Assignment of CS4195

February 16, 2021

Consider the Manufacturing Email data[1], which is given in the following format: each row "a b t" denotes a contact (a temporal link) between node a and b at time step t. This contact network is sampled/measured once every 1 second. Thus, each time step has a duration of 1s, 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 $T = 57791$ 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 $[0, T]$.

Compute the following topological properties of G in 1)-7) without considering the weight of any link and the link weight property in 8).

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? 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.

7) What is the largest eigenvalue (spectral radius) λ_1 of the adjacency matrix?

8) Consider further the weight of each link in G, which is the total number of contacts between the corresponding two nodes within $[0, T]$. Plot the distribution of all the link weights (Choose the scales of the two axes and the bins/bin-size for the distribution such that the plot is insightful for interpretation). Does it 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 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 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.

[1] Radoslaw Michalski, Sebastian Palus, and Przemyslaw Kazienko. Matching organizational structure and social network extracted from email communication. In Lecture Notes in Business Information Processing, volume 87, pages 197–206. Springer Berlin Heidelberg, 2011

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 $t = T = 57791$ 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 number of infected nodes $I(t)$ over time t for each iteration.

9) 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.

10) How influential a node is as a seed node could be partially reflected by the time it takes to reach/infect 80% of the total nodes when this node is selected as the seed node. The shorter the time is, the more influential the seed node is. Using this standard to 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.

11) We are going to explore which nodal level network feature in the aggregated network G could well suggest the nodal influence discussed in 10). 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?

12) Propose another two nodal/centrality features that could possibly well predict nodes' influence. The previous two features are all based on the aggregated network. At least one feature that you are going to propose should take nodal temporal features into account. Compare the two features you proposed and the two features proposed in question 11): which feature better/badly reflects how influential a node is and why? Hint: Similar to 11) plot $r_{RM}(f)$ as a function of f for each metric M you proposed.

13) How influential a node is as a seed node can be also reflected by the average time for the information started by this seed at $t = 0$ to reach a node that belongs to the set $\mathcal{M}$ of $80\% \cdot N$ nodes that are reached earliest in time. Use this standard to rank the influence of all the nodes and record the ranking in a vector $R' = [R'_{(1)}, R'_{(2)}, ..., R'_{(N)}]$. Which metric, R, the degree or the strength, could better predict the influence ranking R' ? Hint: use the method proposed in 11); Influence standard in 13) considers the average time to infect a node in $\mathcal{M}$ whereas the influence standard in 10) considers the maximal time to infect a node in $\mathcal{M}$.

C. Influence of temporal network features on information spreading.

14) Construct the following three temporal networks. G_2 is exactly the same as G_{data} except that the time stamps describing when each temporal link (contact) appears in G_{data} are randomised in G_2. In other words, G_2 is constructed by copying all the temporal links $\{(a, b, t)\}$ from G_{data} but their time stamps t are randomly re-shuffled. The number of contacts between each node pair is the same between G_{data} and G_2. [A time stamp vector v, whose length equals the number of contacts, can be randomly reshuffled to a vector v_2 of the same length by assigning each element in v to a randomly selected position in vector v_2 while avoiding more than one elements from v assigned to the same position in v_2]. G_3 is constructed by the following steps: G_3^* has the same topology as G, which is an unweighted network. Secondly, assign the time stamps in G_{data} to the links in G_3^*, randomly. A link in G_3^* may receive more than one time stamps, meaning that the two nodes contact more than once. A link in G_3^* receives no time stamp means that there is no contact between the corresponding two nodes. G_3 is a temporal network composed of all these contacts and has the same number of contacts as G_{data}

[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 8).

and G_2.

Plot the probability density function (distribution) of the inter-arrival time of two consecutive contacts between a node pair for each of these three networks G_{data}, G_2 and G_3. Hint: For each network, collect the inter-arrival time (time difference) between every two consecutive contacts along each node pair that has at least two contacts and make the corresponding histogram. For a node pair that has 5 contacts, you can collect 4 inter-arrival times. How do the three networks differ from each other? Please interpret your observation.

15) Simulate exactly the same information spreading process on G_2 and G_3 as described in B. On each temporal network, N iterations of the spreading processes are simulated and each iteration starts at a different seed node. Plot the average number of infected nodes $E[I(t)]$ and the standard deviation $\sqrt{Var[I(t)]}$ as a function of the time step t for G_{data}, G_2 and G_3 respectively. Compare and rank the information spreading performance (e.g. prevalence or speed of the spread) on these three temporal networks. Interpret/explain your observation. E.g. which temporal network features could possibly explain the different spreading performance?