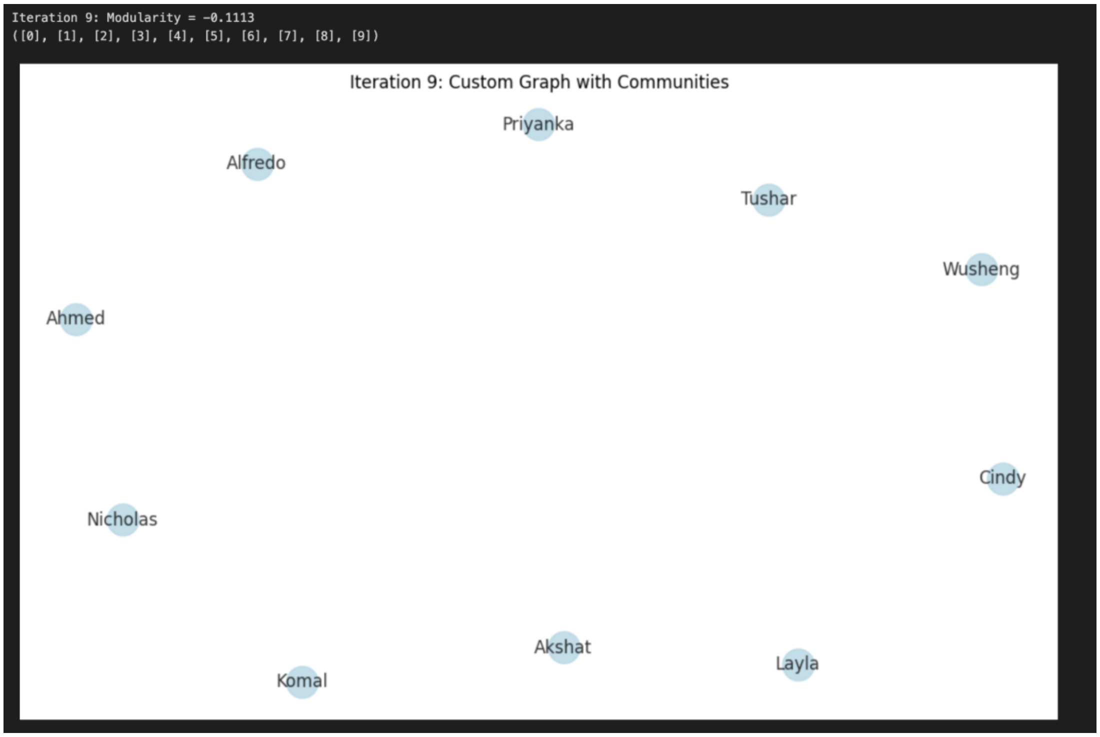
(c) The difference between these two communities is that plexes are more relaxed than cliques. We are comparing node reachability with node degree, which are two measures corresponding to member-based communities.

3. Given the friendship graph from HW#1, use the Girvan-Newman Algorithm to determine the hierarchical clustering dendrogram based on edge betweenness (see Fig 6.9 for example).

This has also been programmatically solved, with the attached code file generating all iterations (1 through 9) of the algorithm until all edges are removed.



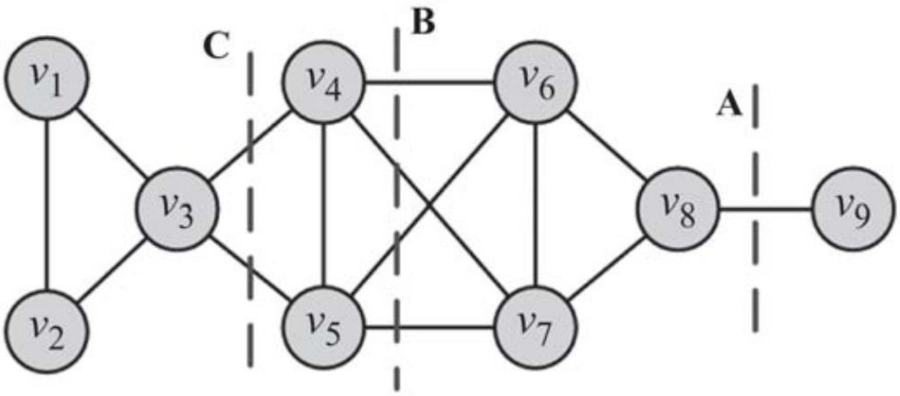
```
Iteration 1: Modularity = 0.0762
([0, 3, 4, 5, 6, 7, 8, 9], [1, 2])
```



```
Iteration 9: Modularity = −0.1113
([0], [1], [2], [3], [4], [5], [6], [7], [8], [9])
```

