2011/10/03

SNA 2011 Fall - Assignment 2 (version 1.1)

Diffusion in Social Networks

Instructor: Shou-De Lin (sdlin@csie.ntu.edu.tw)
TA: Jing-Kai Lou, San-Chuan Hung, and Wei-Shih Lin

Goal

A social network plays a fundamental role as a medium for the spread of the idea, information, and even flu among the entities. Simulation of propagation in a social network is critical to many research areas such as viral marketing. In the class, you have studied several spreading models such as independent cascade model, linear threshold model, SIR model, and so on.
There are two tasks in this assignment.

a. First, you are asked to implement three diffusion models. We provide three real-world social networks for you to examine the models that you have implemented. Part of the nodes in the given social networks will be set as sources to diffuse. Note that you are requested to implement two diffusion models among the followings: IC, LT, SIR, and heat diffusion. Additionally, you are requested to design and implement another one which has to be not among the ones we taught in the class. It is a great chance for you to show cast your creativity.
b. Unfortunately, the virus comes. Some people in the social network are infected. Given the social networks and the diffusion algorithms you have implemented, next you have to think about how to contain the virus (avoid the spread) by blocking critical nodes from networks. The block nodes are the nodes immune from virus. That means, the block nodes will not be infected from other nodes, and therefore will not spread (diffuse) virus. By carefully choosing the block nodes, it is possible to prevent the spreading of a virus.

Dataset

Origin Data Source: http://snap.stanford.edu/data/index.html

Three Social Networks to be used (Directed Graph):

ca-GrQc: sn/ca-GrQc_clean.txt
ca-HepPh: sn/ca-HepPh_clean.txt
ca-HepTh: sn/ca-HepTh_clean.txt

The infected nodes are created based on the following procedure: we first draw 5% of the nodes (the chance a node is selected is proportion to its outdegree) as infected ones, then based on these seed nodes, we execute an IC model for 2 layers to reach 10%~13% infection percentage of nodes as the below table shows. Here we only reveal 60% of the infected nodes, some of them are seed nodes while some of them are not.

Network	p of IC model	# of initial infected nodes	Initial infected percentage
GrQc	0.1	662	12.63%
HepPh	0.02	1385	11.54%
HepTh	0.1	1340	13.57%

ca-GrQc: source_v3/GrQc_reveal.txt
ca-HepPh: source_v3/HepPh_reveal.txt
ca-HepTh: source_v3/HepTh_reveal.txt

These files can be used to test your strategy. However, please note that eventually, we will run your system with the files that contain 100% infected nodes (rather than 60%). Therefore please make sure your code is executable while we replace the above three files with the updated files.

Requirement

Part A: Model Implementation

Please implement three diffusion models (two from the following ones).

- Independent Cascading Model
  - Probability: 0.8
- Linear Threshold Model
  - Node threshold : 0.3

    - Link weight of (u ,v): 1 / in_degree(v)
- Heat Diffusion Model
    - Alpha: 1.0
    - Time: 1.0
    - Threshold: 0.005
    - P: 30
    - Initial Heat:
        - let A0 = {initial infected node}
        - N = #of nodes
        - $heat(v) = \begin{cases} \frac{N}{|a0|} & \text{if} \quad v \in a0 \\ 0 & \quad\quad v \notin a0 \end{cases}$
- SIR Model
    - Birth Rate: 0.5
    - Recover Rate: 0.5

To validate your implementations, please run your models using the source nodes we provided in the given graphs respectively, and show us the infected nodes while the diffusion converges. We will compare your results with gold standard (if exists) to determine your score. Note the results shall be in submitted as the format below:

Infected Node List Format:

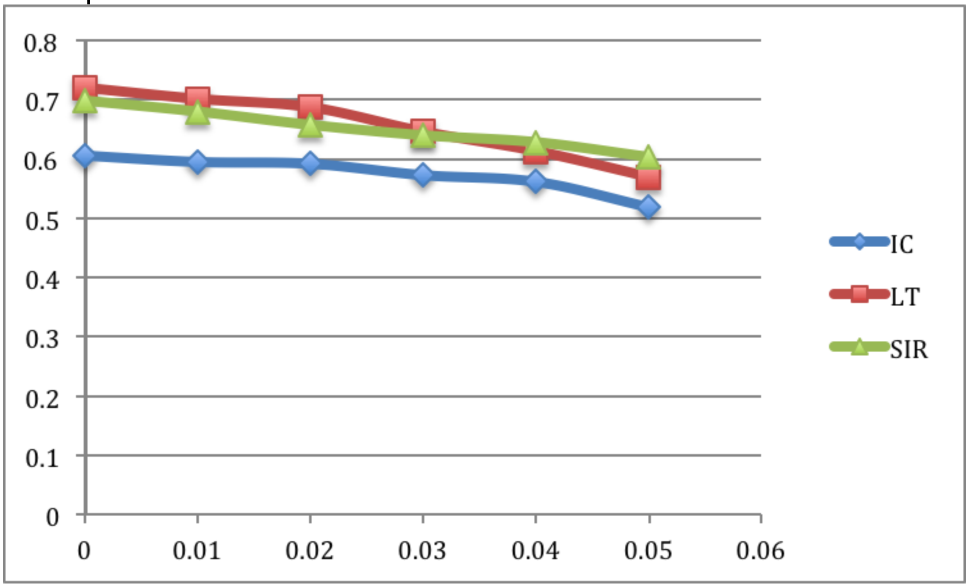
Folder Name: part_a/
File Name: {model}_{graph}.txt
File Content:
Each line contains one infected node.
Example:
File Name:
sir_ca-GrQc.txt
Content:
2345
34456
12
4342
…

Part B: The Restrain of Diffusion

In order to restrain the diffusion, you are able to set 5% (round down) nodes as block nodes. Note that these nodes cannot be the source nodes we assign. We will evaluate the performance with the methods described below.

1. The number of affected nodes while the diffusion converges (i.e. the number of affected nodes will not increase any more. The number at most equals to the graph size) for each diffusion model in each social network. Please create a table to display the numbers. In your report, please also discuss the results (it would be great but not mandatory for you to provide a theoretical justification).
2. We will evaluate the effectiveness of your blocked nodes selection by the area under the curve that shows the ratio of affected nodes with 1%, 2%, 3%, 4%, and 5% block nodes for each diffusion model in three social networks. (Note: x-axis: ratio of block nodes used; y-axis: ratio of affected nodes after converged). Pleases draw such Figure and compute the area under curve for each model in different networks. And please report your block nodes list (note that the order matters since we will start from top nodes to block).

Example:



Block Node List Format:

Folder Name: part_b/
File Name: {model}_{block ratio}_{graph}.txt
File Content:
Each line contains one block node.
Example:
File Name:
sir_0.01_ca-GrQc.txt
Content:
2345
34456

```
12
4342
…
```

Submission

Please compress your report (in PDF or DOC format) , your source code (with a README file to explain how to execute) , block nodes lists to a ZIP file, and submit it to NTU Ceiba system before **2011/10/24 (Mon) 8am**. (Note that late submission will receive at most 60 in grades). Note that eventually we will execute your code using all the infected nodes (rather than 60%). Please make sure your code is executable.

Uploaded Folder Structure

```
{Members id}_{ver.}/ eg. r00944001_r00944002_r00944003_1/
      src/  -- put your code in this folder
      part_a/  --partA results
      part_b/  --partB results
      report.pdf(doc)
      README
```