HW2 Discussion
# Task 1: Top 5

<table>
<thead>
<tr>
<th>Rank</th>
<th>Student</th>
<th>Score</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R97922137 黃安達</td>
<td>0.962854</td>
<td>[Heuristic] : Identify which nodes are colleges, finally <strong>duplicate</strong> part results</td>
</tr>
<tr>
<td>2</td>
<td>R98922069 詹承偉</td>
<td>0.938953</td>
<td>[Heuristic] : Identify which nodes are colleges</td>
</tr>
<tr>
<td>3</td>
<td>R97922055 張道軒</td>
<td>0.915075</td>
<td>[Heuristic] : Identify which nodes are colleges, finally <strong>duplicate</strong> part results</td>
</tr>
<tr>
<td>4</td>
<td>R98922060 翁睿妤</td>
<td>0.911747</td>
<td>[Heuristic] : Identify which nodes are colleges</td>
</tr>
<tr>
<td>5</td>
<td>R98922012 羅亦辰</td>
<td>0.901304</td>
<td>[Heuristic] : Identify which nodes are colleges</td>
</tr>
</tbody>
</table>
# Task 2: Top 5

<table>
<thead>
<tr>
<th>Rank</th>
<th>Student</th>
<th>Score</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R97922137 黃安達</td>
<td>1.09793</td>
<td>[Heuristic] : Based on task 1, run BFS to 2 step, finally <strong>duplicate</strong> part results</td>
</tr>
<tr>
<td>2</td>
<td>R97922028 賴弘哲</td>
<td>1.08506</td>
<td>[Heuristic] : Based on task 1, run BFS to 2 step</td>
</tr>
<tr>
<td>3</td>
<td>R98922069 詹承偉</td>
<td>1.07985</td>
<td>[Heuristic] : Based on task 1, run BFS to 2 step</td>
</tr>
<tr>
<td>4</td>
<td>R97922055 張道軒</td>
<td>1.07175</td>
<td>[Heuristic] : Based on task 1, run BFS to 2 step, finally <strong>duplicate</strong> part results</td>
</tr>
<tr>
<td>5</td>
<td>R98922060 翁睿妤</td>
<td>1.05808</td>
<td>[Heuristic] : Based on task 1, run BFS to 2 step</td>
</tr>
</tbody>
</table>
# Task 3: Top 5

<table>
<thead>
<tr>
<th>Rank</th>
<th>Student</th>
<th>Score</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R98922037 郭彥伶</td>
<td>0.301842</td>
<td>[Heuristic] : remove colleges and departments from task 1, 2, then run Newman (modularity)</td>
</tr>
<tr>
<td>2</td>
<td>R97922041 江宗憲</td>
<td>0.266287</td>
<td>[Heuristic] : remove colleges and departments from task 1, 2, then run Newman (modularity)</td>
</tr>
<tr>
<td>3</td>
<td>B96705036 黃適文 B96705045 嚴恩昞</td>
<td>0.260258</td>
<td>[EM Grouping] : directly apply EM algorithm</td>
</tr>
<tr>
<td>4</td>
<td>B96902118 何柏樟 B96902070 郭英樹 B96902113 林瑋詩</td>
<td>0.259416</td>
<td>[Heuristic + EM] : remove nodes with degree=1, then apply EM algorithm</td>
</tr>
<tr>
<td>5</td>
<td>R97725009 廖耘</td>
<td>0.257813</td>
<td>[SCAN Algorithm] : $\mu = 2$, $\varepsilon = 0.5$</td>
</tr>
</tbody>
</table>

2009/12/1 SNA09F, Prof. Shou-de Lin
Community Methods Failed

• **SCAN**
  – Hard to find decent parameter
  – Too many nodes will be assigned in multiple clusters

• **Hierarchical Clustering**
  – The matrix is too large to compute
  – Hard to find a good distance for the matrix

• **Bridge-Cut and Newman**
  – Hard to determine the threshold for modularity
  – Clusters are densely connected by keywords, so it is hard to have a good partitioning
Community Methods Failed (cont.)

