SNA 2011 Fall - Assignment 3

Role and Community Detection in Heterogeneous Social Networks
Instructor: Shou-De Lin (sdlin@csie.ntu.edu.tw)
TA: Jing-Kai Lou, San-Chuan Hung, and Wei-Shih Lin

Goal

Community detection and role identification are two sort-of orthogonal goals in SNA. Community detection tasks try to identify cohesive subgroups (a.k.a. communities) whose members interact more frequently with each other than with those outside the group. Role-identification tries to group nodes that play similar role in the network.

A bunch of methods have been proposed for community detection in the last decades. However, most of the existing works focus on networks with one kind of actors, or one kind of interaction/relation of actors (a.k.a homogeneous networks). Currently, it is still a challenge to detect communities in a network with various kinds of actors and relations (a.k.a heterogeneous networks). Here you will have to tackle two tasks at the same time

1. identifying the role of nodes
2. detecting the communities in a given heterogeneous network.

Dataset

The given social network describes the relations among people, movies, publish year of movies, and places related to people and movies. There are total 28,960 nodes and 66,505 edges in the network. The information about the given network is revealed as the followings. Figure 1 shows the network topology and the size of each category.

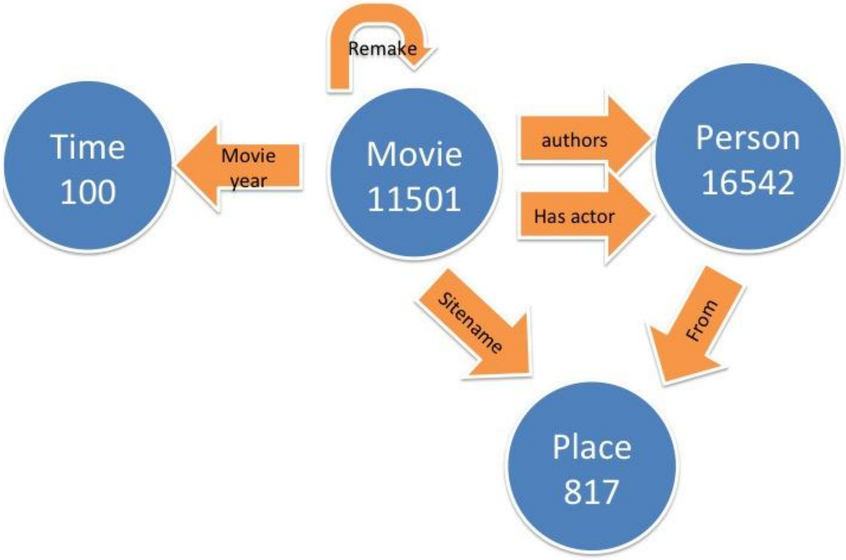


Figure 1

The movie node links to at least one time node, and the person node links to at least one movie node. The place node links to movie node or person node. There is no isolated place node.

There are some noises in dataset. Some movie nodes may have multiple time or place nodes, and some person node may have multiple place nodes. However, the ratios of noise nodes are all smaller than 3.2%.

15.2% movie nodes have ‘remake’ relationship.

The dataset we revealed is **undirected.** The link direction is reverted with a probability 0.5 from original data.

Tasks

Part A: Role Identification of nodes

Please identify the role (time, place, movie, or people) of all nodes in the given network. Upload your results in a text file with file-name <TEAM-INDEX_A.txt>.

The file should contain four partitions containing node IDs which stand for four node types: time, place, movie, and people. Make sure that the four partitions do not overlap with each other and each node is assigned one single role.

The upper bound of accuracy:
1.0

File format:
Each line contains nodes of one type (space as separator).

Example:
===[filename: TEAM-3_A.txt]===
321 23 234 …
1 10 14 1034 …
2 3 4 9 124 …
5 6 7 …
(total 4 lines)

Part B: Community Detection

In the given network, the 100 time-nodes actually stand for 100 consecutive years (i.e. 10 decades). Based on such temporal information, we can then divide this network into 10 communities. Each community shall contain 10 nodes representing 10 consecutive years, the movies released in those 10 years, as well as the corresponding places and personnel involved in those movies. Your goal is to divide the nodes into 10 overlapped (because actors and places can be involved in multiple communities) groups that represent 10 different communities over time.

Upload your result in a text with file-name <TEAM-INDEX_B.txt>. The file should contain exactly 10 lines standing for the 10 communities respectively.

The upper bound of NMI:
1.038

File format:
Each line stands for the node IDs in a community (space as separator). Note that some nodes may appear more than one line. These 10 lines are interchangeable (the order doesn't matter).

