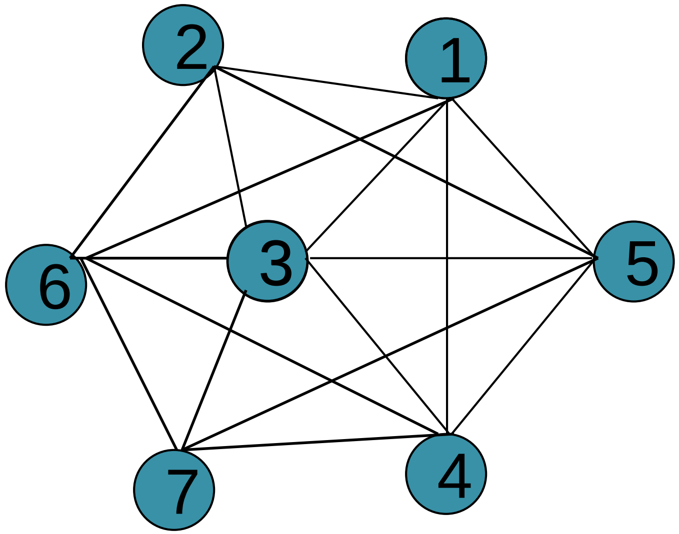
Additional approach for community detection

LCMA: Local Clique Merging Algorithm

- Observation that a maximal dense region covering vertices $\{v_1, \ldots, v_k\}$ in G_{ppi} must necessarily contain the local cliques (if any) of the vertices from $\{v_1, \ldots, v_k\}$.



{1, 3, 4, 5},

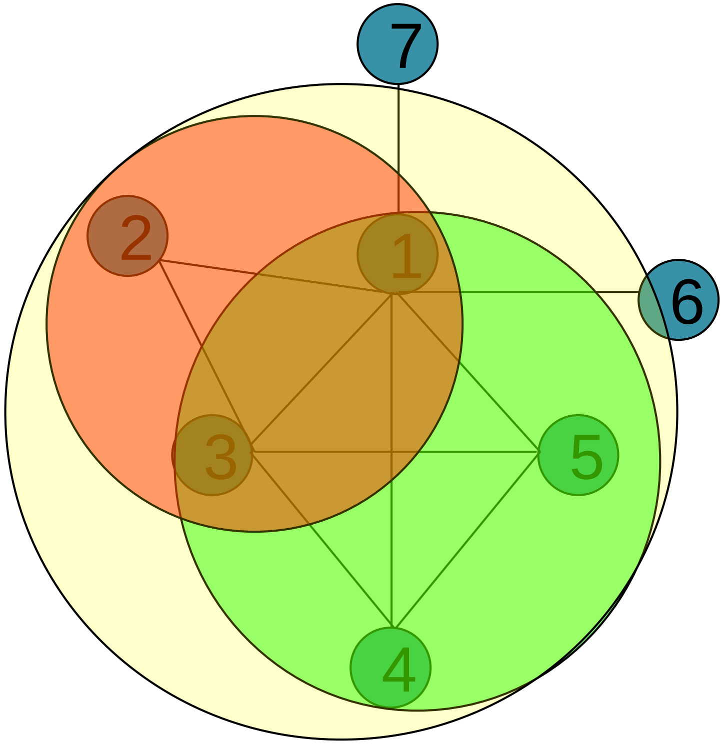
{1,2, 3, 6},

{1,2, 3, 5},

{7, 3, 4, 5},

{7, 3, 4, 6},

Local Clique Merging Algorithm (LCMA): from local cliques to maximal dense subgraphs



