Homework 2, CS 6241 Spring 2019

Instructor: Austin R. Benson
Due Thursday April 18, 2019 at 11:59pm ET on CMS

Assignment

The goal of this assignment is to give you a better understanding of a numerical method for graph data and to provide additional preparation for the final project.

Similar to the first homework, this homework is open-ended, and the assignment has two parts. First, you will implement a numerical method related to the material covered in the Lectures 11–20.[1] Second, you will use your implementation to analyze a "real-world" dataset of your choosing.[2]

[1] See the course content at `http://www.cs.cornell.edu/courses/cs6241/2019sp/`.

[2] Some repositories of graph data are `https://icon.colorado.edu`, `http://snap.stanford.edu/data`, and `http://www.cs.cornell.edu/~arb/data`.

Numerical method

In lectures 11–20, we will have covered numerical methods involving

1. matrices associated with graphs—adjacency, Laplacian, random walk
2. common real-world graph properties—sparsity, clustering, small hop distances, and heavy-tailed degree distributions
3. random graph models
4. graph clustering
5. graph-based semi-supervised learning
6. representation learning for nodes
7. small graph patterns—motifs, applications, and counting
8. centrality and ranking

Pick a numerical method from these topics and implement the method. The method could be one we covered in lecture or it could be a related method that you find on your own. Include a brief description of the numerical method and why it is useful.

Just like the first homework, your implementation should not just call a library function of that method. Instead, you should implement the high-level computations (for this, you can use libraries). The Jupyter notebooks from class follow this pattern.[3] Figure 1 shows an example based on the spectral optimization of the ratio cut objective for graph clustering that we covered in class.

```
,
using DelimitedFiles, LinearAlgebra
using SparseArrays, Arpack

function read_ZKC()
    data = readdlm("zacharys-KC.txt")
    I = convert(Vector{Int64}, data[:,1])
    J = convert(Vector{Int64}, data[:,2])
    n = 34  # number of nodes
    A = sparse(I, J, 1, n, n)
    return max.(A, A')  # symmetrize
end

function RCUT_evec(A::SparseMatrixCSC{Int64,Int64})
    d = vec(sum(A, dims=1))
    L = Diagonal(d) - A  # Laplacian
    evals, evecs = eigs(I + L, nev=2, which=:SM)
    return evecs[:, 2]
end

function main()
    ZKC_graph = read_ZKC()
    v2 = RCUT_evec(ZKC_graph)
end
```

Figure 1: Code for spectral algorithm for optimizing the two-way ratio cut objective.

[3] `https://github.com/arbenson/cs6241_2019sp`

Data analysis

Next, pick a "real-world" (i.e., not synthetic) dataset on which to apply your numerical method. Provide a brief description of the dataset and explain what the numerical method reveals about the data. Provide some qualitative insight.

Preparation & submission guidelines

Typesetting. The homework should be prepared with LaTeX.

Code. Part of the assignment involves implementing a numerical method. You need to include your code in your submission, and you can easily do so using the `listings` package. This is how Figure 1 was created. You do not need to include code for the qualitative data analysis component.

Collaboration. You are encouraged to discuss and collaborate on the homework to the extent of exchanging, formulating, and discussing ideas as a group. However, you have to write your own homework submission completely on your own and also understand what you are writing. You must also list your collaborators on your homework.

Academic Integrity. I expect you to maintain academic integrity in the course. For example, follow the collaboration guidelines mentioned above and do not copy someone else's software implementation. Failure to maintain academic integrity will be penalized severely. Plagiarism is a form of academic misconduct, so make sure to provide proper citations. Cornell has a number of guidelines on plagiarism.[4]

[4] https://plagiarism.arts.cornell.edu/tutorial/index.cfm

Submission. Your homework should be submitted as a single PDF that includes the following:

1. Your name, along with names of any collaborators (if applicable).
2. A brief description of the numerical method that you are using.
3. Your code (as outlined above).
4. A brief description of data that you used.
5. Qualitative analysis of the dataset obtained via the numerical method.

Submit your PDF on CMS.[5]

[5] https://cmsx.cs.cornell.edu