4. Text Ch 6 question 7 - For Figure 6.8:

- **Compute Jaccard and Cosine similarity between nodes v4 and v8, assuming that the neighborhood of a node excludes the node itself.**
- **Compute Jaccard and Cosine similarity when the node is included in the neighborhood.**



$$\sigma_{\text{Jaccard}}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|},$$

$$\sigma_{\text{Cosine}}(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}.$$

- $\sigma_{Jaccard}(v4,\ v8)\ =\ \left|\frac{N(v4)\cap N(v8)}{N(v4)\cup N(v8)}\right|$

$$\sigma_{Jaccard}(v4,\ v8)\ =\ \left|\frac{\{v3, v5, v6, v7\} \cap \{v6, v7, v9\}}{\{v3, v5, v6, v7, v9\}}\right|$$

$$\sigma_{Jaccard}(v4,\ v8)\ =\ \left|\frac{\{v6, v7\}}{\{v3, v5, v6, v7, v9\}}\right|$$

$$\sigma_{Jaccard}(v4,\ v8)\ =\ 0.4$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ \left|\frac{N(v4) \cap N(v8)}{\sqrt{(|N(v4)||N(v8)|)}}\right|$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ \left|\frac{\{v3, v5, v6, v7\} \cap \{v6, v7, v9\}}{\sqrt{(|\{v3, v5, v6, v7\}||\{v6, v7, v9\}|)}}\right|$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ \left|\frac{\{v6, v7\}}{\sqrt{(|\{v3, v5, v6, v7\}||\{v6, v7, v9\}|)}}\right|$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ 0.577$$

- $\sigma_{Jaccard}(v4,\ v8)\ =\ \left|\frac{N(v4)\cap N(v8)}{N(v4)\cup N(v8)}\right|$

$$\sigma_{Jaccard}(v4,\ v8)\ =\ \left|\frac{\{v3,v4,v5,v6,v7\}\cap\{v6,v7,v8,v9\}}{\{v3,v4,v5,v6,v7,v8,v9\}}\right|$$

$$\sigma_{Jaccard}(v4,\ v8)\ =\ \left|\frac{\{v6,v7\}}{\{v3,v4,v5,v6,v7,v8,v9\}}\right|$$

$$\sigma_{Jaccard}(v4,\ v8)\ =\ 0.2857$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ \left|\frac{N(v4)\cap N(v8)}{\sqrt{(|N(v4)||N(v8)|)}}\right|$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ \left|\frac{\{v3,v4,v5,v6,v7\}\cap\{v6,v7,v8,v9\}}{\sqrt{(|\{v3,v4,v5,v6,v7\}||\{v6,v7,v8,v9\}|)}}\right|$$

$$\sigma_{Co\sin e}(v4,\ v8)\ =\ \left|\frac{\{v6,v7\}}{\sqrt{(|\{v3,v4,v5,v6,v7\}||\{v6,v7,v8,v9\}|)}}\right|$$

$\sigma_{Co\sin e}(v4,\ v8)\ =\ 0.447$

5. Text Ch 6 question 9 - Normalized mutual information (NMI) is used to evaluate community detection results when the actual communities (labels) are known beforehand.

