Te Pūkenga | Whitireia NEW ZEALAND | WelTec

Bachelor of Information Technology

DS6501
Social Data Analytics

Assignment 2

Social Network Analysis

School of Innovation, Design & Technology

Learning Outcomes

LO	Description
3	Identify and explain the visual analytical concepts applied to large social data sets
4	Analyse and discuss current social, ethical, security and privacy issues relating to large-scale social data analytics

Due Date

23:59, 02 June 2024

Submission Details

Please upload your *zip file* containing both your *report and R code* to the designated subsection titled "Report and Code" within the "Assignment 2" section under the Assessments banner on Moodle. Additionally, to check for plagiarism, please upload your report to the subsection labeled "Similarity Report" within the same "Assignment 2" section under the Assessments banner on Moodle. Detailed guidelines for uploading files to Turnitin are also available for your reference.

Please refer to the course outline for extensions, resit/resubmit details.

Grading

- This assignment is worth 30% of the total mark of the course.
- The assignment will be marked out of 100.
- To pass a grade of 50% must be achieved.
- The assignment will be marked based on the code demonstration. If you do not understand your assignment's code and is not able to reproduce or modify it during the demonstration, you can only get 50% of your total assignment marks.

Terms and Conditions

Please ensure that you submit your assessment as ***your original work***. Answers created by large language models, like *ChatGPT*, are *strictly prohibited*. Any content generated by AI in your assessments will be treated as *plagiarism*. You are required to hand in assignments on the due date and time. Extensions of time will only be granted for students who have an acceptable documented reason for not completing the assessment by the specified due date.

Please see details of terms and conditions in the Bachelor of Information Technology handbook 2024.

Purpose

This assignment aims to visually depict and explore the dynamics of interactions among members within a London street gang. The analysis will involve delineating connections and nodes, attributing social network characteristics, streamlining the network through group filtering, identifying communities, and assessing centrality measures. The findings of the analysis should be documented within a comprehensive report, as outlined in the Tasks and Assignment Deliverables sections below.

Problem Statement

The dataset under scrutiny pertains to a London street gang that emerged in 2005. Comprising 54 young men, the gang operated in criminal activities in and around a social housing estate situated in an economically disadvantaged inner London borough. Data was sourced from anonymized records of police arrests and convictions obtained from a Borough Operational Command Unit of London's Metropolitan Police Service, covering the period from 2005 to 2009. The dataset includes demographic profiles of gang members (e.g., age, ethnicity, and ranking), alongside information on the severity of their criminal involvements.

Criminologists suggest that the majority of criminal activities occur within pairs or small groups, a phenomenon referred to as co-offending. Numerous researchers have analyzed this dataset to investigate the presence of ethnic homophily—the tendency for individuals to co-offend more frequently with others of the same ethnic background. Upon examination, it becomes apparent that the street gang as a whole exhibits ethnic diversity. Through network diagrams and community detection methodologies, you will evaluate this hypothesis and explore additional characteristics to gain insight into how members of this gang interact with one another.

Tasks

T1

a) Import the data set "StreetGangData" available at the Moodle page under assignment 2 Section. The data set consists of two files, StreetGangLinks' and 'StreetGangsNodes'. It contains the links and nodes of a network relating to a London street gang. Further details of the data attributes are available in the attributes.txt file.

b) Create an igraph object based on the data files. To avoid additional clutter within the network diagrams, you should treat links as **undirected**.

c) Inspect the attributes of the network and describe the nodes and links.

d) Plot the network using the default settings and describe the main problem with this plot in terms of its readability.

e) Plot the network again using the following attributes settings:

  - The width of the links between nodes should be set to the value of their weight attribute.

- The size of each node should equal to the Age attribute divided by 2.

Although issues still remain, describe how the network plot has been improved.

T2

a) Explore measures of centrality within the network by calculating the degree, betweenness, and closeness measures of each node.
b) Repeat the analysis performed in (**a)** to display nodes in descending order of their value for each centrality measure. Identify the top 3 nodes with the highest values for each centrality measure. Explain why identifying these gang members may be of interest to the police.

T3

a) Simplify the network by removing nodes with a degree **less than** 15. Now remove edges from this simplified network whose **weight** attribute is less than 3. Plot this adjusted network using layout option of **'layout_nicely'**. For consistency, use this layout style for each network plotted in the remainder of this assignment.
b) Briefly describe the network plotted in (**a**). Your description should discuss the groupings within the network and you should identify nodes that act as bridges between these groups. Based on this analysis and the analysis performed in **T2**, which node is likely to be the most influential street gang member.

