SMM638 — Network Analytics

FINAL PROJECT

The target business analytics problem of your choice (500 words)

Goal: Through this project, I intend to propose the potential of introducing a circular economy system using the Transport for London (TFL).

Background: In a linear economy, A product goes through the standard process of manufacture, consumption, and disposal. With this process, there will be mounting waste which will either be incinerated or landfilled, neither of which is a sustainable solution for the longer term. Consider this, what if, instead of disposing, this waste could be reintroduced into the manufacturing process as a raw material. The term "Circular Economy" was coined in 1988 (with its conceptual presence dating back to 1966) describing it as "an economic system where waste at extraction, production, and consumption stages is turned into inputs"[1]. Countries have recently started adopting the same after discoveries of rising environmental issues, with Europe as the leading continent on circular economy and Netherlands, France, and Spain at the forefront of it[2].

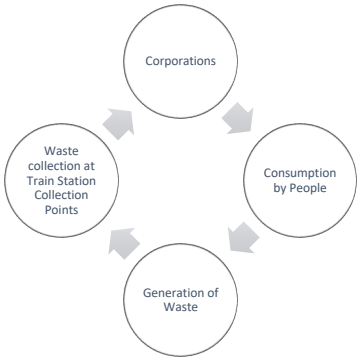
Speaking of waste, plastic products, specifically plastic bottles and single use plastic cannot be further used after they serve their purpose. Finding ways to safely dispose them is an impending challenge but most importantly, reintroducing them back into a manufacturing process would remove the scope of waste completely.

So, we now have the concept in place (circular economy) and the object that will undergo the process (waste). The third part is the focus of this analysis, i.e., the location of collection. I wanted to choose places which are not far from the depositor of waste but could be slightly far for the collector. Therefore, I chose the tube stations of London to be potential locations for said collection bins, beginning with a pilot project with few stations.

Scope of this analysis:

- The primary focus of this project would be on London and the train transport system within.
- The analysis is done from the perspective of the Government (Mayor of London).
- The underlying assumption of considering the network and footfall of each station is that the more the number of people, the more the waste generated. Therefore, waste data is not separately investigated.
- The cost of placing these bins is also not considered in this analysis.

Proposed project overview:
The devised process is such that, the corporations introduce plastic-based products into the consumption ecosystem which are single use. After consumption, the consumers dispose the waste product in a collection bin placed at one of the chosen tube stations. At a designated day and time, companies collect the waste from these stations and reintroduce the same into the manufacturing process.



[1] The Economics of Natural Resources: by Allen V. Kneese

[2] How countries are striving to build their circular economy: sustainabilitymag.com

The justification for the choice of the problem (300 words)

'There just aren't enough bins and the bins we do have aren't emptied often enough,'[3] these words have been picked up verbatim from an article discussing the prevalent waste collection issues in the United Kingdom. I was intrigued to understand the waste collection system in UK specifically after I observed a lack of waste bins at tube stations. An article[4] dated 2011 issued by the Transport for London (TfL) stated that they are consciously increasing the number of bins at stations and that the waste is sorted and recycled[5]. However, a recycling economy still does not reduce the amount of waste and only delays it[6]. Thus, to add to the green cause, I wanted to suggest a system where this waste could be reintroduced into the production process itself.

But why the tube? The tubes[7] are the second most frequently (first being transportation by bus) used mode of public transport as revealed by the Transport Statistics Great Britain report of 2021[8]. I did not choose bus stops as a potential spot for collection since the network of buses is denser and usually have public waste bins around, so it is just a matter of timely collection of the same. To start small, we need to run the pilot project in populated but easily identifiable spots. In addition, not only do tube stations do need more bins[9], but they also serve the purpose of "convenience" for the consumers since so many commute using the same through different tube lines.

In order to complete the circle, it is imperative to tie-up with corporations who require said waste as raw materials. The onus of transporting, segregating, and re-introducing the material into the manufacturing system will be on the corporations so that they take responsibility of introducing these materials in the consumption space in the first place. This would also lead to the companies thinking of alternative materials to create products.

