Network Data Analysis Coursework 2 Instructions

This coursework asks you to analyse the road network and road events in the centre of a UK city, Leeds. The coursework consists of a series of connected tasks. In each task, you should consider and apply the ideas and techniques learnt in the module along with your technical skills. The tasks are not a set of options: **all tasks must be completed by each group**. Like Coursework 1, the submission will be a single written report in the form of a scientific paper. Instructions for typesetting and structuring this document are below.

Primary Data:

The road network of Leeds is available from OpenStreetMap, while datasets on road traffic accidents in Leeds over a few years are available from the following source:

https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents

Task A (spatial networks and planarity):

In this task, you construct and investigate the road network. You are asked to choose an area of roughly 1 square kilometre around the centre of Leeds for your analysis. You should look for an area where a significant number of the recorded road accidents occurred in the area, e.g. 300 or more total across multiple years, and show you have tested this in your report. You only need to consider roads used for driving, not walking paths or private roads (investigate the osmnx documentation for how you might do this filtering). Your report should give the coordinates of the area chosen. You should then answer the following questions:

1. What are the characteristics of this road network? Include, at least, the spatial diameter of the network, the average street length, node density, intersection density, and edge density.
2. What is the average circuitry of the network? What does this tell you about the efficiency of using roads in this area?
3. Is the network planar? Why/why not? Provide examples and argue your answer considering the conditions of planarity.

Task B (road accidents):

In this task, you should investigate the road accidents on the road network.

1. Plot the distribution of road accidents on your road network and visualise this. Aggregate across multiple years of accident data. You do not need to consider or represent when the accidents happened, only their location; but you are welcome to add information about time if you believe there is something interesting to show.
2. Investigate whether a high number of accidents on one road correlates with a high number on connecting roads. Calculate the k-function and the Moran's I values for the above spatial graph. What inferences can you draw from this analysis?
3. Investigate whether accidents happen nearer to intersections or partway along roads. Consider this as asking at what fraction of the road length away from the nearest intersection do accidents typically occur.

For the above, we suggest using the spaghetti library as shown in your lessons. The third question will require investigating the API of the library to find the relevant functions to answer the question.

Task C (Voronoi diagrams):

Despite its accidents, the city of Leeds is ideal for organising marathons. The city mayor would like to organise a day of parallel, simultaneous marathons in different parts of the city. The mayor would like to maximise the participation of citizens by organising these marathons in diverse locations of the city, dividing the city into various areas (or “cells”) so that every person can join a marathon that is close to their home. Within each of these cells, a path of exactly 42 Km is needed. Assuming that the mayor would like to organise N=10 simultaneous marathons:

1. Select the initial set of 10 cell seed points. For this, you can use several criteria, such as being far away from frequent accident roads, being close to public transport, being evenly spread, etc. (explain your choice in the report).
2. Visualise the cells yield by your selection of seed points in a Voronoi diagram. What kind of Voronoi diagram (edge planar, node network, or edge points network) is most useful for this problem, and why?
3. Find 2 or 3 cells for which you can find at least one path (or more, if possible) that is (a) exactly 42 Km long, and (b) finishes at the same point where it starts. Visualise both the cells and the found paths.
4. Try to extend the previous step to all cells. Can you find at least one such a path for every cell?
5. If for steps 3-4 there were cells with no such path, what different options could you consider to increase the number of cells that include such paths? (Hint: think about the number and location of seed points; the size of the area under consideration; etc.) Choose one of such options, repeat steps 3-4, and report the results you obtain, explaining your reasoning.

Task D (TransE, PROV, PageRank):

The mayor's office is also interested in finding an efficient way of representing the provenance of important events in the road network of Leeds, such as accidents and marathons, and how this could be used for insights.

1. How would you represent such an event using the W3C PROV provenance data model standard[1]? What would correspond to Agents, Entities, and Activities? Provide a diagram illustrating your modelling, and create a (not necessarily spatial) network with (not necessarily real) data, with at least 20 nodes in 1 single connected component.
2. Compute the PageRank value for all the nodes of this network. What do these PR values tell you about the events in the city?
3. Train and evaluate TransE, RotatE and GCN embeddings for the previous provenance network, visualise them, and evaluate them with the CoDExMedium dataset. Tune hyper-parameters to improve performance. What could these embeddings be used for from a practical point of view? What kind of problems could they help address?

