Assignment of DSAIT4310

September 12, 2025

Consider the processed "SFHH" face-to-face contact network at a conference[1], which is given in the following format: each row "a b t" denotes an undirected contact (a temporal link) between node a and b at time step t. This contact network is sampled/measured once every 20 seconds during a conference. Thus, each time step has a duration of 20s, which is though not relevant for our analysis in this assignment. We denote this temporal network as G_{data}.

A. Explore properties of network G that is aggregated over all the $T = 3259$ steps. Specifically, the aggregated network G is composed of all the nodes that have ever appeared in the dataset and any two nodes are connected by a link if they have at least a contact over the whole period $[1, T]$.

Compute the following topological properties of G in 1)-6) without considering the weight of any link and the link weight property in 7).

1) What is the number of nodes N, the number of links L, the average degree $E[D]$ and standard deviation of the degree $\sqrt{Var[D]}$?

2) Plot the degree distribution. Which network model, Erdős-Rényi (ER) random graphs or scale-free networks, could better model this network with respect to degree distribution? Why?

3) What is the degree correlation (assortativity) ρ_D? What is its physical meaning?

4) What is the clustering coefficient C?

5) What is the average hopcount $E[H]$ of the shortest paths between all node pairs? What is the diameter $H_{\max}$?

6) Has this network the small-world property? Justify your conclusion quantitatively (Hint: Lecture 2).

7) Consider further the weight of each link in G, which is the total number of contacts between the corresponding two nodes within $[1, T]$. Plot the probability density function (distribution) of the link weight (Choose the scales of the two axes and the bins/bin-size for the distribution such that the plot is insightful for interpretation). The probability density function $f_W(x)$ of the weight W of a link is defined as $f_W(x) = lim_{\Delta x \to 0} \frac{Pr[x<W\leq x+\Delta x]}{\Delta x}$, the probability that the variable (or the percentage of links whose weight) is within each range or bin $(x, x+\Delta x]$ normalized by the size of the bin Δx. Does W follow a power-law distribution? Why?

Hint: All metrics computed for the network G are recommended to put into a table.

B. Information spreading on a temporal network

We consider the following information spreading process, which is actually a simplified Susceptible-Infected model but on a temporal network. Initially, at time $t = 0$, a single node s is infected meaning that this node possesses the information whereas all the other nodes are Susceptible, thus have not yet perceived the information. Node s is also called the seed of the information. Whenever an infected node i is in contact with a susceptible node j at any time step t, the susceptible node becomes infected during the same time step and could possibly infect other nodes only since the next time step via its contacts with susceptible nodes. Once a node becomes infected, it stays infected forever. For example, assume that the seed node s has its first contact at time $t = 5$ and that contact is with node m. Although node s gets infected since $t = 0$, it infects a second

[1] C. Cattuto et al., "Dynamics of person-to-person interactions from distributed rfid sensor networks", PloS one 5(7), e11596 (2010)

node, i.e. node m only at $t = 5$ when it contacts m. Infection happens only when an infected node and a susceptible node are in contact. The number of infected nodes is non-decreasing over time.

Simulate the information spreading process on the given temporal network G_{data} for N iterations. Each iteration starts with a different seed node infected at $t = 0$ and ends at $t = T = 3259$ the last time step that the network is measured. Via the N iterations, we consider the spreading process that starts at every node $i \in [1, N]$. Record the total number of infected nodes $I(t)$ at each step t for each iteration.

8) Taking all the N iterations into count, plot the average number of infected nodes $E[I(t)]$ together with its error bar (standard deviation $\sqrt{Var[I(t)]}$) as a function of the time step t.