Two Steps of LCMA Algorithm

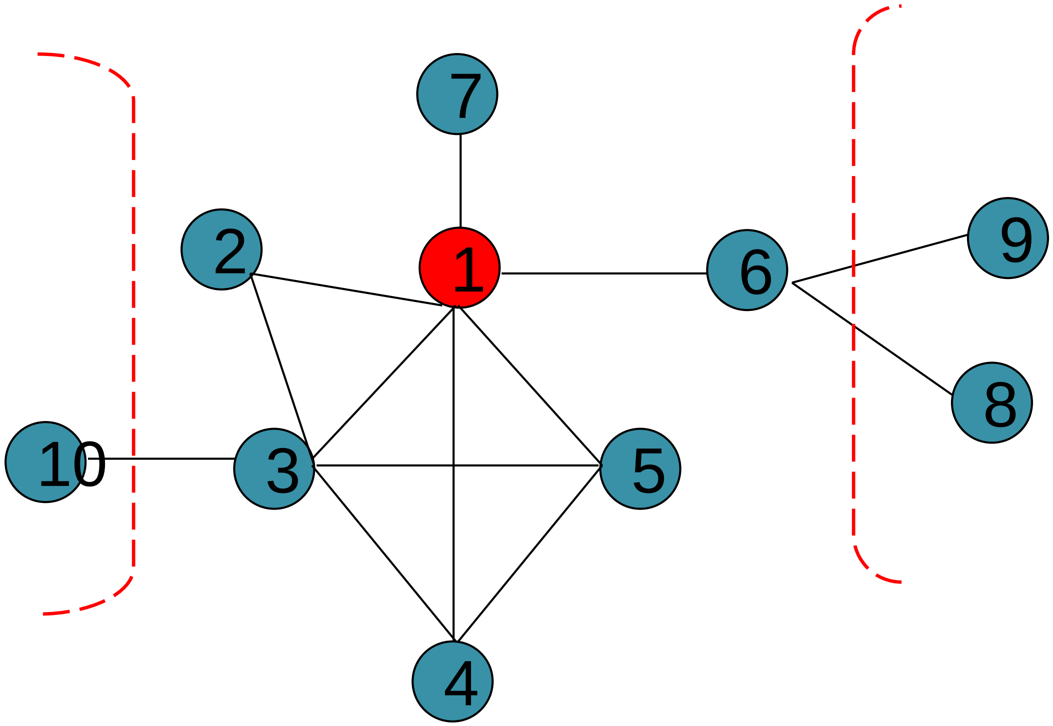
- 1. Computes the local cliques for all the vertices in G_{ppi}.
- 2.merge these local cliques to form maximal dense graphs.

Local neighborhood graph

For each vertex v_i from graph G_{ppi}, we first get its initial local neighborhood graph - namely, v_i, all its neighbors and the edges between the neighbors in graph G_{ppi}.

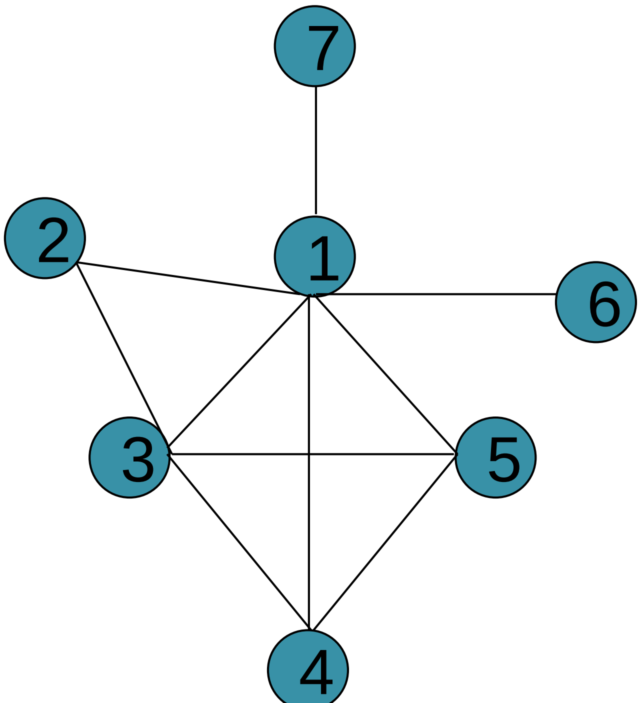
Definition 2 *Let a graph* $G = (V, E)$. *For each vertex* $v_i \in V$, *its local neighborhood graph* $G_{v_i} = (V_{v_i}, E_{v_i})$, *where* $V_{v_i} = \{v_i\} \cup \{v | v \in V, (v, v_i) \in E\}$, $E_{v_i} = \{(v_j, v_k) | (v_j, v_k) \in E, v_j, v_k \in V_{v_i}\}$.