Paper structuring and typesetting

Your paper must be typeset to follow the ACM SIGCHI Conference Proceedings Format, and we strongly advise you write it using LaTeX/Overleaf. A Template that you can directly edit is available here: https://www.overleaf.com/latex/templates/chi2020-proceedings/qtdvrwbtqxww

The paper must contain the following sections:

- **Title and abstract**. These summarise the full contents of your coursework in just a few words (title) and in a few sentences (abstract). Instructions on how to write a good abstract are available e.g. here: https://www.anu.edu.au/students/academic-skills/research-writing/journal-article-writing/writing-an-abstract
- **Authors**. Write down your names and affiliations (King's College London). As a footnote to your names, describe your individual contributions to the work and the division of labour. This can be statements like:

[1] https://www.w3.org/TR/prov-primer/

    - "All members contributed to this work equally"
    - "X developed the initial code, Y improved bugs and extended functionality, Z wrote the first draft of the related work"
    - Etc.
    - Make your best effort for all members to contribute an equal amount of work
- **Related work**. This must contain a description of research papers and books related to the problem you are analysing. You can start with citing the papers discussed in the slides in class, the ones listed as Supplementary reading on KEATS for topics 5-8, and by searching for related papers on Google Scholar[2]. Try to find papers that also analyse the social networks behind other collaborative projects such as Wikipedia, StackOverflow, Quora, etc. and include them. **Read** these papers, **describe** what they do and find, and answer the question: **what does the work in your analysis borrow or reuse from these papers, and what does it newly contribute?** In other words: **how is your analysis similar to this previous work, and how is it different?**
- **Methods.** Describe in your own words the tasks A, B, C and D, and explain what methodology you are going to follow to answer them. This means, only say what method, algorithm, metric, etc. you will use to solve the tasks, but do not report on the results yet—this is done in the following section. Cite other papers as appropriate when you reference these methods, algorithms, metrics, etc.
- **Results.** Report the answers to Tasks A, B, C, and D, and as succinctly as possible. You can use subsection headers to help organise your reporting. Include Figures and Tables to support your explanations. Do not include any code snippets, but **add a link to your GitHub repository.** Make sure all your code and commit history is accessible.
    - Remember that all members should contribute code and this will be checked against your GitHub commit history, which should reflect continuous updates and comments rather than bulk, one-time uploads.
    - Make sure all commits come from your King's GitHub account, not a personal account.
- **Discussion.** Discuss the results you obtain from Tasks A, B, C, and D, answering the tasks' questions.
- **Conclusion.** Wrap up your report, recalling the summary you made in the abstract and highlighting the results of your analysis according to what you have written in the discussion.

You can ignore the 'Author keywords' and 'CCS Concepts' sections of the template.

Assessment

The following criteria will be used to determine the mark for each submission:

- Demonstrated understanding of network data analysis concepts and how they can apply to the questions in the coursework tasks.
- Technical ability in using programming to tackle a data analytics problem, showing ability to research and apply data manipulation techniques as required for the problem.
- Creative thinking about the problems described in the coursework and specifically the network-related aspects.
- Clarity of explanation of what code does and why and what results mean, plus good use of visualisations and presentation.

[2] https://scholar.google.nl/

- Succinctness of reporting, i.e. conveying a lot of substance clearly in a short amount of space. Note that marks will be deducted for exceeding the page limit.
- Ability to interpret, explain and position research papers related to the tasks in the coursework.

Submission instructions

Deadline. You have approximately 4 weeks between release of the coursework and submission. The strict deadline is **11.55pm UK time on 24 April**.

Size limits. The report should be at most 8 pages in length including figures, but excluding any references, for which you have unlimited space. You are welcome to add an appendix beyond the 8 pages if you want to document more work you have done but this appendix will not be considered by the markers. The amount of space used per task may vary depending on how much you find to say on each. Follow the template format as described above (ACM SIGCHI template).

Submission format. The report should be submitted on KEATS as a single PDF document. Ensure you are clear well in advance of the deadline how you will turn your working document(s) into a single PDF file. Reports are co-authored by all students in a group. Every student must submit a copy of their group report.

Plagiarism, collusion and technical support. You are not allowed to submit anyone else's work as your own (plagiarism), which is a serious matter of misconduct. Do not copy text from other papers without appropriate academic practice (citation and reference to the paper **and** quotation marks for all text copied). Do not copy text from ChatGPT or other text generators. You and you alone are entirely responsible for the text contained in your report.

You are allowed and encouraged to discuss specific technical problems you face with the coursework and how to solve them with the rest of the class via the KEATS discussion forum. For full guidance on what is acceptable to ask the class and what must be done individually see the 'Coursework questions and collusion' page in the Assessment section of the KEATS page and feel free to ask clarifying questions in the tutorials or the discussion forum.