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

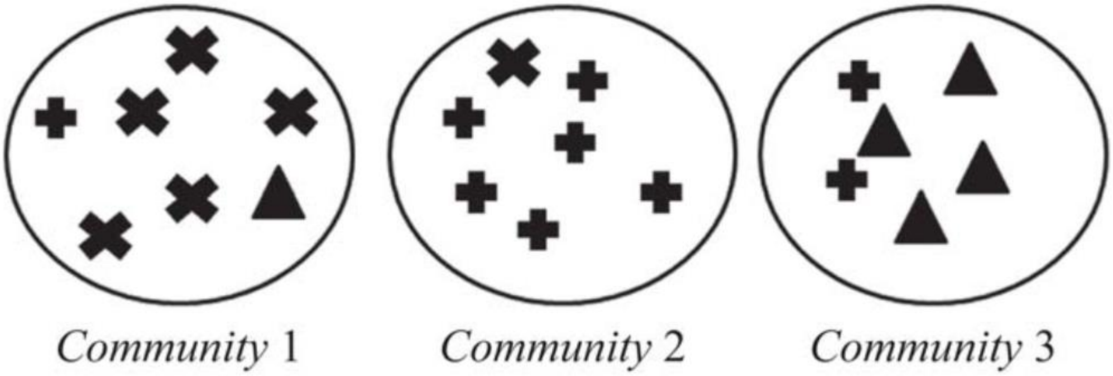
- **Explain how NMI works (describe the intuition behind it).**

- NMI ranges between 0 and 1 due to the normalization process. An NMI value close to one indicates a high similarity between communities found and labels. A value close to zero indicates a long distance between them. ￼

- NMI is based on mutual information between nodes based on information theory. When two nodes don't have any mutual information, their MI value is 0. Since MI is unbounded, NMI is a normalized value that ranges between 0 to 1. In the MI formula, we consider the nodes in the community (n_h), nodes with a particular label (n_l), and nodes in the community with that label ($n_{h,l}$). To normalize this, we use entropy functions. And after simplification, we end up with:

$$NMI = \frac{\sum_{h\in H}\sum_{l\in L} n_{h,l}\log\frac{n\cdot n_{h,l}}{n_h n_l}}{\sqrt{(\sum_{h\in H} n_h \log\frac{n_h}{n})(\sum_{l\in L} n_l \log\frac{n_l}{n})}}.$$

6. Text Ch 6 question 10 - Compute NMI for Figure 6.15.



Let's say we're going through labels x, +, and Δ through communities 1, 2, and 3.

$$\left(5\cdot\left(\log\left(\frac{(20\cdot 5)}{(7\cdot 6)}\right)\right)+1\cdot\left(\log\left(\frac{(20\cdot 1)}{(7\cdot 9)}\right)\right)+1\cdot\left(\log\left(\frac{(20\cdot 1)}{(7\cdot 5)}\right)\right)\right)$$
$$+\left(1\cdot\left(\log\left(\frac{(20\cdot 1)}{(7\cdot 6)}\right)\right)+6\cdot\left(\log\left(\frac{(20\cdot 6)}{(7\cdot 9)}\right)\right)+0\cdot\left(\log\left(\frac{(20\cdot 0)}{(7\cdot 5)}\right)\right)\right)$$
$$+\ \left(0\cdot\left(\log\left(\frac{(20\cdot 0)}{(7\cdot 6)}\right)\right)+2\cdot\left(\log\left(\frac{(20\cdot 2)}{(6\cdot 9)}\right)\right)+4\cdot\left(\log\left(\frac{(20\cdot 4)}{(6\cdot 5)}\right)\right)\right)$$

Community 1: $\left(5\cdot\left(\log\left(\frac{(20\cdot5)}{(7\cdot6)}\right)\right)+1\cdot\left(\log\left(\frac{(20\cdot1)}{(7\cdot9)}\right)\right)+1\cdot\left(\log\left(\frac{(20\cdot1)}{(7\cdot5)}\right)\right)\right)$