[3] Is it just us or… is London covered in rubbish? : timeout.com
[4] More bins on London overground for a tidier tube: tfl.com
[5] TfL announcement
[6] Circular Economy v/s Linear Economy : Blog Post
[7] The words- "tube" and "train" are interchangeably used during this research.
[8] Transport Statistics Great Britain: gov.uk
[9] Here's why there are so few trash cans in London: businessinsider.com

The network dataset suited to address the chosen problem (500 words)

I joined **three datasets** for my analysis:

1. **London Train Network [Undirected, Unipartite, Unweighted network]:**
   Since I wanted to explore the network structure of the train lines in London, I searched for a dataset that consists of the connections within the network and not the geographic placements of the stations and hence, did not look for spatial data. I chose 2 datasets out of the set of datasets from the .mat file called "London Underground dataset" provided by Mr. Austin R. Benson[10]:
   a. **"Labelled_Network" [315*315 Square Matrix]:** This matrix consisted of the 315 station codes (codes that were assigned to each station)*b* represented as vertices and hence were the rows and columns' values. If one station is connected to another, then the corresponding common cell of the said row and column would have the tube line code(s) (codes that were assigned to each tube line) as an integer or list. The diagonal will be zero since stations cannot be connected to themselves.
   b. **"Station_Names" [315*1 Matrix]:** The 315 Station codes and their corresponding station names.

   Since the network connectivity must be considered, I did not include the tube lines associated with the stations and only considered the adjacency. In addition, it must be noted that only a sample of 315 stations were considered which consisted of stations from the following tube lines: 'Bakerloo', 'Central', 'Circle', 'District', 'Hammersmith & City', 'Jubilee', 'Metropolitan', 'Northern', 'Piccadilly', 'Victoria', 'Waterloo', 'Overground' and 'DLR'.
2. **Transport for London (TFL) Annual Counts for Underground, Overground and DLR (Annualised) 2021 [Node Attribute]:**
   The station entry / exit counts in this file represent the entry / exit or boarding / alighting count at each station annualised to an annual entry / exit total and rounding it off to the nearest unit (since footfall cannot be in decimals). The dataset consists of footfall data for all the stations for the year 2021 (which was further filtered for relevant data). It includes the footfall data for the London Underground, London Overground, Docklands Light Railway and TfL Rail. The sources[11] for the footfall count are:
   a. APC: Automatic passenger counters, device installed either at platforms or on trains to count the number of passengers passing through in each direction (in / out)
   b. Loadweigh: Train loading data collected and converted from the train self-weighing system
   c. Manual Count: Infrequent observations made at stations, mainly for ungated locations or for boarding / alighting flows that do not necessarily go through gatelines
   d. Mixed: Mixed method where different data sources (Loadweigh, Manual Count etc) are combined to estimate the station count
   e. Scaled: Count from a different year, normally for stations that have been manually counted, adjusted for year-on-year growth

The three datasets were joined based on the common station names (cleaned the data for bring uniformity) and then assigned the station codes to the footfall data as well to serve as a node attribute.

[10] Detecting Core-Periphery Structure in Spatial Networks by Junteng Jia and Austin R. Benson: Dataset

[11] Annual Station Counts/2021/: TfL Data

Main steps of the analysis (300 words)

The following steps were followed for the analysis:

- Chose to analyse the data at train station (node) level.
- The network is an undirected, unweighted, and unipartite network.

Network Data

1. Data Cleaning:
    a. Checked the data for any null values and found none.
    b. Filtered the footfall data to match the sample of 315 stations.
    c. Checked the data for difference in names of stations across the datasets. Discrepancies were solved by renaming the relevant footfall data to match the Network data
2. Data Pre-Processing:
    a. Prepared an adjacency matrix using the network matrix by converting any non-zero value to 1.
    b. Assigned footfall data as a node attribute to the nodes.