- **K-means**
  - Features: PageRank, z1, z2, z2-z1

- **EM-iterative algorithm**
  - Get a little better result (compare to others)
  - Cannot handle heterogeneity of nodes
Heuristics Performs Better

• Using heuristics to identify college is very helpful for task 1 and task 2

• Based on identified colleges, departments can be derived by BFS to dept 2

• According to the given academic hierarchy, there exists different structural features which can be used to distinguish diverse types of nodes
Heuristic Features

- **College Node**
  - $CC1 = 0$, $CC2$ is low
  - $1 < Z1 < 20$, $Z2 > 1$, $Z2-Z1 > 0$
  - Shortest distances between two colleges should be smaller than 5 (College – Department – Professor – Professor – Department – College)
  - Rare connections between its second neighbors
  - If there exists connections between its second neighbors, it’s these two second neighbors must have an another common neighbor
Another Kind of Heuristics

• Guess which nodes are students
• The neighbors of students must be professors or students, others must be colleges and departments
• The next is to distinguish colleges and departments by some above heuristic features
Observation to Guess Students

- **Nodes have CC greater than 0.33 are students**
  - 1. CC(college) = 0
  - 2. CC(department) > 0 or = 0, but very small
  - 3. CC(keyword) > 0 or = 0, but very small
  - 4. CC(professor): the following observations

![Diagram of network with nodes and connections]

李曜成
Another Observation to Guess Students

(d) Professor with two keywords and two students
   $CC = 0.2$

(e) Professor with two keywords and three students
   $CC = 0.2$

李曜成
Observations to Guess Student Nodes

(clustering coefficient = 0.67 degree = 3 )

(clustering coefficient = 0.5 degree = 4 )

clustering coefficient = 1 and degree = 2.

(clustering coefficient = 0.33 degree = 5 )
### Guess Students by Degree and CC

<table>
<thead>
<tr>
<th>CC\Deg</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>S(k=3,p=1) 5149</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>.66666</td>
<td>X</td>
<td>X</td>
<td>S(k=2,p=1) 8311</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>1</td>
<td>X</td>
<td>S(k=1,p=1) 6804</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>.4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>S(k=4,p=1) 4555</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>0.33333</td>
<td>X</td>
<td></td>
<td>417</td>
<td>541</td>
<td></td>
<td>2328</td>
</tr>
<tr>
<td>0</td>
<td>S</td>
<td>12000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2009/12/1  SNA09F, Prof. Shou-de Lin
Other Interesting Methods

• After removing nodes with degree=1 and CC!=0, those nodes with **unchanged** degree could be colleges

• Remove triangles + heuristics
Other Interesting Methods

• Using Liblinear classifier to classify node types
  – Training data: http://www.cetd.com.tw
  – Features: degree, clustering coefficient, mean and variance of 1-step and 2-step neighbors
  – Test data: graph data of HW2
Common Strategy for Task 3

• Step 1. **Remove detected colleges, departments, nodes with degree=1, or other heuristics**
  – Then directly apply Newman method (modularity)
    => Get best NMI score

• Step 2. Choose nodes with **highest degree as seeds**
  (keywords could have very high degree)

• Step 3. (a) Apply existing clustering methods, or
  (b) directly perform BFS to a certain step

For clustering methods, the distance matrix e.g.,
• Length of shortest path
• The number of common neighbors
• Belong to the same department
• Jaccard similarity, etc.
Tricky Method

- **NMI Attack**: for task 1 and task 2, just duplicate final result could increase the NMI score since the NMI formula will count twice for the right answer
  - NMI is not for clustering allowing overlapping
Some Thoughts for HW2

雖然最後沒有一個結果可以用的，分數也是全班最後一名，但是我在這次作業中建立了許多如Num 類別和 heap 方法等程式碼，將來的研究都可以用得到，此外還親身體驗到這些演算法應用上可能會碰到的問題，也找到進一步的研究方向，還是覺得收穫豐富。

Life is just like a box of chocolates; eventually we couldn’t find the correct pattern in this graph. Still, we learned a lot from trying to find the patterns and had a lot of fun😊

事情才沒有這麼簡單。就在剩下兩三天 deadline，星期一還有期中考沒唸的情況下，這個作業真的只能宣告失敗了。如果給我更多時間，我一定可以把東西做好，但是現實是有限的，這次交出去的成績這麼差，期末可能要好好努力了，我可不希望最後 SNA 被當掉，呃！