Local neighborhood graph



LCMA 1: Mining for local cliques

- Iteratively remove the loosely connected vertices



Density=8/12=0.67

|V|=7, |E|=10.

LCMA 1: Mining for local cliques from local neighborhood graph

- For each node in its local neighborhood graph, **iteratively remove the loosely connected vertices** until the density of local neighborhood graph does not increase.

- Paper proved that **the resulting graph is a fully connected graph, namely, clique**.

LCMA 2: Merging local cliques for maximal dense neighborhoods

To detect *larger* dense graphs which can match the larger complexes better, the LCMA algorithm performs a merging step after the local cliques have been identified

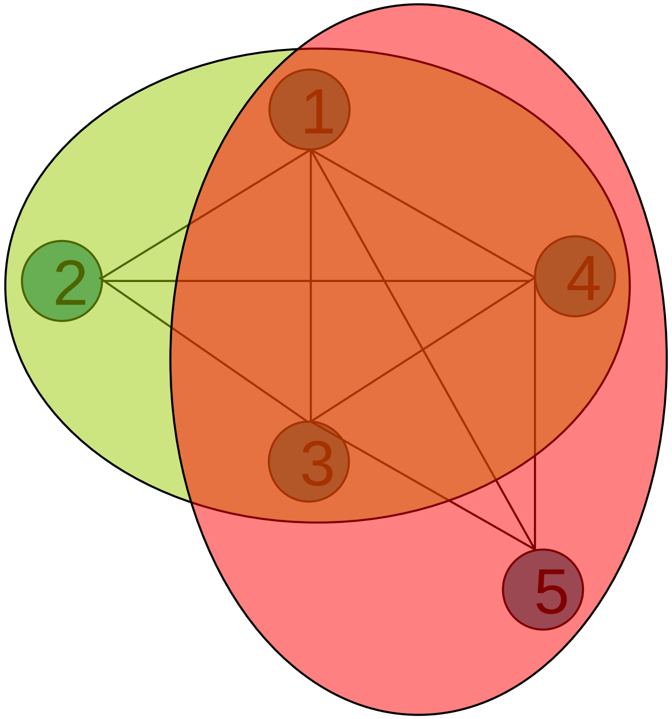
Definition 3 *Neighborhood Affinity. Given two neighborhoods (subgraphs) A and B, we define the Neighborhood Affinity NA between them as*

$$NA(A,B) = \frac{|A \cap B|^2}{|A| * |B|} \tag{10}$$

Equation quantifies the degree of similarity between neighborhoods. If two neighborhoods have larger intersection sets and similar sizes, then they are more similar and have bigger neighborhood affinity.

NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

Example of the Neighborhood Affinity



$$A = (V_1, E_1)$$

$$V_1 = \{1,2,3,4\}$$

$$E_1 = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$$

$$B = (V_2, E_2)$$

$$V_2 = \{1,3,4,5\}$$

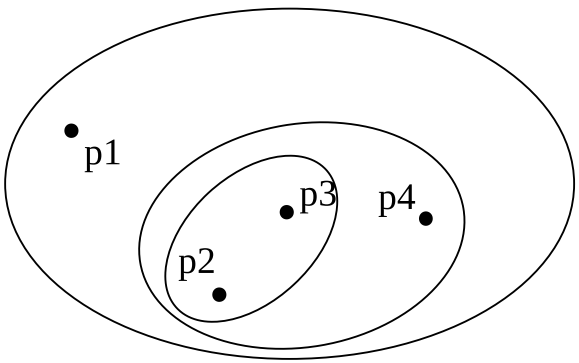
$$E_2 = \{(1,3),(1,4),(1,5),(3,4),(3,5),(4,5)\}$$

