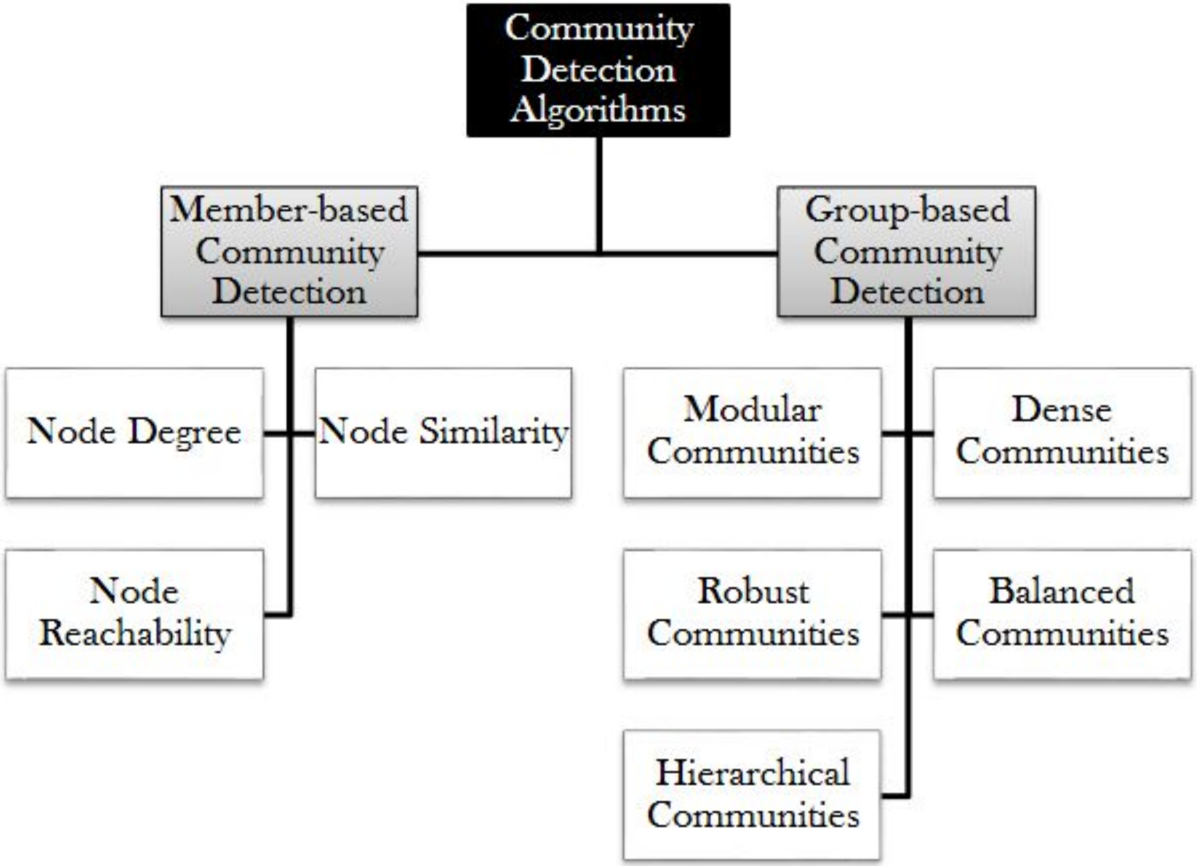
# Community Detection Algorithms



# Community Evolution

# Network and Community Evolution

- How does a **network** change over time?
- How does a **community** change over time?
- What properties do you expect to remain roughly constant?
- What properties do you expect to change?