9) How influential a node is as a seed node could be partially reflected by the time it takes to reach/infect 80% of the nodes in the network when this node is selected as the seed node. The shorter the time is, the more influential the seed node is. Plot the influence of every node in an decreasing order (i.e., plot the time for a node to reach 80% of the nodes in the network for every node in an increasing order). Rank the influence of all the nodes and record the ranking in a vector $R = [R_{(1)}, R_{(2)}, ..., R_{(N)}]$ where $R_{(i)}$ is the node index of the $i - th$ most influential seed node and $R_{(1)}$ is the most influential node that infects 80% nodes in the shortest time. Note that you don't need to provide this vector in your report.

10) We are going to explore which nodal level network feature in the aggregated network G could well suggest the nodal influence discussed in 9). Compute the degree and strength[2] of each node in the aggregated network G and rank the importance of the nodes according to these two centrality metrics respectively. You obtain the ordered vector $D = [D_{(1)}, D_{(2)}, ..., D_{(N)}]$ and $S = [S_{(1)}, S_{(2)}, ..., S_{(N)}]$, where $D_{(i)}$ is the node having the $i - th$ highest degree and $S_{(i)}$ is the node with the $i - th$ highest strength. How precisely a centrality metric e.g. the degree could predict seed nodes' influence could be quantified by the top f recognition rate $r_{RD}(f) = \frac{|R_f \cap D_f|}{|R_f|}$ where R_f and D_f are the sets of nodes ranking in the top f fraction according to their influence and degree respectively and $|R_f| = fN$ is the number of nodes in R_f. Plot $r_{RD}(f)$ and $r_{RS}(f)$ as a function of f where $f = 0.05, 0.1, 0.15, ..., 0.5$. Which metric, the degree or the strength could better predict the influence of the nodes? Why? Attention: It is possible that many nodes have the same rank; Therefore, the choice of R_f or D_f may be not unique; In this case, $r_{RD}(f)$ should be computed as the average of 1000 iterations and within each iteration, R_f and D_f are chosen independently and randomly from their corresponding possible choices, based on which the recognition rate is derived; The same holds for the computation of any recognition rate. Consider, for example, a network with degree sequence $\{d_1 = 3, d_2 = 1, d_3 = 1, d_4 = 1\}$, $D_{f=0.5}$ has three possible choices $\{1, 2\}, \{1, 3\}$ and $\{1, 4\}$. A random choice $D_{f=0.5}$ from possible sets could also be obtained by firstly choosing the rank of nodes $2, 3$ and 4 in degree as a randomized/reshuffled vector of $[2, 3, 4]$ since node 1 has rank 1 and then selecting the 2 nodes out of the 4 with the highest rank.

11) Consider the third nodal property: the time Z when a node has its first contact in the network, and use this metric to predict nodes' influence. Plot the recognition rate $r_{RZ}(f)$ as a function of f. Compare this nodal property with degree and strength: which feature better/worse reflects how influential a node is and why?

12) How influential a node is as a seed node could also be partially reflected by the time it takes to reach/infect 20% of the nodes in the network when this node is selected as the seed node. Use this standard to rank the influence of all the nodes and record the ranking in a vector $R^* = [R^*_{(1)}, R^*_{(2)}, ..., R^*_{(N)}]$. How influential a node j is as a seed node can be also reflected by the average time $E[\tau(j)] = \frac{\sum_{i \in \mathcal{M}} \tau_i(j)}{0.8N}$ for the information starts from the seed j at $t = 0$ to reach the nodes that belong to the set $\mathcal{M}$ and the set $\mathcal{M}$ contains the $80\% \cdot N$ nodes that are reached earliest in time by the information starts at seed j. Note that $\tau_i(j)$ denotes that time when node i gets infected and j is the seed node. Use this standard to rank the influence of all the nodes and record the ranking in a vector $R' = [R'_{(1)}, R'_{(2)}, ..., R'_{(N)}]$. Consider the three rankings of influence R, R' and R^*. Evaluate the similarity between every two rankings, via plotting their recognition rate curve, as described in 10). Explain and interpret your observations.

[2] The strength of a node is the sum of the weight over all the links that are connected to the node and the weight of a link has been defined in 7).