$$NA(A,B)$$

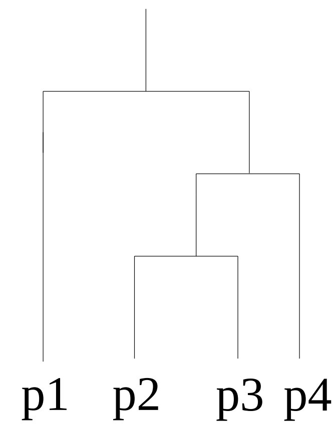
$$= \frac{|A \cap B|^2}{|A| * |B|}$$

$$= \frac{3*3}{4*4} = \frac{9}{16} = 0.5625$$

Agglomerative methods



Traditional Hierarchical Clustering

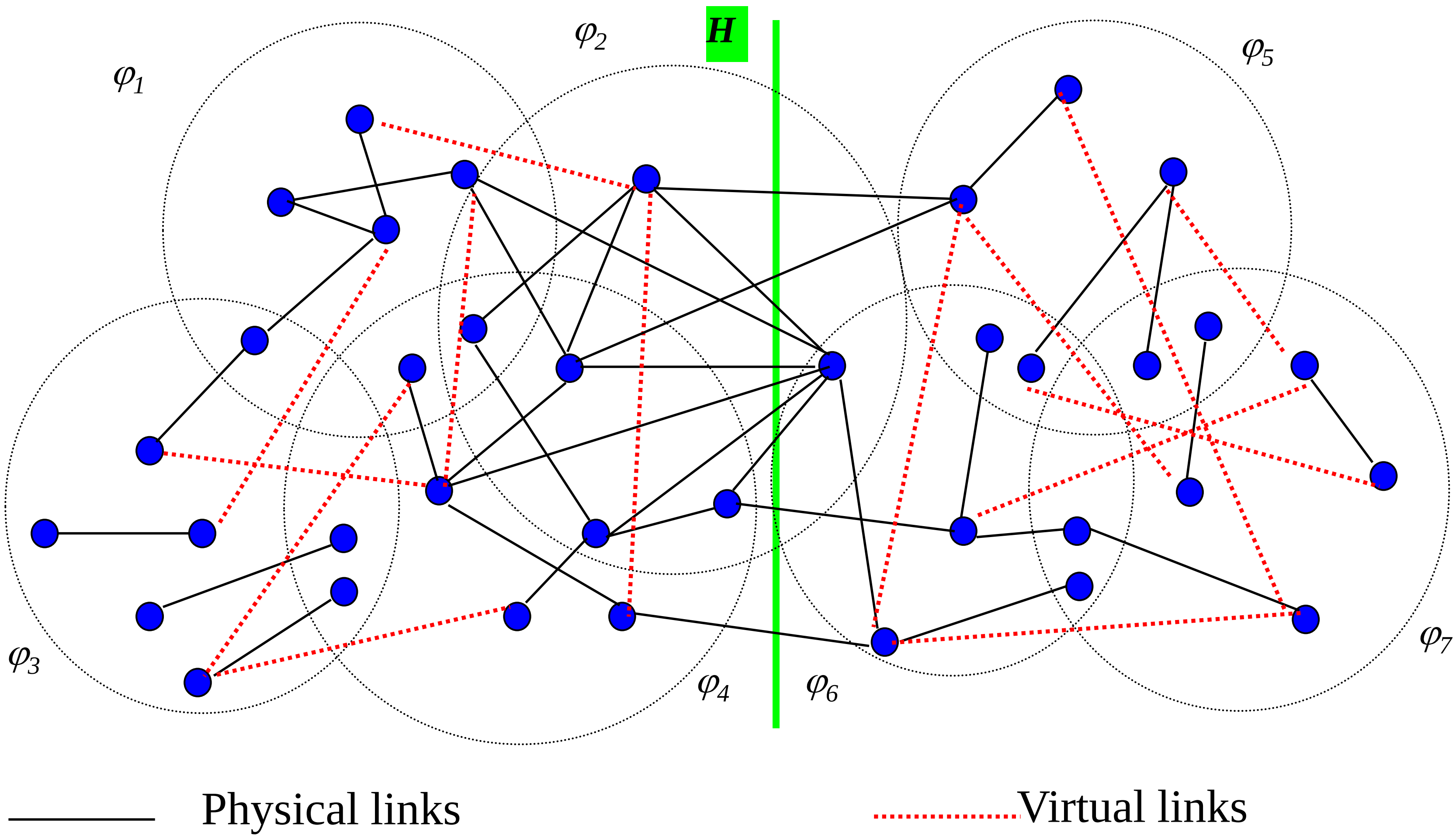


Traditional Dendrogram

Less links among the community members

- Social entities only *interact with limited community members*. As such, *there exist communities which do not have very dense connections among all its members*. This will make existing algorithms (mostly density-based) suffer.

Virtual Links enhance the connectivity among members within same communities



Virtual Links are content based

How to add virtual links

- For each node (crawl Web to get its additional information)
- Compute pair-wise content similarity
- If they are larger than certain threshold (i.e. average similarity among the known community members), then we regard them have virtual links