# How Networks Evolve?

# Network Growth Patterns

1. Network Segmentation
2. Graph Densification
3. Diameter Shrinkage

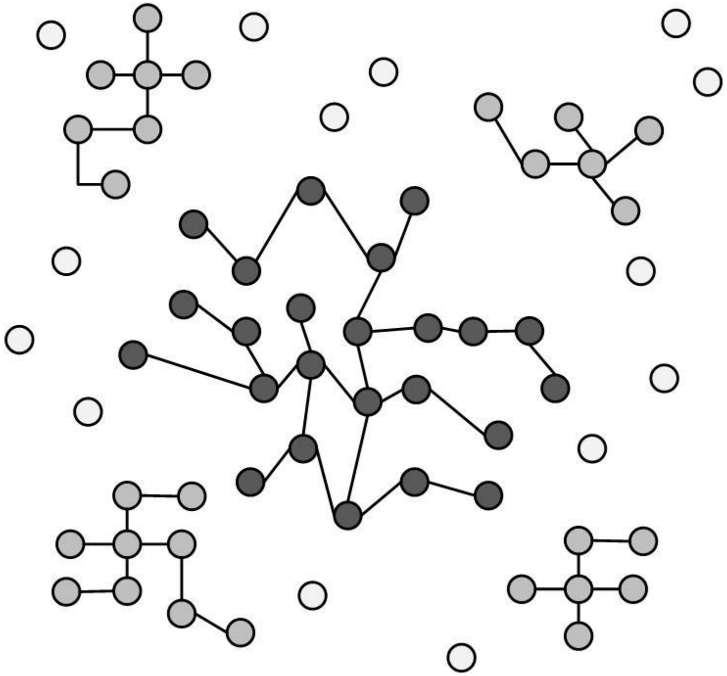
# 1. Network Segmentation

- Often, in evolving networks, segmentation takes place, where the large network is decomposed over time into three parts

1. **Giant Component**: As network connections stabilize, a giant component of nodes is formed, with a large proportion of network nodes and edges falling into this component.

2. **Stars**: These are isolated parts of the network that form star structures. A star is a tree with one internal node and n leaves.

3. **Singletons**: These are orphan nodes disconnected from all nodes in the network.



# 2. Graph Densification

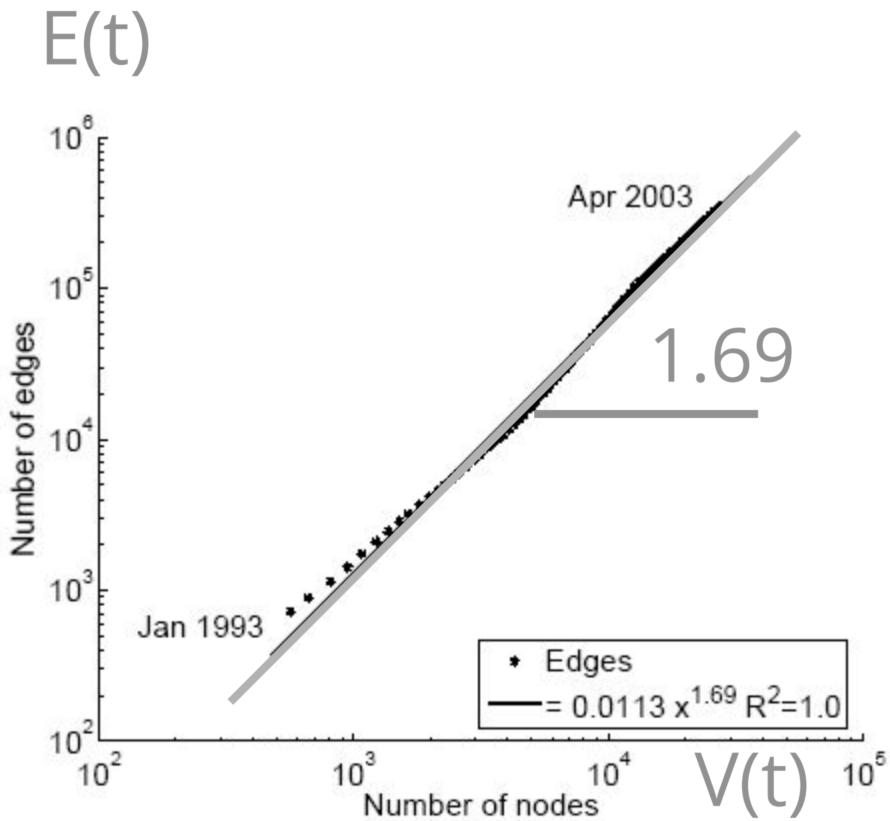
- The density of the graph increases as the network grows
  - The number of edges increases faster than the number of nodes does