Community 1: $(1.883-0.498-0.243)\ =\ 1.142$

Community 2: $\left(1\cdot\left(\log\left(\frac{(20\cdot1)}{(7\cdot6)}\right)\right)+6\cdot\left(\log\left(\frac{(20\cdot6)}{(7\cdot9)}\right)\right)+0\cdot\left(\log\left(\frac{(20\cdot0)}{(7\cdot5)}\right)\right)\right)$

Community 2: $(-0.322+1.679+0)\ =\ 1.357$

Community 3: $\left(0 \cdot \left(\log\left(\frac{(20\cdot 0)}{(7\cdot 6)}\right)\right) + 2 \cdot \left(\log\left(\frac{(20\cdot 2)}{(6\cdot 9)}\right)\right) + 4 \cdot \left(\log\left(\frac{(20\cdot 4)}{(6\cdot 5)}\right)\right)\right)$

Community 3: $(0 - 0.260 + 1.703) \;=\; 1.443$

Entropy(H) = $\left(\left(7 \cdot \log\left(\frac{7}{20}\right)\right) + \left(7 \cdot \log\left(\frac{7}{20}\right)\right) + \left(6 \cdot \log\left(\frac{6}{20}\right)\right)\right)$

Entropy(H) = $(-3.191 - 3.191 - 3.137) \;=\; -9.519$

Entropy(L) = $\left(\left(6 \cdot \log\left(\frac{6}{20}\right)\right) + \left(9 \cdot \log\left(\frac{9}{20}\right)\right) + \left(5 \cdot \log\left(\frac{5}{20}\right)\right)\right)$

Entropy(L) = $(-3.137 - 3.121 - 3.010) \;=\; -9.268$

$$NMI \;=\; \frac{(Community1 + Community2 + Community3)}{\sqrt{\left((EntropyH)\cdot(EntropyL)\right)}}$$

$$NMI \;=\; \frac{(1.142 + 1.357 + 1.443)}{\left(\sqrt{\left((-9.519)\cdot(-9.268)\right)}\right)}$$

$$NMI \;=\; \frac{3.942}{9.392} \;=\; 0.419$$

7. Text Ch 6 question 11 - Why is high precision not enough? Provide an example to show that both precision and recall are important.

High precision alone is not sufficient because precision and recall are complementary metrics that provide a more comprehensive understanding of the performance of a classifier or information retrieval system.

Precision measures the accuracy of the positive predictions made by a model. It is calculated as the number of true positive results divided by the total number of positive predictions (true positives + false positives). A high precision indicates that the model's predictions are mostly correct when it predicts an instance as positive.

Recall, on the other hand, measures the model's ability to identify all relevant instances correctly. It is calculated as the number of true positive results divided by the total number of actual positive instances (true positives + false negatives). A high recall indicates that the model can capture most of the positive instances in the dataset.

Let’s say we take a spam classifier, which has:

TP = 20

FP = 5

FN = 0 (no spam classifiers were missed)

Then $\Pr e\,cision = \frac{(20)}{(20+5)} = 0.8$

$$Recall = \frac{(20)}{(20+0)} = 0$$

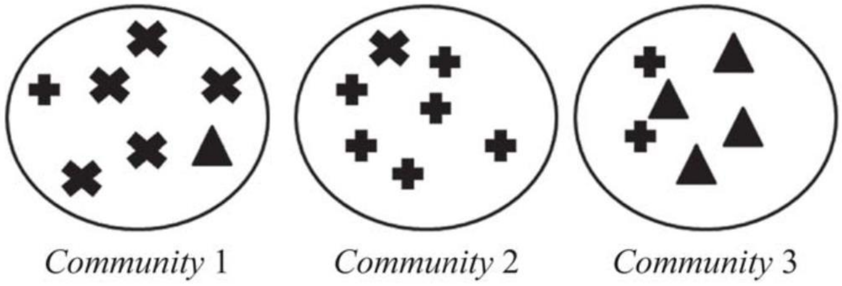
Here, even though precision is high, it does not imply that the system is good since the recall being 0 implies the system misses out on actual spam.