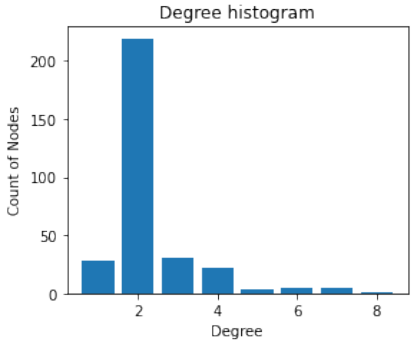
Exploratory Data Analysis using Network Theory

1. Calculated the degree distribution of the network
2. Created the matrix of the network
3. Calculated the degree (degree centrality) of each node
4. Visualised the network
5. Calculated the following centrality measures for each node:
    a. Betweenness Centrality
    b. Eigenvector Centrality
    c. Closeness Centrality
6. Calculate sum or average of all centrality measures [point 5]
7. Create a Correlation Model to check highly correlated sum v/s footfall or average v/s footfall and choose potential stations for placing collection bins.

Scope for potential analysis

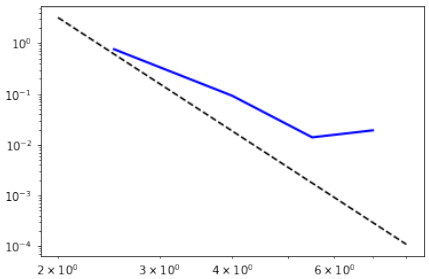
Calculating the clustering coefficient would be beneficial if the spatial data of the network is being checked to identify the cluster of stations near each other. The potential collection bins in this case could be at an approximately equidistant spot from the nodes of a cluster (e.g the approximate centroid). This is an hypothesised solution and needs to be tested before being considered as a potential solution.

The justification for each step (600 words)

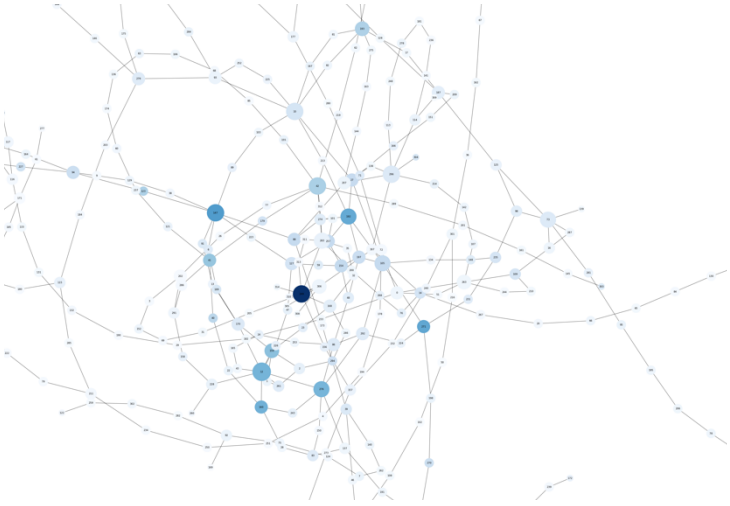


(Figure 1)

1. The data was analysed at the node-level since the goal of this analysis is to choose which train stations would be optimal for setting up the pilot program.
2. To understand the network, the degree distribution is a very good indicator of the distribution of the number of connections (degree) among the nodes. This helps in understanding the future course of action for instance, the distribution of our network shows that maximum number of nodes have a lesser degree (lesser number of connections and fewer nodes have more connections. We can see this as the length of the bar is smaller when degrees are higher. The points with high degree are called hubs and could be the potential stations wherein the program could be initiated. This also shows that the network is a highly centralised network.
3. Even though the nature of the histogram pointed towards the distribution potentially being power law (distribution of scale-free networks), graphing the same could confirm that it is very similar to a power law distribution but is actually somewhere in between a small world and scale-free network.



4. Furthermore, creating the matrix of the network helps us understand whether there is any specific structure followed by the network for example, “core-periphery”, since transportation networks usually do exhibit such a structure. In case the network follows a core-periphery structure, it would mean that there is high connectivity at the core so the pilot could be targeted at the stations that exhibit the nature of the core. This also means that the network may exhibit a weak community structure and hence, the business solution approach must not be from a community perspective.



