Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CMT224
Module Title: Social Computing
Lecturer: Dr Liam Turner
Assessment Title: Social Computing Problem Sheet
Assessment Number: 1
Date Set: 26th February 2024
Submission Date and Time: by 25th April 2024 at 9:30am
Feedback return date: 24th May 2024

If you have been granted an extension for Extenuating Circumstances, then the submission deadline and return date will be later than that stated above. You will be advised of your revised submission deadline when/if your extension is approved.

If you defer an Autumn or Spring semester assessment, you may fail a module and have to resit the failed or deferred components.

If you have been granted a deferral for Extenuating Circumstances, then you will be assessed in the next scheduled assessment period in which assessment for this module is carried out.

If you have deferred an Autumn or Spring assessment and are eligible to undertake summer resits, you will complete the deferred assessment in the summer resit period.

If you are required to repeat the year or have deferred an assessment in the resit period, you will complete the assessment in the next academic year.

As a general rule, students can only resit 60 failed credits in the summer assessment period (see section 3.4 of the academic regulations). Those with more than 60 failed credits (and no more than 100 credits for undergraduate programmes and 105 credits for postgraduate programmes) will be required to repeat the year. There are some exceptions to this rule and they are applied on a case-by-case basis at the exam board.

If you are an MSc student, please note that deferring assessments may impact the start date of your dissertation. This is because you must pass all taught modules before you can begin your dissertation. If you are an overseas student, any delay may have consequences for your visa, especially if it is your intention to apply for a post study work visa after the completion of your programme.

NOTE: The summer resit period is short and support from staff will be minimal. Therefore, if the number of assessments is high, this can be an intense period of work.

This assignment is worth 100% of the total marks available for this module. If coursework is submitted late (and where there are no extenuating circumstances):

1. If the assessment is submitted no later than 24 hours after the deadline, the mark for the assessment will be capped at the minimum pass mark;
2. If the assessment is submitted more than 24 hours after the deadline, a mark of 0 will be given for the assessment.

Extensions to the coursework submission date can *only* be requested using the [Extenuating Circumstances procedure](). Only students with *approved* extenuating circumstances may use the extenuating circumstances submission deadline. Any coursework submitted after the initial submission deadline without *approved* extenuating circumstances will be treated as late.

More information on the extenuating circumstances procedure and academic regulations can be found on the Student Intranet:

https://intranet.cardiff.ac.uk/students/study/exams-and-assessment/extenuating-circumstances
https://intranet.cardiff.ac.uk/students/study/your-rights-and-responsibilities/academic-regulations

By submitting this assignment you are accepting the terms of the following declaration:

I hereby declare that my submission (or my contribution to it in the case of group submissions) is all my own work, that it has not previously been submitted for assessment and that I have not knowingly allowed it to be copied by another student. I declare that I have not made unauthorised use of AI chatbots or tools to complete this work, except where permitted. I understand that deceiving or attempting to deceive examiners by passing off the work of another writer, as one's own is plagiarism. I also understand that plagiarising another's work or knowingly allowing another student to plagiarise from my work is against the University regulations and that doing so will result in loss of marks and possible disciplinary proceedings[1].

[1] https://intranet.cardiff.ac.uk/students/study/exams-and-assessment/academic-integrity/cheating-and-academic-misconduct

Assignment

You are tasked with analysing various datasets representing different types of social and communication behaviour. These datasets are provided as files and can be found alongside this coursework pro-forma on Learning Central. You should ONLY use the files provided as they are intentionally modified versions of public datasets[2].

Alongside the dataset files, there are 3 (THREE) IPython notebooks, named part-1.ipynb, part-2.ipynb, and part-3.ipynb, which you should solely use to complete the assignment and submit these in line with the Submission Instructions section above. The cells in each completed notebook will be ran in the order that they appear. You do not need to resubmit the dataset files.

You are required to address 16 total questions across the 3 parts. Each part is made up of 1 or 2 tasks containing multiple questions. These questions are also listed below for convenience.

For EACH question in EACH notebook:

1. Complete the cell below each question marked with "#CODE:" with the Python code needed to generate any new information you need for your answer. This information should be outputted when the cell is ran and any floating-point values should be presented to 2 decimal places unless they are less than 0.01.

2. Complete the cell below this marked with "ANSWER:" with your answer to the question, referring to the information outputted above (as well as any previous cell if needed). In doing so, briefly explain your approach and methods/measures used to answer the question and justify any choices made. Each answer cell should (ideally) be no more than 125 words.

Each question is worth 6 marks (making a total of 96/100 possible marks) and a further 4 marks (4/100) will awarded for the overall usability and readability of the notebooks submitted. Marks will be awarded using the criteria described in the Criteria for assessment section below.