8. Text Ch 6 question 12 - Discuss situations where purity does not make sense.

Purity considers that the most frequent label represents the community, and we take an average of such majorities over the total number of nodes. While this might seem fair, when someone has one node per community, the total purity would be 1, which is inaccurate. Similarly, if we have large communities of similar nodes, we will still have purity close to 1, which does not make sense.

9. Text Ch 6 question 13 - Compute the following:

- **precision and recall**
- **F-measure**
- **NMI**
- **purity**



Precision and Recall

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$TP = \left(\binom{4}{2} + \binom{2}{2}\right) + \binom{4}{2} + \left(\binom{3}{2} + \binom{3}{2}\right)$$

$$TP = 7 + 6 + 6 = 19$$

$$FP = (2 \cdot 4 + 4 \cdot 1 + 2 \cdot 1) + (4 \cdot 1) + (3 \cdot 3) = 27$$

$$FN = (4 \cdot 1) + \big((1 \cdot 4) + (4 \cdot 3) + (1 \cdot 3)\big) + (2 \cdot 3) = 4 + 4 + 12 + 3 + 6 = 29$$

$$P = \frac{19}{19 + 27} = 0.413$$

$$R = \frac{19}{19 + 29} = 0.395$$

F-Measure:

$$F = 2 \cdot \left(\frac{P \cdot R}{R + P}\right)$$

$$F = 2 \cdot \left(\frac{0.413 \cdot 0.395}{0.413 + 0.395}\right) = 0.4037$$

Purity:

$$Purity = \frac{1}{N} \cdot \sum_{i=1}^{k} \max_j \left|C_i \cap L_j\right|$$

$$Purity = \frac{1}{18} \cdot [4 + 4 + 3] = 0.611$$

NMI:

$$NMI = \frac{-2 \cdot \sum_{i=1}^{S} \sum_{j=1}^{R} C_{ij} \log\left(\frac{C_{ij} \cdot N}{C_i \cdot C_j}\right)}{\sum_{i=1}^{S} C_i \cdot \log\left(\frac{C_i}{N}\right) + \sum_{j=1}^{R} C_j \cdot \log\left(\frac{C_j}{N}\right)}$$

$N = 18, S = 3, R = 3$

$$\sum_{i=1}^{S} C_i \log\left(\frac{C_i}{N}\right) \ = \ 5 \cdot \log\left(\frac{5}{18}\right) \ + \ 8 \cdot \log\left(\frac{8}{18}\right) + \ 5 \cdot \log\left(\frac{5}{18}\right) \ = \ -8.3804$$

$$\sum_{j=1}^{R} C_j \log\left(\frac{C_j}{N}\right) \ = \ 8 \cdot \log\left(\frac{8}{18}\right) \ + \ 5 \cdot \log\left(\frac{5}{18}\right) + \ 6 \cdot \log\left(\frac{6}{18}\right) \ = \ -8.4617$$

A1	A2	A3	Ci	
B1	4	1	0	5
B2	1	4	3	8
B3	2	0	3	5
Cj	8	5	6	

$$\sum_{i=1}^{S}\sum_{j=1}^{R} C_{ij} \log\left(\frac{C_{ij} \cdot N}{C_i \cdot C_j}\right)$$

$$= \ 4 \cdot \log\left(\frac{4 \cdot 18}{5 \cdot 18}\right) \ + \ 1 \cdot \log\left(\frac{1 \cdot 18}{5 \cdot 5}\right) + 0 + \ 1 \cdot \log\left(\frac{1 \cdot 18}{8 \cdot 8}\right) + 4 \cdot \log\left(\frac{4 \cdot 18}{5 \cdot 8}\right)$$

$$+ \ 3 \cdot \log\left(\frac{3 \cdot 18}{8 \cdot 6}\right) + 2 \cdot \log\left(\frac{2 \cdot 18}{5 \cdot 8}\right) \ + 0 + \ 3 \cdot \log\left(\frac{3 \cdot 18}{5 \cdot 6}\right) \ = \ 1.0589$$

$$NMI \ = \ \frac{-2 \cdot 1.0589}{-8.3804 - 8.4617} \ = \ 0.1257$$