5. Visualised the network with station codes as labels, the size of the circle showing the degree of the node and the colour showing the range of footfall. This visualisation shows the few hubs (big circles) which have high connectivity and have darker hues could be chosen for the pilot. The high degree could also correspond to the different tube lines the node has which means people may have to get down at these stations to change the line. Furthermore, the high footfall also confirms this hypothesis, and these central nodes/hubs could be the potential base

for waste collection. Higher the degree would also mean that the footfall would not be affected much if there is a strike on one of the lines.

6. A helpful centrality measure that could further confirm the importance of the nodes would be the "betweenness" centrality measure. The stations with high betweenness centrality would fall in between most of the shortest paths of the network. This means that the node would be a frequently travelled node.
7. The most influential nodes could be defined by checking the eigenvector centrality which showing the connectivity of each node with other influential nodes. The eigenvector centrality could help in identifying the highly influential central stations that are connected to other stations with high frequency of journeys.
8. The closeness centrality measure shows how close each station is to all the other stations and may support the other measures. However, as mentioned in the paper checking the correlation between centrality measures, "*Overall degree, closeness, and continuing flow centrality were strongly intercorrelated, while betweenness remained relatively uncorrelated with the other three measures (Bolland, 1988).*"[12], therefore closeness centrality could just be looked at as a metric and be dropped as a variable while creating the model to avoid repetition for further analysis.
9. Network Density measures or Network Mechanisms cannot be very helpful in this analysis since we are not trying to optimise the network rather are choosing the best points on the network for implementation of the project.
10. Central nodes would also be influential points in the city and hence, would be ideal for waste pick-up and transportation for the corporations.
11. Since we cannot be seeing the centrality measures one-by-one to compare between each other (as it will produce multiple results), the sum or the average of the centrality measures would be a great indicator of how central the node is.
12. Checking the correlation between the footfall data of the stations with the highest total/ average of the centrality measures could give us a potential set of stations to begin the pilot project with.

[12] "How Correlated Are Network Centrality Measures?" by Thomas W. Valente, Kathryn Coronges, Cynthia Lakon, and Elizabeth Costenbader

Set of possible actionable business analytics emerging from the project (300 words)

From the above-mentioned steps, we can easily identify the potential stations where the collection bins need to be placed. The focus is primarily on the central nodes to increase diffusion and efficiency. Diffusion here refers to communicating the presence of these bins (in addition to marketing campaigns) to a greater number of people about the program (through word-of-mouth) since the nodes already have so many people visiting it already.

The cost of placing and maintaining these bins must be considered as well. However, the cost of timely pick-up, transportation, sorting and re-introduction into the manufacturing process is the responsibility of the corporations. For this to effectively happen, the government must collaborate with certain corporations (for the pilot) depending on nature of waste and proximity to the manufacturing unit.

An important metric that we have not considered in this analysis[13] but should be considered in further analysis is the amount of waste already being generated and times of collections at these stations. This would give us a potential size/number of bins to be placed or increasing/decreasing frequency of collection or assignment of companies to collect at certain times. The nature of waste could also help in understanding the type of collection points, for example, if more plastic bottles are collected at station A and nearby stations, a collection bin dedicated for plastic bottles should be placed. The placement of these bins within the stations is another problem which could either be solved by network analytics using the most visited spaces in the network of all significant places in a station, or a more advanced approach would be by using computer vision on the camera recordings to identify the same.

If the pilot is successfully carried, we can use our technology to incentivise the public in being involved in this act by connecting oyster/contactless card readers to the collection points so that for every waste disposed, the card holder gets one free ride per day. Of course, this is just a suggestion and a “cost v/s benefit” analysis is required but incentivisation and increase in convenience are the key areas to look at to successfully shift to a circular economy with the help of public infrastructure and the public.

[13] Due to lack of public data regarding the same.