$$E(t) \propto V(t)^{\alpha}$$

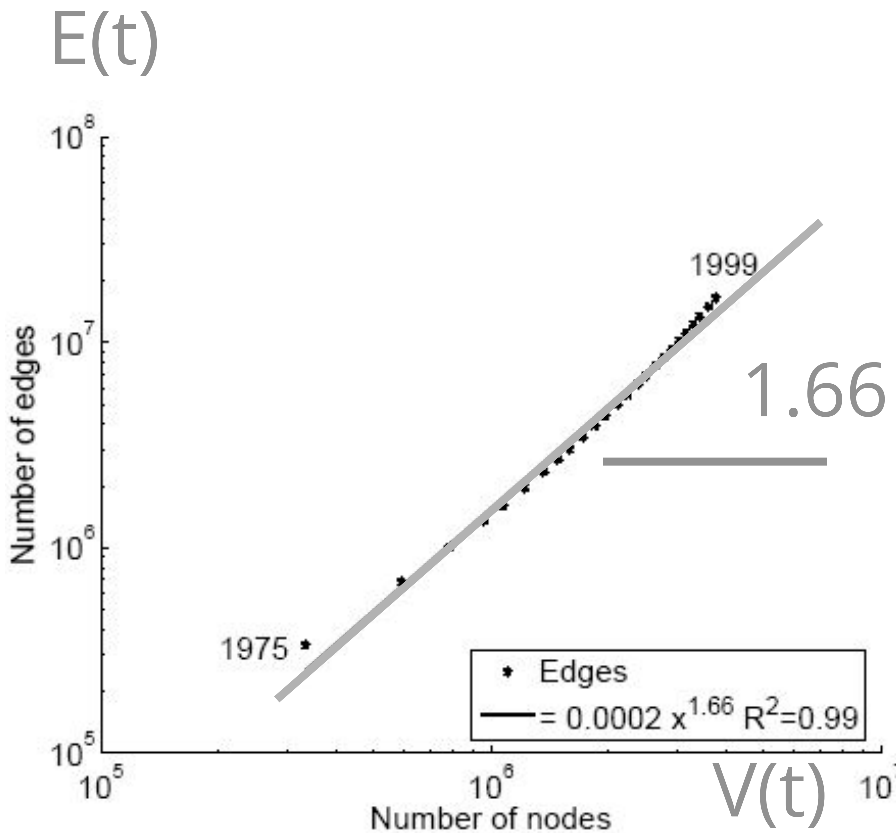
- Densification exponent: $1 \leq \alpha \leq 2$:
  - $\alpha = 1$: linear growth – constant out-degree
  - $\alpha = 2$: quadratic growth – clique