- Virtual links can be used to do friend recommendation

Virtual Link - Predicting friendship

- Input: two people
- Output: should they be Facebook friends?

Virtual Links

- <u>Input</u>: two people
- <u>Output</u>: should they be Facebook friends?

- **Features:**
  - **friends list**
  - **school**
  - **home town**
  - **Music**
  - **hobbies**
  - **...**

Peter, Julia, …
NUS, IIT,
Hyderabad, India
Rock
Tennis, running
…

Celia, Julia, …
NUS, Tsinghua
Beijing, China
Rock
Tennis,
…

Big SIMILARITY?

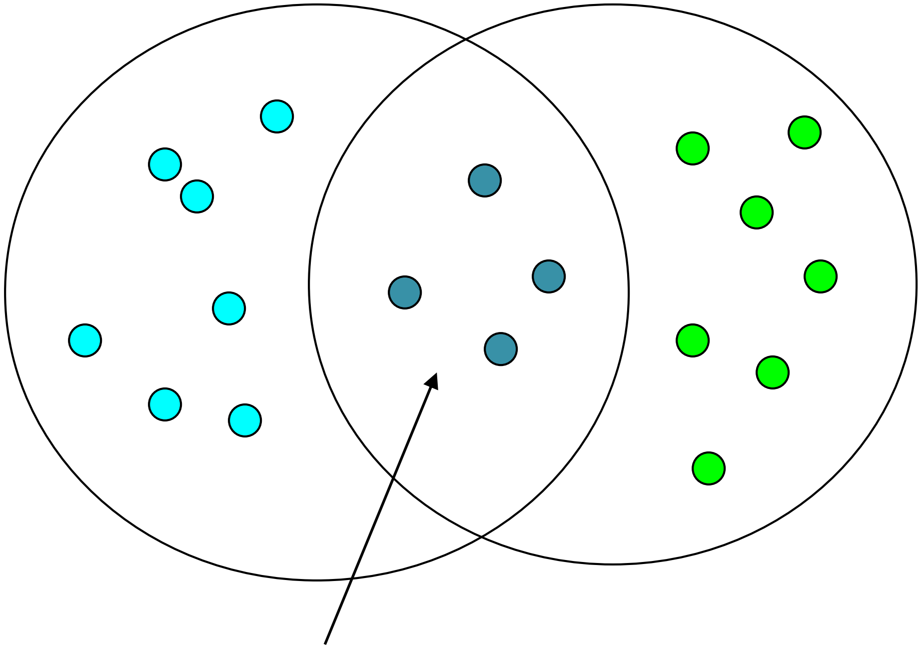
Yes!

Merge small dense graphs by computing similarity measures

- Communities can consist of the people from different small dense graphs . It is thus necessary to combine them together to form those bigger communities.
- We evaluate the similarities between graphs by the following *three different similarity measures*.

