Bertinoro International Spring School 2022

Opinions and conflict in social networks: Models, computational problems and algorithms

Assignment

Aristides Gionis, KTH

Deadline: May 10, 2022

Course website: `https://aris-gionis.github.io/biss-2022-ocsn/`

Instructions

To pass the course you should answer correctly (*i*) **one** of the questions 1, 2, or 3, (*ii*) **both** questions 4 and 5, and (*iii*) the **first three parts** of the programming project.

Answering more questions will lead to an opportunity for higher grade, if this is supported by your University, better understanding of the topic, and (hopefully) greater satisfaction.

Return the assignment by the deadline by email to Federico Cinus at `cinus@diag.uniroma1.it`. Please make sure that the subject of your email contains the substring BISS2022.

You should return two items: (1) a pdf file with all your answers and your report for the programming project; and (2) a zip file with your implementation and all scripts for the programming project **or** a link to a github project.

We strongly discourage handwritten solutions, and strongly encourage solutions edited in Latex.

For the programming project you are free to use the programming language of your choice.

If you have any question you can ask your course instructor at `argioni@kth.se`. Please make sure that the subject of your email contains the substring BISS2022.

Math / analytical questions

Question 1

1.1. Let $G = (V, E)$ be an undirected d-regular graph, let A be the adjacency matrix of G, and let $L = I - \frac{1}{d}A$ be the normalized Laplacian of G. Prove that for any vector $\mathbf{x} \in \mathbb{R}^{|V|}$ it is

$$\mathbf{x}^T L \, \mathbf{x} = \frac{1}{d} \sum_{(u,v) \in E} (x_u - x_v)^2. \tag{1}$$

1.2. Show that the normalized Laplacian is a positive semidefinite matrix.

1.3. Assume that we find a non-trivial vector $\mathbf{x}_*$ that minimizes the expression $\mathbf{x}^T L \, \mathbf{x}$. First explain what non-trivial means. Second explain how $\mathbf{x}_*$ can be used as an embedding of the vertices of the graph into the real line. Use Equation (1) to justify the claim that $\mathbf{x}_*$ provides a meaningful embedding.

Question 2

Let $G = (V, E)$ be an undirected graph. Define the cut function $cut : 2^V \to \mathbb{N}$ as $cut(S) = |\{(u, v) \in E$ such that $u \in S$ and $v \in V \setminus S\}|$, for each subset of vertices $S \subseteq V$. In other words, $cut(S)$ is defined as the number of edges of the graph G that have one endpoint in S and the other endpoint in $V \setminus S$.

Prove that the $cut(\cdot)$ function is submodular.

Is the $cut(\cdot)$ function monotone? Prove or disprove your claim.

Propose an approximation algorithm for finding the maximum cut in a graph. What is the approximation factor of your algorithm? Justify your answer.

Question 3

Consider the *independent-cascade model* discussed in lectures 2 and 3. In particular, we are given a directed graph, $G = (V, E, p)$, where $p : E \to [0, 1]$ are edge probabilities, so that for each edge $(u, v) \in E$, p_{uv} represents the probability that node v adopts an action given that u adopts the action.

For a subset of *seed* nodes $A \subseteq V$, we write $\sigma(A)$ to denote the number of total nodes in the graph that adopt an action, when the seed nodes A adopt the action initially. Since $\sigma(A)$ is a random variable over the edge probabilities p, we consider its expectation $f(A) = \mathbb{E}[\sigma(A)]$, where expectation is also taken over the edge probabilities p.

Show that the *influence* function $f(\cdot)$ is submodular.

Is the function $f(\cdot)$ monotone? Prove or disprove your claim.

We want to find a set of seed nodes A of size $|A| = k$, where k is an input parameter, to maximize the expected number of nodes in a network that adopt an action. That is, we want to maximize the function f.

Propose an approximation algorithm for this *influence-maximization problem*. What is the approximation factor of your algorithm? Justify your answer.

Question 4

Propose your own measure of polarization in social media. It may use content, network structure, or both. However, it has to be a different than the methods appearing in the slides. Be creative!

Question 5

Reflect on the ethical aspects of breaking filter bubbles and maximizing information diversity in social media. In particular, is it ethical for a social-media platform to adjust its filtering and ranking algorithms in order to break filter bubbles and maximize information diversity? Shouldn't people be free to be exposed to any information they wish to receive without any algorithmic intervention? Provide some arguments for and against algorithmic interventions of this type. What is your final verdict?

Programming project

The objective of the programming project is to experiment with the ideas and methods proposed in a recent paper on opinion formation and filter bubbles. You will have the opportunity to process some of the data used in this paper and reproduce some of their results. We also ask you to discuss the paper and highlight its strengths and weaknesses. The paper is:

Uthsav Chitra and Christopher Musco. *"Analyzing the impact of filter bubbles on social network polarization."* In Proceedings of the 13th International Conference on Web Search and Data Mining, 2020.

You should produce a report describing what you did and the results you obtained. You should include the necessary visualizations and plots. Your report should be self-contained but also precise and succinct. Discussion should be in English — do not include any code, your code will be provided in the supplementary zip file.

Part 1. (No programming in this part) Discuss the paper of Chitra and Musco in terms of approach, methodology, and conclusions. Try to be critical and highlight the strong points of the paper, as well as its shortcomings and limitations.

Part 2. Obtain the two datasets used in the paper. Use the sentiment analysis tools pointed by the authors (you may need to go back to previous papers) to estimate expressed opinions, and then solve the linear system to obtain innate opinions.

Produce a visualization of the two networks where the innate opinions are shown as colors of the network nodes in a colormap.

Part 3. According to the *homophily hypothesis*, "individuals tend to associate and bond with similar others," and thus, we would expect that neighboring nodes in the graph tend to have similar innate opinions.

How would you test the homophily hypothesis in the two networks?

Part 4. Implement the algorithm described in Section 3.1 of the paper, where the network nodes update their expressed opinions and the network administrator modifies the graph structure in order to minimize disagreement, and the process continues iteratively until convergence.

Use your implementation to reproduce the Figure 2 of the paper, which indicates the emergence of filter bubbles and the increase of polarization in a network.

Discuss whether your results match the ones of the paper or whether there are any discrepancies. If the latter is the case, what would be the reasons for the discrepancies?