$E(t)$ and $V(t)$ are numbers of edges and nodes respectively at time $t$

# Densification in Real Networks



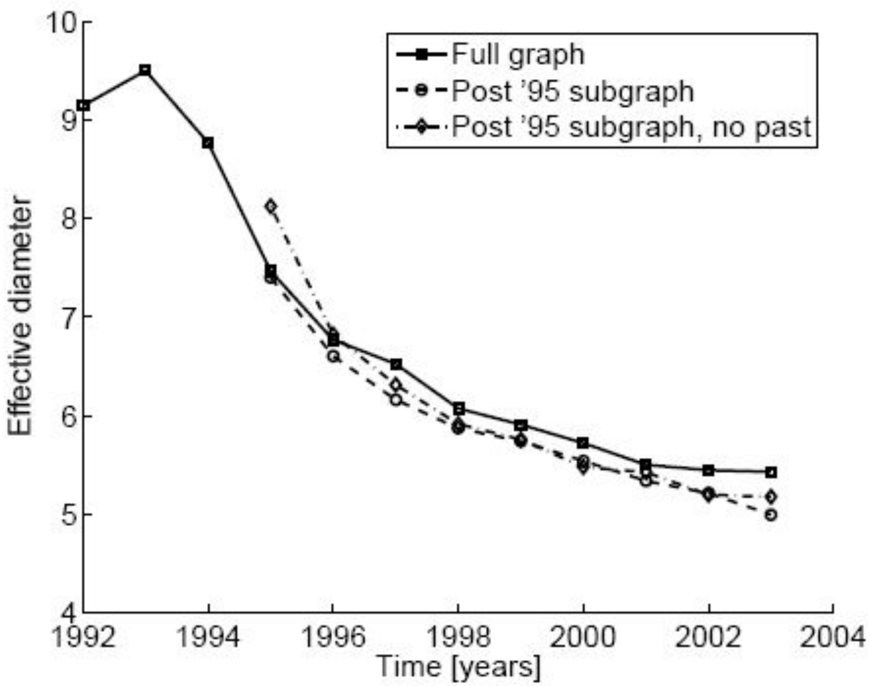

**Physics Citations**



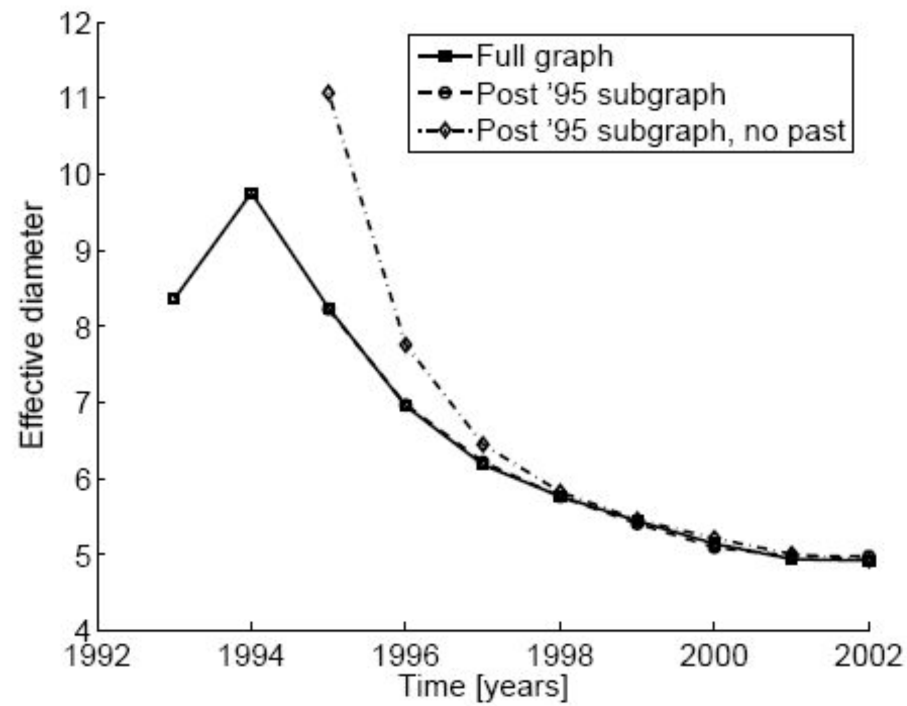

**Patent Citations**

# 3. Diameter Shrinking

- In networks diameter shrinks over time




**ArXiv citation graph**

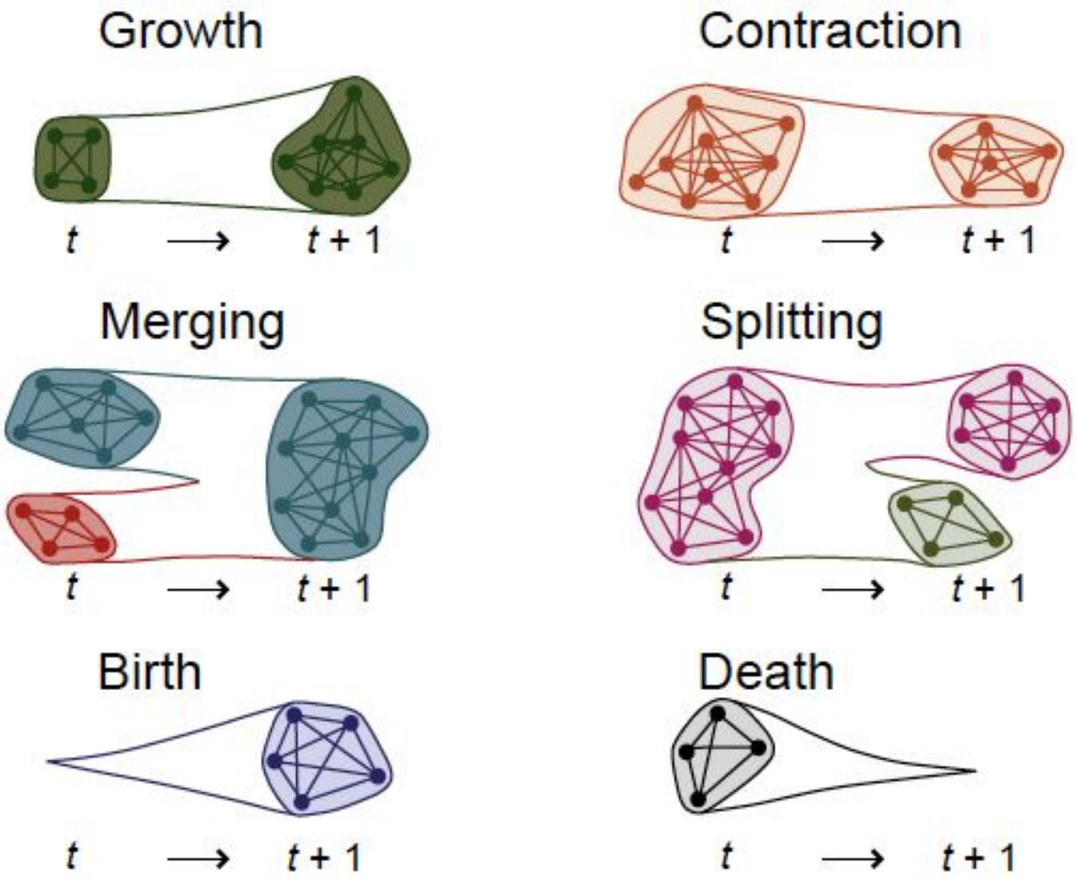



**Affiliation Network**

# How Communities Evolve?

# Community Evolution

- Communities also expand, shrink, or dissolve in dynamic networks

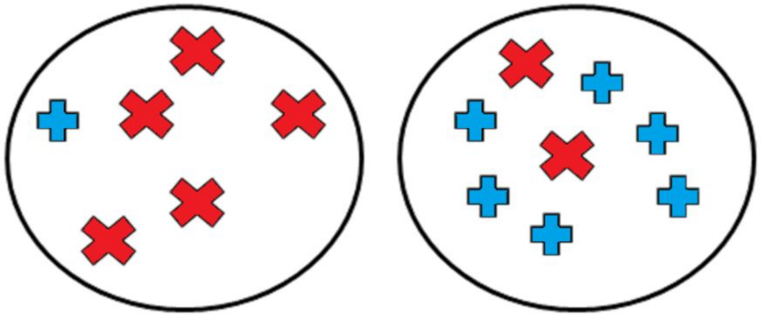


# Community Evaluation

# Evaluating the Communities

We are given objects of two different kinds (+, ×)

- **The perfect community:** all objects inside the community are of the same type