Vertex overlapping based similarity

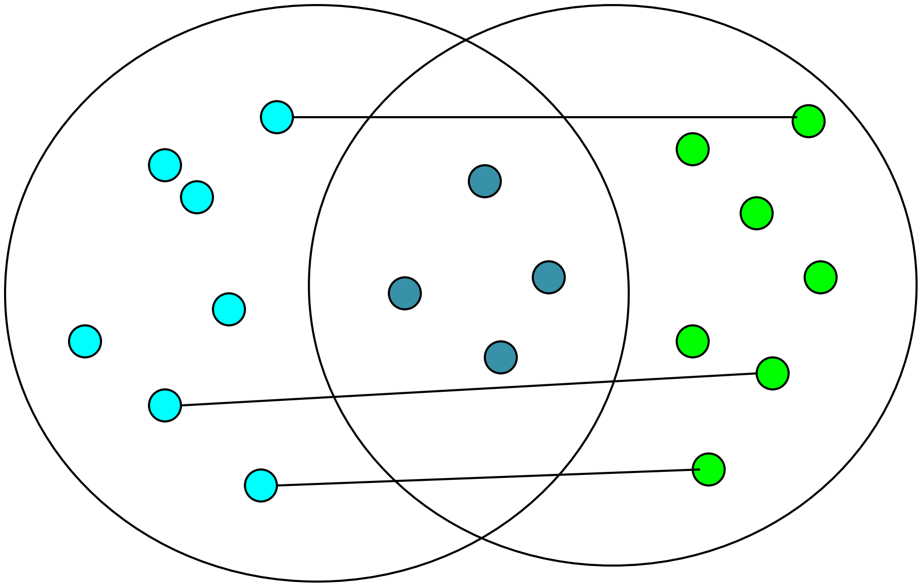
$$Vex\_sim(\varphi_i, \varphi_j) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|} / K_{vex}$$



If two graphs share a high proportion of members, then they should be combined into same community.

Physical link based similarity

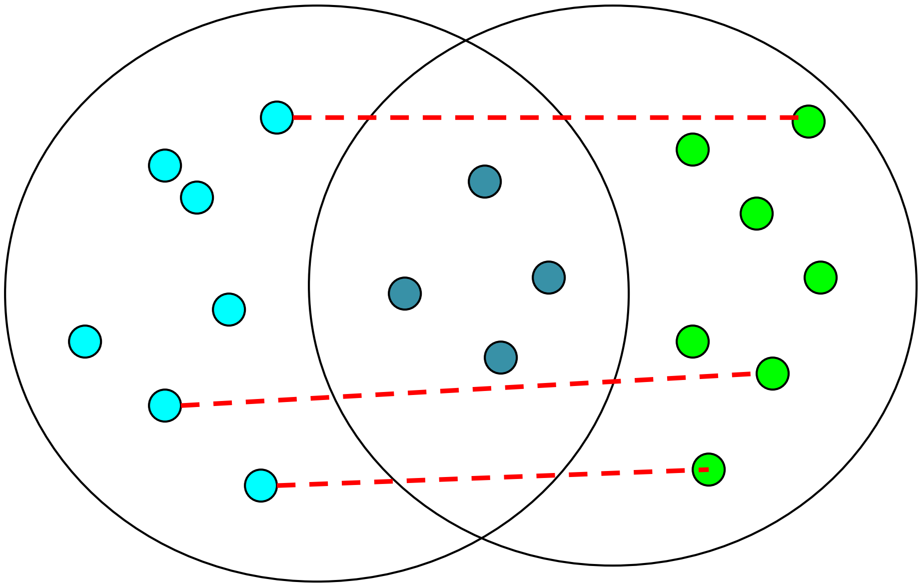
$$PL\_sim(\varphi_i,\varphi_j)=\frac{|\{(v_i,v_j)\,|\,(v_i,v_j)\in\varphi_k, k\neq i, k\neq j, v_i\in V_i\setminus V_j, v_j\in V_j\setminus V_i\}|}{|V_i\setminus V_j|*|V_j\setminus V_i|}/K_{PL}$$



Physical link based similarity: evaluates how closely the members from different graphs interact with each other.