You may use any Python packages installable via pip.

"%pip install <some_package>" commands should be placed in the cell below "Install Python packages (pip only)" provided at the top of each notebook.

"import <some_package>" lines for all packages required for the notebook to be ran successfully should be placed in the cell under "Import Python packages" provided at the top of each notebook.

You may add additional cells throughout the notebooks, but this should be minimised.

---

[2] Jure Leskovec, & Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection.
http://snap.stanford.edu/data

Questions (duplicated from the three notebook files)

Part 1: Social media behaviour data

Task 1 of 1

Examine the Graph Modelling Language (gml) files "socialmedia_cmt224_reply_network.gml" (reply network) and "socialmedia_cmt224_social_network.gml" (social network) which represent Twitter data between a sample of users over several days at the time of the Higgs boson particle discovery. Both networks are directed and share the same ids for nodes (anonymised Twitter users). However, the shared user ids are contained within the "label" attribute in the .gml files, not the node "id" attribute of each individual .gml file.

In the reply network, an edge from a node, u, to some other node, v, indicates that u replied to a Tweet made by v during the time period. Replies are also Tweets. Edges are weighted with the weight representing the number of times this happened over the time period.

In the social network, an edge from node u to v indicates that u follows v on the social media platform.

Using these networks, answer the following questions:

Q1. What fraction of users do not reply to or follow any other user, but have had others reply to their Tweets?

Q2. How does the topological structure of the reply network differ from the social network in terms of overall sparsity of edges between users and the number of connected groups of users?

Q3. Does the number of users a user follows in the social network correlate with the number of replies that they make?

Q4. Is a user that replies to another user's Tweet multiple times more likely to follow that user in comparison to if they only replied once?

Q5. How many users have only mutual following connections (i.e., every user they follow also follows them) and only mutual reply connections with these same users?

Part 2: Email behaviour data

Task 1 of 2

Examine the file "emails_cmt224.edgelist" which represents email behaviour at an organisation. Each line contains two numbers, u and v, separated by a blank space. Consider each number as an identifier for an individual in the organisation, with the space on each line representing that the individual, u, sent at least one email to the other individual, v, at some point. Model the data using an appropriate, directed network representation and answer the following questions:

Q1. Do the majority of individuals have a higher or lower ratio of mutual connections than average in the network?

Q2. Using the largest, strongly connected component (where at least one path exists between each individual and all others). Could the connectivity of the component be suggested to be reflective of a small world phenomenon in comparison to the typical connectivity of 10 comparative random networks?

Q3. Are occurrences of induced, connected subgraphs of 3 individuals (triads) with only mutual connections more abundant in the largest, strongly connected component than those with a mixture of asymmetric and mutual connections? What does this suggest about how mutual connections are distributed in the component?

Task 2 of 2

Examine the JSON file "emails_cmt224_departments.json" (departments file). Keys in the departments file represent individuals using the same ids as in the "emails_cmt224.edgelist" file in Part 2, Task 1 and the values represent a department id that the individual can be attributed to. Using the contents of the departments file in combination with the network in Part 2, Task 1, answer the following questions:

Q1. Using the connections that individuals have in the network, are they more likely to mix with others in their department or those with a similar number of connections?

Q2. Are all departments with 15 or more members more tightly connected amongst themselves in comparison to all individuals across the overall network irrespective of their department? Where in this context, 'more tightly connected' is defined as having more mutual AND clustered connections. In addition to answering the overall question as yes or no, provide a list of departments this is true for (if any) and not true for (if any).

Part 3: Peer-to-peer message behaviour data

Task 1 of 2

Examine the file "p2p_msg_cmt224.csv" which represents messaging behaviour between users on a messaging platform. Each row has four columns, representing a single event where a person (person_a) messaged another person (person_b) on some date (date) at some time of day (time). From this, answer the following questions:

Q1. Build a suitable network to represent social connections based on the messaging behaviour that took place in the first 28 days. In doing so, assume that one or more messages from one person to another represents a mutual underlying social connection (i.e., regardless of whether person_a messaged person_b, person_b messaged person_a, or both at some point).

Q2. Using the largest connected component of the network constructed in Task 1, Q1. What is the mean, median and the standard deviation of the differences between the maximum degree of separation of each individual and the average distance

between the individual and all others?

Q3. Build another suitable network to represent social connections based on ALL message behaviour in the dataset. In doing so, assume that one or messages from one person to another represents a MUTUAL underlying social connection (i.e., regardless of whether person_a messaged person_b, person_b messaged person_a, or both at some point). Can the social phenomenon, 'Triadic Closure', be supported for the common nodes that exist in both the network created from behaviour for the first 28 days (i.e., from Task 1, Q1) and the network built from all message behaviour?