- **Evaluation with ground truth**
- **Evaluation without ground truth**

# Evaluation with Ground Truth

- When ground truth is available
  - We have partial knowledge of what communities should look like
  - We are given the correct community (clustering) assignments

- **Measures**
  - Precision and Recall, or F-Measure
  - Purity
  - Normalized Mutual Information (NMI)

# Precision and Recall

$$Precision = \frac{Relevant\ and\ retrieved}{Retrieved}$$

$$Recall = \frac{Relevant\ and\ retrieved}{Relevant}$$

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

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

**True Positive (TP) :**

- When similar members are assigned to the same communities
- A **correct** decision.

**True Negative (TN) :**

- When dissimilar members are assigned to different communities
- A **correct** decision

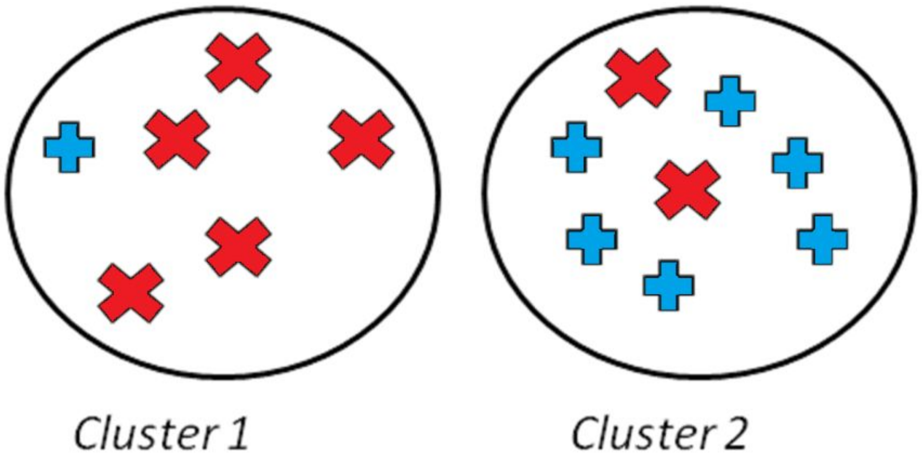
**False Negative (FN) :**

- When similar members are assigned to different communities
- An **incorrect** decision

**False Positive (FP)** :

- When dissimilar members are assigned to the same communities
- An **incorrect** decision

# Precision and Recall: Example




$$
\begin{aligned}
TP &= \binom{5}{2} + \binom{6}{2} + \binom{2}{2} = 26, \\
FP &= (5 \times 1) + (6 \times 2) = 17, \\
FN &= (5 \times 2) + (6 \times 1) = 16, \\
TN &= (6 \times 5) + (2 \times 1) = 32.
\end{aligned}
$$

$$P = \frac{26}{26+17} = 0.60$$

$$R = \frac{26}{26+16} = 0.61$$

# F-Measure

Either $P$ or $R$ measures one aspect of the performance,

- To integrate them into one measure, we can use the harmonic mean of precision of recall

$$F = 2 \cdot \frac{P \cdot R}{P + R}$$

For the example earlier,

$$F = 2 \times \frac{0.6 \times 0.61}{0.6 + 0.61} = 0.60$$

# Purity

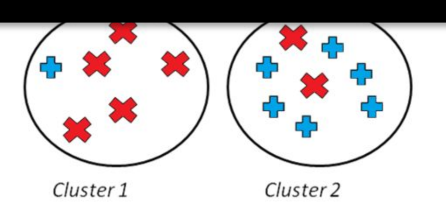
We can assume the majority of a community represents the community

– We use the label of the majority against the label of each member to evaluate the communities

Purity can be easily **tampered** by

- Points being singleton communities (of size 1); or by
- Very large communities

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




*purity is:* $\frac{6+5}{14} = 0.78$

- $k$: the number of communities
- $N$: total number of nodes,
- $L_j$: the set of instances with label $j$ in all communities
- $C_i$: the set of members in community $i$