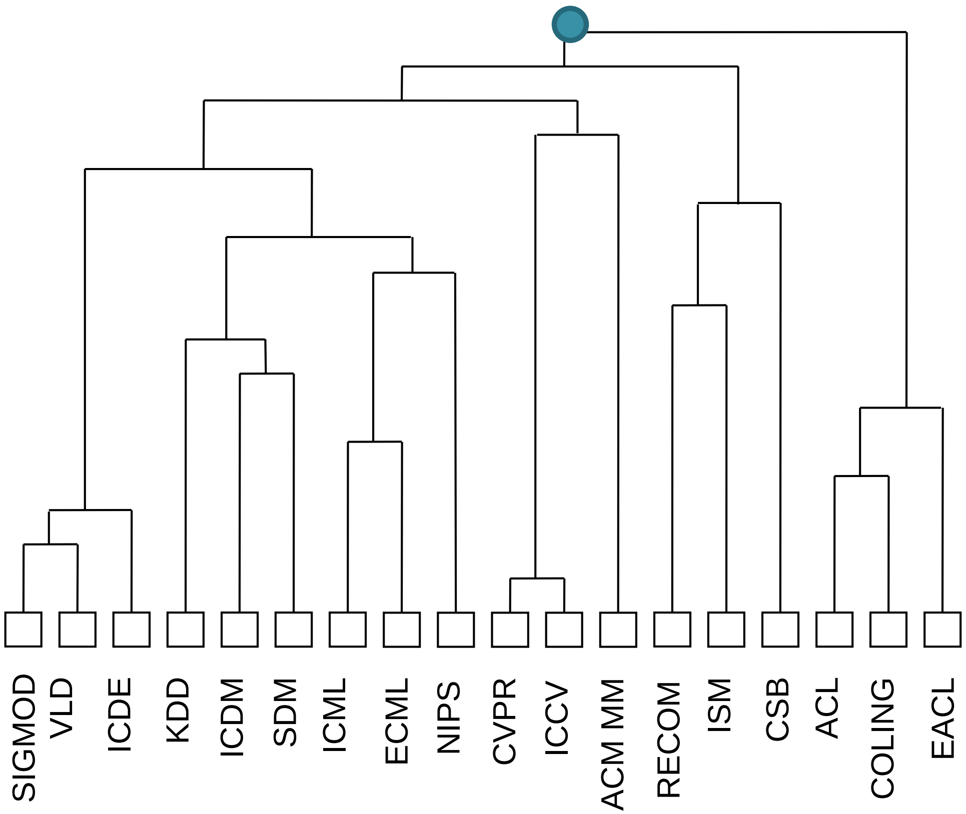
Virtual link based similarity

$$VL\_sim(\varphi_i,\varphi_j)=\frac{|\{(v_i,v_j)\,|\,v_i\in\varphi_i, v_j\in\varphi_j, consim(v_i,v_j)>\delta, v_i\in V_i\setminus V_j, v_j\in V_j\setminus V_i\}|}{|V_i\setminus V_j|*|V_j\setminus V_i|}/K_{VL}$$



