UNC – Charlotte: Machine Learning with Graphs (ITCS 6010/8010) Fall 2018

Assignment 2

Due 11:59pm ET December 23, 2020

No late days are allowed for this problem set. This problem set should be completed individually.

General Instructions

Submission instructions: You should submit your answers via Canvas.

Submitting answers: Prepare answers to your homework in a single PDF file and submit it via Canvas. Please submit your code as well. It can be a Jupyter notebook (preferred).

Questions

Signed networks [50 points]

Online networks are very useful in analyzing the social theory of structural balance and status inequality. For this in-class assignment/homework. Do the following

- Read: Signed Networks in Social Media.
- Read: Predicting Positive and Negative Links in Online Social Networks.

Download the Slashdot dataset and conduct analysis to answer the following questions:

Questions

1. Compute the number of triangles in the network.
2. Report the fraction of balanced triangles and unbalanced triangles. (assume network is undirected; if there is a sign for each direction, randomly pick one.)
3. Compare the frequency of signed triads in real and "shuffled" networks (refer slides) (assume network is undirected; if there is a sign for each direction, randomly pick one.)
4. Compute "Gen. Surprise" (assume directed signed networks) for each of the 16 types
5. Rewrite the formula for "Rec. Surprise" using the idea introduced in "Gen. Surprise"
6. Compute "Rec. Surprise" for all each of the 16 types.

The SIR Model of Disease Spreading [50 points]

In this question, you will explore how varying the set of initially infected nodes in the SIR model can affect how a contagion spreads through a network.

For the 2005 Graph Drawing conference, a data set was provided of the IMDB movie database. We will use a reduced version of this dataset, which derived all actor-actor collaboration edges where the actors co-starred in at least 2 movies together between 1995 and 2004. Please download the required file from:

- imdb_actor_edges.tsv
- imdb_actors_keys.tsv

We will be comparing our results to two other null models, the Erdős-Rényi graph and the Preferential Attachment graph, with the same number of nodes and expected degree. Please download the networks from:

- SIR_erdos_renyi.txt
- SIR_preferential_attachment.txt

Recall from lecture that under the SIR model, every node can be either susceptible, infected, or recovered and every node starts as either susceptible or infected. Every infected neighbor of a susceptible node infects the susceptible node with probability β, and infected nodes can recover with probability δ. Recovered nodes are no longer susceptible and cannot be infected again. Algorithm 1 describes for pseudo-code of this process.

1. For a node with d neighbors, what is its probability of getting infected in a given round?

2. Implement the SIR model above and run 100 simulations with $\beta = 0.05$ and $\delta = 0.5$ for each of the three graphs. Initialize the infected set with a single node chosen uniformly at random. Record the total percentage of nodes that became infected in each simulation. Note that a simulation ends when there are no more infected nodes; the total percentage of nodes that became infected at some point is thus the number of *recovered* nodes at the end of your simulation divided by the total number of nodes in the network.

   Inspecting the data, you should see that some simulations die out very quickly, while others manage to become *epidemics* and infect a large proportion of the networks. For all three graphs:

   Compute the proportion of simulations that infected at least 50% of the network; we will consider these events epidemics. To compare the likelihood of an epidemic starting across graphs, and more importantly, test whether or not the observed differences are actually significant, use pairwise Chi-Square tests. For each pair of networks, compute:

   ```
   scipy.stats.chi2 contingency([[e1, 100-e1],[e2, 100-e2]]),
   ```

   where e1 is the number of trials where ≥50% were infected in network 1 and e2 is the number of trials where ≥50% were infected in network 2. Report both the $\chi 2$ statistic and p-values. See the documentation linked above for details on interpreting the output of the function call.

Algorithm 1: Pseudo-code for simulating the SIR model on a graph $G = (V, E)$

```
Input: initial set of infected nodes I
S ← V\I        // susceptible nodes
R ← ∅          // recovered nodes
while I ≠ ∅ do
    S' ← ∅     // nodes no longer susceptible after the current iteration
    I' ← ∅     // newly infected nodes after the current iteration
    J' ← ∅     // nodes no longer infected after the current iteration
    R' ← ∅     // newly recovered nodes after the current iteration
    foreach node u ∈ V do
        if u ∈ S then
            foreach (u, v) ∈ E with v ∈ I do
                With probability β: S' ← S' ∪ {u}, I' ← I' ∪ {u}, and break for loop
        else if u ∈ I then
            With probability δ: J' ← J' ∪ {u} and R' ← R' ∪ {u}
    S ← S\S'
    I ← (I ∪ I')\J'
    R ← R ∪ R'
```

Finally, answer the following questions about the two synthetic networks:

- Does the Erdős-Rényi graph appear to be more/less susceptible to epidemics than the Preferential Attachment graph?
- In cases where an epidemic does take off, does Erdős-Rényi graph appear to have higher/lower final percentage infected?
- Overall, which of these two networks seems to be more susceptible to the spread of disease?
- Give one good reason why you might expect to see these significant differences (or lack thereof) between Erdős-Rényi and Preferential Attachment? (2–3 sentences)

We highly recommend that you debug your code with a smaller number of simulations and only run with 100 simulations once you are confident in your code. Running ~100 simulations is necessary to ensure statistical significance in some of the comparisons.