# Mutual Information

- **Mutual information (MI).** The amount of information that two random variables share.
  - By knowing one of the variables, it measures the amount of uncertainty reduced regarding the others

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

- $L$ and $H$ are labels and found communities;
- $n_h$ and $n_l$ are the number of data points in community $h$ and with label $l$, respectively;
- $n_{h,l}$ is the number of nodes in community $h$ and with label $l$; and $n$ is the number of nodes

# Normalizing Mutual Information (NMI)

- Mutual information (MI) is unbounded
- To address this issue, we can normalize MI
- How? We know that

$$MI \leq min(H(L), H(H)),$$
$$(MI)^2 \leq H(H)H(L).$$
$$MI \leq \sqrt{H(H)}\sqrt{H(L)}.$$

- $H(.)$ is the entropy function

$$H(L) = -\sum_{l \in L} \frac{n_l}{n} \log \frac{n_l}{n}$$
$$H(H) = -\sum_{h \in H} \frac{n_h}{n} \log \frac{n_h}{n}.$$

## Normalized Mutual Information

$$NMI = \frac{MI}{\sqrt{H(L)}\sqrt{H(H)}}.$$

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

## We can also define it as

Note that $MI < 1/2(H(H) + H(L)$

$$NMI = \frac{I(H;L)}{\frac{1}{2}(H(L) + H(H))}$$

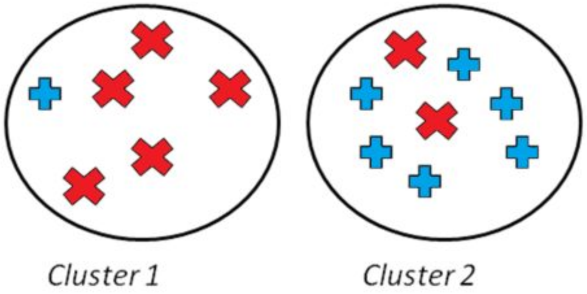
# Normalized Mutual Information

$$NMI = \frac{\sum_{h,l} n_{h,l} \log \frac{n \cdot n_{h,l}}{n_h n_l}}{\sqrt{(\sum_h n_h \log \frac{n_h}{n})(\sum_l n_l \log \frac{n_l}{n})}}$$

- where $l$ and $h$ are known (with labels) and found communities, respectively
- $n_h$ and $n_l$ are the number of members in the community $h$ and $l$, respectively,
- $n_{h,l}$ is the number of members in community $h$ and labeled $l$,
- $n$ is the size of the dataset

- **NMI** values close to one indicate high similarity between communities found and labels
- Values close to zero indicate high dissimilarity between them

# Normalized Mutual Information: Example




Found communities (H)

- [1,1,1,1,1,1, 2,2,2,2,2,2,2,2]

Actual Labels (L)

- [2,1,1,1,1,1, 2,2,2,2,2,2,1,1]

$n = 14$

| | $\mathbf{n}_h$ |
|---|---|
| h=1 | 6 |
| h=2 | 8 |

| | $\mathbf{n}_l$ |
|---|---|
| l=1 | 7 |
| l=2 | 7 |

| $\mathbf{n}_{h,l}$ | l=1 | l=2 |
|---|---|---|
| h=1 | 5 | 1 |
| h=2 | 2 | 6 |

# Evaluation without Ground Truth


(a) U.S . Constitution


(b) Sports

- **Evaluation with Semantics**
  - A simple way of analyzing detected communities is to analyze other attributes (posts, profile information, content generated, etc.) of community members to see if there is a coherency among community members
  - The coherency is often checked via human subjects.
    - Or through labor markets: Amazon Mechanical Turk
  - To help analyze these communities, one can use word frequencies. By generating a list of frequent keywords for each community, human subjects determine whether these keywords represent a coherent topic.

- **Evaluation Using Clustering Quality Measures**
  - Use clustering quality measures (SSE)
  - Use more than two community detection algorithms and compare the results and pick the algorithm with better quality measure