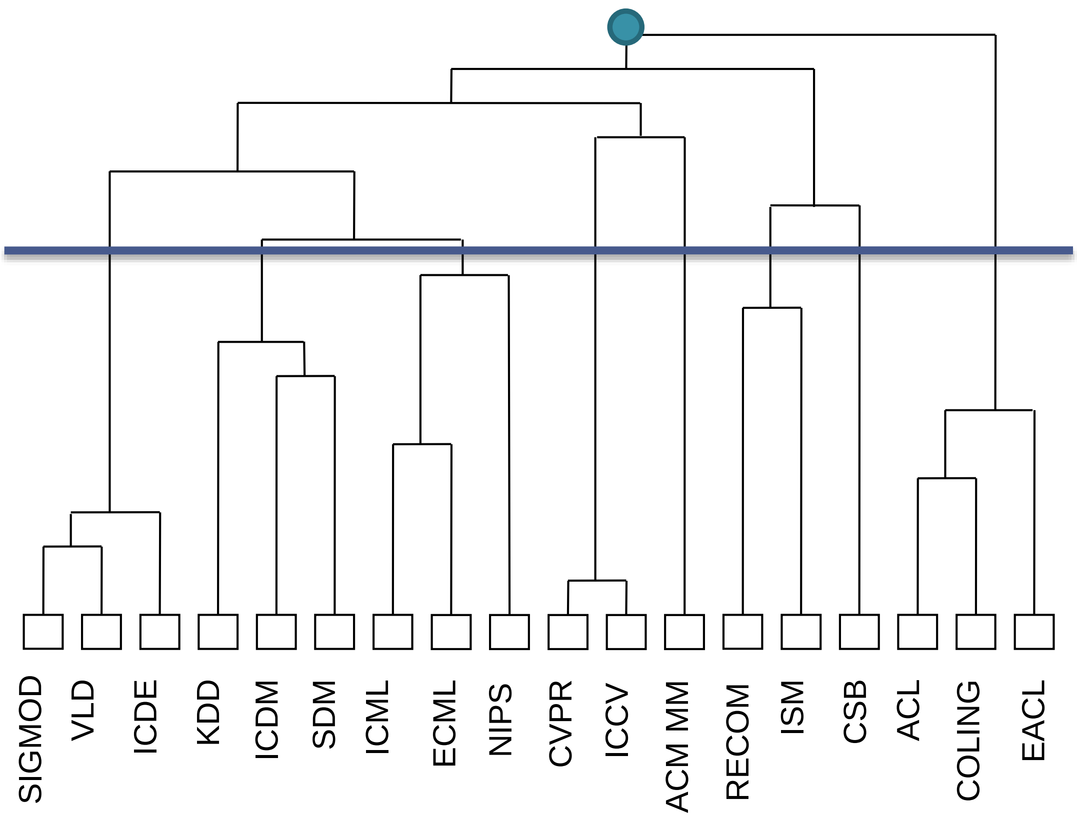
Connect those people from different graphs whose content similarity is equal to or higher than the average similarity between people within randomly selected events.

Overall Idea of ECODE Algorithm



- Hierarchical clustering approach
- Detect similar graphs in terms of *overlapping vertices* and *physical/virtual links*, and then merging them to form bigger communities

ECODE Algorithm



Automatic stopping criteria: automatically terminates when the *quality of the detected communities become maximal*

Compute the quality of the current level of the tree
Link based method

- The hierarchical clustering will result in a tree (one big community). The merging process can be stopped if *the current merging step does not improve the quality of the current level of tree*.
- Newman has proposed a quality function Q (**modularity**) to evaluate the goodness of a tree

$$Q = \sum_i (e_{ii} - a_i^{\,2})$$

where e_{ii} is the number of links in the same group connecting the vertices (intralinks) and a_i is the sum of edges from the vertices in group i to another group j (interlinks)

Compute the quality of the current level of the tree
Content based method

- There are many interactions across different communities, instead of using the physical links, we use the content/feature-based approach.

$$Q = \sum_i (cossim(i,i) - \sum_j cossim(i,j)^2)$$

- It favors a community substructure which has in overall bigger intra-similarity and less inter-similarity in terms of their topics and content.
- The algorithm can stop at a level of tree with the maximal *Q* value.

References

1. Xiao-Li Li, Chuan-Sheng Foo, Kar Leong Tew, See-Kiong Ng, "Searching for Rising Stars in Bibliography Networks", DASFAA 2009, Australia.
2. Xiao-Li Li, Soon-Heng Tan, Chuan-Sheng Foo and See-Kiong Ng. "Interaction Graph Mining for Protein Complexes Using Local Clique Merging." in *Genome Informatics, Vol. 16, No.2.* 2005.
3. Xiao-Li Li, Aloysius Tan, Philip S. Yu, See-Kiong Ng, ECODE: Event-Based Community Detection from Social Networks, DASFAA 2011, Hong Kong.