Analysing the spreading of a meme on social media

Context and project description:

Networks are very useful to model epidemics and, more generally, any kind of spreading process (for example, diffusion of ideas, news, memes, among many others). For instance, one of the simplest spreading models on networks, the Susceptible-Infected (SI) model, assumes that every node of the network can be either in the Susceptible (S) or Infected (I) state. Initially, every node is susceptible (grey nodes in Figure 1) except one that is in the I state (red node in Figure 1). At each time step, the infected node can spread the disease to its neighbours with a probability *p.* The long-term evolution emerging from this model is predictable: every node of the network will eventually get infected. However, how this is achieved highly depends on the structure of the network.

Regarding meme spreading, we can easily extend the SI model by having nodes refer to users or communities that "infect" each other through the spreading (sharing) of the meme.

In this project, we aim to explore how a hypothetical meme spreads across a real social network, the Reddit hyperlink network (described below). By using simple epidemic spreading models like SI, we aim to investigate how the network structure influences the spreading patterns. Are memes quickly spread throughout the whole network? Or, rather, is their spreading concentrated into small clusters or communities of nodes?

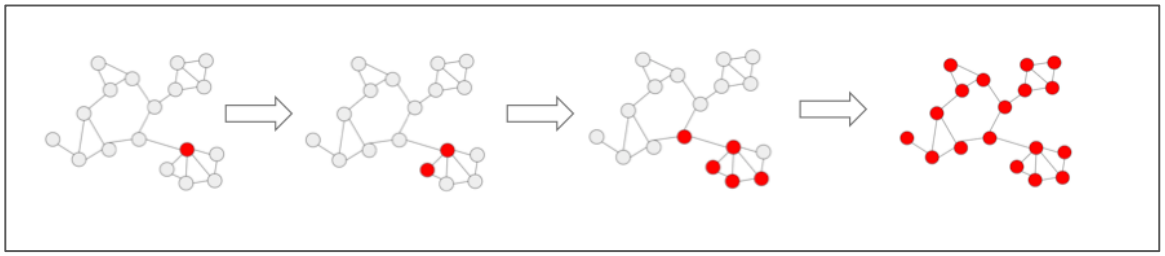

Figure 1: example of spread according to the Susceptible-Infected (SI) model

Dataset:

We will use the Reddit Hyperlink Network, provided in [1] and described below:

"The Reddit Hyperlink (RH) Network represents the directed connections between two subreddits (a subreddit is a community on Reddit). We also provide subreddit embeddings. The network is extracted from publicly available Reddit data of 2.5 years from Jan 2014 to April 2017." [2]

In the RH network, each node is a (anonymous) subreddit and each edge represents a mention (hyperlink) from one subreddit to another. Note that the RH network is a directed network where the source node corresponds to the subreddit where the hyperlink was posted, and the target node corresponds to the subreddit the hyperlink directs to.

Tasks:

Task 1:

a) Use the community detection algorithm provided in Gephi [3] to obtain communities from the network. Use 5 different "Resolution" values and show how each value is influencing on the resulting community structure (number of communities, average and standard deviation on the size of communities, among others).

b) Take one of the 5 partitions that you created, discuss how similar or dissimilar the detected communities are in this specific partition. Use basic network metrics (e.g. different centrality measures, clustering, assortativity) to guide the discussion.

Note: you have to *explain why* you chose this partition over the other 4 that you created

Task 2:

Use the SI epidemic spreading model to simulate the spreading of a meme in this network (Hint: the EoN library [4] from Python has the *fast_SIR()* function that could be helpful; also for epidemic spreading models, have a look at the lecture slides and bibliography, and [5]).
In particular you are required to:

a) Choose two different centrality measures (you have to justify your choosing; refer to the lecture slides and bibliography, and chapter 7 in [6] for a good review on centrality measures). For each one, run 10 times the SI model starting with only the most central node being infected. Then, run 10 times the same model but now with the least central node initially infected. You can use an arbitrary value of the transmission rate as long as you keep the value consistent throughout the whole analysis.

b) From the above simulations, create a plot of T_i vs D_i, where each data point corresponds to a specific node, and where T_i is the time steps it took for node i to get infected, and D_i is the distance from node i to the initially infected node.

Task 3:

Take the 5 largest communities found with your chosen partition in Task 1. For this part, you are required to conduct an analysis of how the spreading happens within and across these communities. To this end, you are asked to perform the following for each one of the 5 communities:

a) Run a SI model where the first infected node is a randomly chosen node from the community. Plot N_c vs t, where N_c is the number of nodes infected for community c (with $c=1,..,5$).

b) Create a random graph (Hint: you can use *erdos_renyi_graph() or gnm_random_graph()* from the Python NetworkX library) with the same number of edges and nodes than the original network. Run the same simulations than in (a). Compare the plots obtained in (a) and (b), and discuss the differences.

From the results above, discuss how the properties of each community (size, clustering, average centrality) might affect the spreading of the meme within, away from, or towards the community.

Task 4:

*This Task should be conducted ONLY by PG students (and **not** by UG students)*.

For this Task, you will use ONLY the largest community found in Task 1.

a) Randomly remove 5% of the network's nodes. Then, run a SI model where the first infected node is a randomly chosen node from the largest community. Plot N_c vs t, where N_c is the number of nodes infected for community c (with $c = 1, \ldots, 5$). Then repeat the process for 10%, 15%, 20% and 25%. How is the evolution changing depending on the percentage of removed nodes?
b) Repeat what you have done in (a) but now, instead of removing X% of random nodes, remove the X% of the nodes with the highest eigenvector centrality.

Compare and discuss the differences obtained between (a) and (b).
How can we identify the key nodes that are enabling a flow of memes between communities?
Discuss which other network metrics could be useful in this context.

References:

[1] Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018, April). Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference* (pp. 933-943).

[2] https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[3] Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E*, *80*(5), 056117.

[4] https://epidemicsonnetworks.readthedocs.io/en/latest/index.html

[5] Duan, Wei, et al. "Mathematical and computational approaches to epidemic modeling: a comprehensive review." Frontiers of Computer Science 9.5 (2015): 806-826.

[6] Newman, M. (2018). *Networks*. Oxford university press