T4.

a) Set the colour of each node in the network (using the network you created in **T1**) based on the value of the node's **'Ranking'** attribute (e.g. blue for nodes with a Ranking value of 1, red for Ranking 2 etc.).
Plot this network and briefly describe the pattern observed in terms of the placement of nodes with the same ranking. How does a gang member's ranking (assigned by the police) appear to relate to the degree of the nodes and the seriousness of co-offending with other gang members?
Include a legend to explain the colour coding.
b) Using the network you created in **T1**, re-assign the colour of each node within the whole street gang based on the **'Birthplace'** attribute. Now simplify this network so that it includes only those gang members who have served time in prison. Plot this simplified network and describe the interactions observed between gang members. Your description should discuss evidence of any grouping based on ethnicity and whether gang members of differing ethnicity interact with one another. A legend should be added to the plot which explains the colour coding you have defined.
c) Using the network created in (**b**), delete nodes where a gang member's ranking is **less than 3**. Now determine the hub score of gang members within this network. Using this network create a two-panel plot. In the first panel, plot the network where the size of each node is set to 15 times the value of the hub score. In the second panel, display the communities within the network using the **cluster_optimal()** function.

   Using these two plots identify the nodes within two communities where gang members possess higher hub scores, as compared to other communities within this network. Describe these two communities in terms of their ethnicity and also in terms of the nature of their interactions (i.e. friends, co-offenders etc.), both within each community and between these two communities.

T5.

Simplify the network you created in **T1** based on the **'Ranking'** attribute. Create **five** networks in total, one for each ranking, so that each network only contains gang members with the same ranking value. Plot each network and discuss (with justifications) whether you would generally agree that the assigned ranking value reflects the seriousness of the co-offending committed among gang members within each network.

T6.

a) Create three networks that show the criminal interactions between UK gang members and each other ethnicity i.e. West African and UK, Caribbean and UK and finally, East African and UK. Use the network created in **T1** as the initial starting point. Remove links with a weight value equal to 1 (i.e. friends), then plot each network.

b) Using the network created in (**a**) for UK and Caribbean interactions, calculate the **Authority** scores of gangs members. Set the size of each node to 10 times the value of authority score. Create a two panel plotting window and plot this network in the first panel. Now identify the communities with this network using the **cluster_optimal()** function and plot these communities within the second panel.

c) Based on the networks created in (**a)** and (**b)**, what evidence is there that supports the hypothesis of the research paper (available on Moodle) that co-offending occurs mostly among gang members of the same ethnicity? Is there evidence contrary to this hypothesis? You should ignore any isolates and nodes too small to distinguish their colour coding within the first panel plot in (**b)**.

Prepare a comprehensive analysis report documenting the findings and observations from the above tasks. Your report should present all the plots and results of the analysis generated by the functions you have used.

Assignment Deliverables

Each student is required to submit the R code and analysis report bundled as a single zip file.

1. *R Code:* An R script containing all the code used to analyse the network of gang members. The script should include comments that help to explain the tasks performed. The script should include comprehensive comments clarifying the tasks performed. Before submission, ensure to clear all objects/variables and plots from your R Studio workspace using the 'sweeping brush' icons, then rerun your script to ensure its functionality on my machine.
2. *Analysis Report*
    a. An executive summary providing an overview of the analysis.
    b. The results of each stage of the analysis performed in task T1 to task T6, including all plots and observations

c. Interpretation of analysis results, addressing the tasks outlined above.
d. A conclusion section summarizing key insights gained from analysing this network.

Marking Schedule:

Student(s):

Task	Description	Max Mark	Student Mark
1	An executive summary that provides a high-level overview of the report and highlights the major findings and implications	**10**	
2	Importing and creating igraph object with appropriate treatment of links, inspecting attributes. Plotting network with default settings and identifying readability issue, improving readability with adjustments	**15**	
3	Calculating centrality measures for each node and identifying top 3 nodes, explaining relevance to police	**10**	
4	Simplifying network, removing nodes and edges, describing resulting network and influential node	**8**	
5	Colouring nodes based on 'Ranking' attribute, describing pattern observed, analysing relation with degree. Colouring nodes based on 'Birthplace' attribute, simplifying network, describing interactions, adding legend	**15**	
6	Determining hub score, plotting network and communities, identifying nodes with higher hub scores	**8**	
7	Simplifying network based on 'Ranking' attribute, creating and plotting networks for each ranking.	**8**	
8	Creating networks for criminal interactions between UK gang members and other ethnicities, plotting networks.	**10**	
9	Calculating Authority scores, plotting network and communities, analysing evidence for research hypothesis	**8**	
10	Conclusion – summary of major findings and future directions.	**8**	
	Total	**100**	