Q4. What hypothetical, non-existent edges would need to be added to the network representing all message behaviour (i.e., from Task 1, Q3) such that a message could pass along a path from any node to any other? In doing so, aim to minimise the number of edges that would be needed as well as the longest shortest path in the network as a result.

Task 2 of 2

Using the largest connected component of the social network constructed from all data in Task 1, Q2, assume the role of an outsider with complete visibility of the network that now wishes to spread a hypothetical message such that everyone in the component would know the information it contained as quickly as possible. Assume that messages will now spread in sequential timesteps using the following mechanism. If an individual is told the message at timestep t, the individual will forward the message to all of their direct connections at timestep $t+1$. Individuals can therefore be told the message more than once. From this, answer the following questions:

Q1. If you could only select 1 individual to tell at timestep 0, what set of nodes could you select from which would result in the message being received by everyone in the fewest timesteps as possible and what would the number of timesteps be?

Q2. If you had to select any 5 individuals to tell at timestep 0, can the message be received by everyone in fewer timesteps than the single individual selection in Q1? In determining your answer, use one or more appropriate network connectivity measures, rather than an exhaustive search through every combination of nodes in the network.

Learning Outcomes Assessed

1. Analyse fundamental traits of complex networks by synthesising theoretical concepts and methodologies from graph theory.
2. Evaluate and implement computational approaches to model and visualise complex social phenomena.
3. Design and create software to investigate or support human interaction behaviour.

Criteria for assessment

Credit will be awarded against the following criteria. There are 100 marks available for this assignment. Each of the 16 questions are worth 6 marks, split between up to 3 marks for the approach and implementation and up to 3 marks for the explanations and justifications of the approach and implementation. This totals 96/100 possible marks. Marks will be awarded using the following criteria:

0 marks	1 mark	2 marks	3 marks
Unsuitable implementation that does not address the question. OR Non completion of the question.	Partially completed implementation that uses some appropriate selection of appropriate methods and measures.	Completed implementation with mostly appropriate selection and implementation of appropriate methods and measures.	Complete implementation with appropriate selection and implementation of methods and measures.
0 marks	1 mark	2 marks	3 marks
Little to no explanation of the approach taken in the implementation. Or the explanation is incorrect. OR Non completion of the question.	Partially incorrect description/explanation of the approach taken in the implementation. OR A brief description of the overall approach used in the implementation, but with missing or limited explanations of the why the specific methods or measures were used.	Some explanation of approach taken in the implementation with an explanation of why the specific methods or measures were selected in the implementation, but little to no explanation for why they are the most appropriate choice.	A clear, concise explanation and justification of the approach taken, including comparison against alternative, and potentially worse, choices of the methods or measures used where relevant.

A further 4 marks (4/100 possible marks) will be awarded for the overall usability and readability of the notebooks, using the following criteria:

0 marks	1-2 marks	3-4 marks
No notebooks are runnable without modification due to errors.	All cells in some notebooks are runnable without modification due to errors. Some or most cells are clearly formatted without excessive commented out code and white space.	All cells in all notebooks are runnable without modification due to errors. Most or all cells are clearly formatted without excessive commented out code and white space. Floating point values are presented to 2 decimal places unless the value is less than 0.01.

Your total mark for this assignment will be the sum of marks for all 16 questions plus the overall usability and readability mark.

The total mark awarded for this assessment aligns with the percentage boundaries for the following levels of attainment:

Distinction (70-100 marks)
Merit (60-69 marks)
Pass (50-59 marks)
Fail (0-50 marks)

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned on 24th May 2024 via email.

Feedback from this assignment will be useful for your Dissertation.

Submission Instructions

Description		Type	Name
Part 1 Notebook (Using the template provided on Learning Central)	Compulsory	One IPython Notebook file (.ipynb)	[student_number]-part-1.ipynb
Part 2 Notebook (Using the template provided on Learning Central)	Compulsory	One IPython Notebook file (.ipynb)	[student_number]-part-2.ipynb
Part 3 Notebook (Using the template provided on Learning Central)	Compulsory	One IPython Notebook file (.ipynb)	[student_number]-part-3.ipynb

Any code submitted will be run on a system equivalent to your University provided laptop and must be submitted as stipulated in the instructions above.

Any deviation from the submission instructions above (including the number and types of files submitted) may result in a mark of zero for the assessment or question part.

<u>Staff reserve the right to invite students to a meeting to discuss coursework submissions</u>

Support for assessment

Questions about the assessment can be asked at the beginning or end of synchronous sessions in Weeks 5-10.