Example:

```
===[filename: TEAM-3_B.txt]===
1  3  9  10  ...
2  3  4  9  124  ...
4  23  234  ...
8  10  14  1034  ...
5  7  ...
8  12  22  …
…
(total 10 lines)
```

Evaluation Criteria:

The results will be evaluated via the methods described below.

Role identification: Accuracy for Pairwise Community Memberships

- Consider all the possible pairs of nodes and check whether they reside in the same community
- An error occurs if
  - Two nodes belonging to the same community are assigned to different communities after clustering
  - Two nodes belonging to different communities are assigned to the same community
- Construct a contingency table or confusion matrix
  - accuracy = (a + b) / (a + b + c + d) = 2 * (a + b) / (|V| (|V| - 1))

a= # of pairs that are supposed to be in the same group are predicted as in the same group

b= # of pairs that are supposed to be in different groups are predicted as in different groups

c= # of pairs that are supposed to be in the same group are predicted as in different groups d= # of pairs that are supposed to be in different groups are predicted as in the same

Community Detection: Normalized Mutual Information (NMI)

See class slides for the definition.
Reference: http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

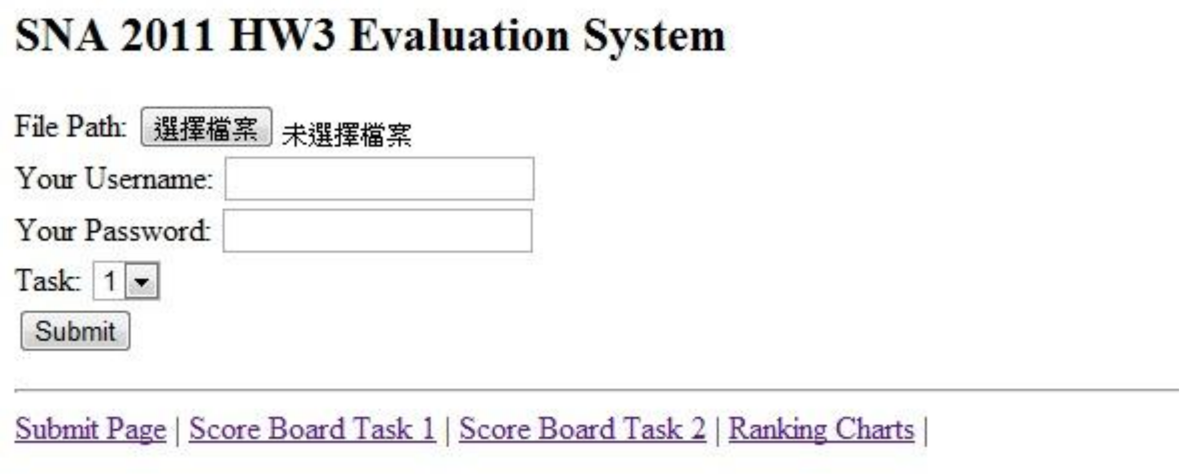
Submissions

The assignment will be held as a competition. 50% of your grade will be based on how well you do in the competition (another 50% will be based on the novelty and soundness of your algorithm written in the report). The top team(s) for each task will receive “secret presents”. Please note that you are requested to design **at least one method** that utilizes the prediction of the other task to improve the outcome of the current task. For instance, you can design a bootstrapping algorithm that utilizes the results of role-prediction to improve the quality of community detection, and then use the improved community detection results to further improve the role-prediction, and keep going until on further improvement on both sides. If you can do such successfully, we will help you to publish the results.

You need to submit not only the results but also a report (in .doc, .txt or .pdf) that clearly describes your method. Note that you don’t need to submit your code this time.

Online Leader-board

We have created an online leader-board that allows you to judge how well you does (http://mslab.csie.ntu.edu.tw/SNA2011/EvalSys3/index.php).



SNA 2011 HW3 Evaluation System

File Path: 未選擇檔案

Your Username:

Your Password:

Task: 1

選擇檔案
Submit
Submit Page | Score Board Task 1 | Score Board Task 2 | Ranking Charts |

For each task, each team is allowed to submit your results 30 times in the first two weeks of competition (10/30-11/13), and 25 times in the final week of competition (11/14-11/20). Please carefully process your upload. The leader board to show the ranking of uploaded results, and the last submission before the deadline will be considered as the final submission of your team.

Please note that for task A, we will use only 50% of the pairs (randomly chosen) for leader-board outcome. However, eventually the competition outcome is based on the accuracy of the 100% pairs. For task B, the leader-board outcome directly reflects the final outcome.