School of Mathematics

Mathematics project

Graph Analysis of Dynamic National Data Exchange Networks

Supervisor
Philip Greulich

Candidate
Andrius Matsenas

Academic year 2021/2022

Contents

1 Introduction **4**
1.1 Broad background 4
1.2 Motivation 4
1.3 Research objective 5

2 Subject-specific Background **6**
2.1 General background 6
2.2 Graph Theory Background 6
2.3 Network Science Background 7
2.3.1 Total number of nodes and links 7
2.3.2 Adjacency matrix of weighted networks 7
2.3.3 Degree, Distributions, Strength 8
2.3.4 Random Networks, Critical point 9
2.3.5 Sparseness 9
2.3.6 Power Law Degree Distribution 10
2.3.7 Clustering Coefficient 10
2.3.8 Clustering and Communities 11

3 Results **12**
3.1 Understanding the data 12
3.2 Real Network Characteristics 14
3.3 Hubs 17
3.4 Average and Global Clustering coefficients 19
3.5 Communities 20

4 Conclusions and Outlook **23**
4.1 Conclusions 23
4.2 Outlook 24

List of Tables

1 Overview of X-tee with key structure metrics on Jan 26, 2022 from 8AM-8PM EEST 15
2 X-tee hubs throughout Jan 26, 2022 by largest incoming-degree 18

List of Figures

1 Visual representation of 20 most active members on Jan 26, 2022 from 10am-11am EEST displayed with their member names (in Estonian) and colored by their member class. Red - Government; Blue - Commercial; Yellow - Non-Government Organisation 13

2 A heat-map for the same 20 most active members on Jan 26, 2022 from 10am-11am EEST displayed with their member codes. Horizontal axis displays outward direction and the vertical axis displays inward direction activity. Scale: $0.2 < \ln w_{ij} < 13.7$, absence of color tile indicates no link ($w_{ij} = 0$). 14

3 Log Distribution against Log of Incoming Degrees fitted with a straight line of slope $-\gamma_{\text{in}} = -1.95$ 16

4 Log Distribution against Log of Outgoing Degrees fitted with a straight line of slope $-\gamma_{\text{out}} = -1.49$ 16

5 Average and Global Clustering coefficients on a dual y-axis depicted throughout Jan 26, 2022, x-axis shows hours since midnight 19

6 50 most active members grouped into 5 optimal communities, Jan 26, 2022, 10am-11am data, each node is sized in proportion to it's strength, legend identified by hand 21

7 Interactions between the 5 optimal communities leads to a complete graph. Further analysis could make link thickness proportional to query volume. 22

1 Introduction

1.1 Broad background

Networks have diffused into everyday life through now-familiar realities like the Internet, social networks, blockchains and more. Commercial- as well as public institutions form digital bridges between each other to exchange key data and thus most data transactions could, at least in theory, be viewed and analysed as networks or graphs, in Mathematics' terms.

The *net* in "network" hints at the visual representation of such system to capture the idea of its members and their interconnectedness. As a result, there is a critical need for all sorts of insights from network analysis, both common and sophisticated. But it's been difficult to research big, often cross-border, systems due to their complexity and lack of a common standard.

Luckily for us, some such systems have consolidated under a single infrastructure where activity- and relational data could be collected and hence also analysed. Good examples include social networks like Facebook where network analysis methods have been tried and tested [1], and what I aim to cover in this project are national data exchange networks like X-Road. The X-Road protocol is an ecosystem solution that provides unified and secure data exchange between companies, government bodies, not-for-profits and other organisations [18]. It has been applied in multiple countries on a national level and thus can be considered one of the first more comprehensive overviews of a data economy. Therefore in this project, I'm exploring how to measure a network that represents the data exchange infrastructure of a whole country.

1.2 Motivation

In fact, the X-Road protocol has been applied on a national level in Estonia, Finland, Iceland and the Faroe Islands [16] therefore potentially affecting over 7 million people (the sum of the populations) and in reality even more as foreigners also reap benefits of a connected data economy.

Estonia, with a population of 1.3 million, has been relying on their instance of X-Road, called X-tee, since 2001 [4] which means these data exchange processes have become an integral part of most Estonians' day-to-day lives. Examples include instantaneous checking of a person's registered living address on bus ticket validation, seamless flowing of all personal health data between private and public health institutions, voting online, and automatic generation of annual tax forms based on income data, to name a

few use cases.

Estonia's X-tee has been collecting monitoring (activity) data since 2016 [13] and is automatically releasing it as open data. Although the releases have been slightly reformatted to fit data protection regulations by censoring personally identifiable information and delaying data flow, this data still enables a whole country to be modelled, to some extent, as a big network of different organisations interacting with each other.

Understanding the patterns of these interactions and how these organisation are split into industries or sectors in this so-called data economy can help us identify the relative importance and relevance of these organisations and sectors. This in turn could eventually be applied to a real-time system to spot anomalies as quickly as they arise, either in a cyber threat perspective or economics perspective.

1.3 Research objective

The Estonian X-tee data offers a great starting dataset to build analysis methodology and metrics that could later be applied to other similar infrastructures (e.g. the Finnish instance of X-Road). Therefore, this project aims to first describe the graph structure with some key metrics, and detect possible communities and their characteristics.

The objective is to represent the plain transaction data as a network, analyse it's distribution (hence deduce applicable models), connectedness, most active members, implied communities, and how some of these measures change in time. The aim for the results is to add value to the already existing infrastructure insights [6]. The hypothesis is that network science and graph analysis methods could be an insightful perspective to help improve the decision making and monitoring of the Estonian X-Road network but could also be applied to other data exchange networks already serving millions of people.

In addition to the other promising benefits mentioned previously, these analysis metrics could increase the confidence of new administrations of countries or municipalities to adopt such secure data exchange layers. This decreases bureaucracy, paperwork, and increases the overall happiness of citizens which could soon add up to tens of millions more people on similar data interoperability infrastructures [17].

2 Subject-specific Background

2.1 General background

When dealing with networks, we often come across two sets of terminology. In Network Science, we have networks that are comprised of nodes and related pairs of nodes, called links. And in Graph Theory, we have graphs that are comprised of vertices and related pairs of vertices, called edges. The subtle difference in those two terminologies is the field of application. When talking about real world systems the {*network, node, link*} terminology is most appropriate and the {*graph, vertex, edge*} terminology is used when discussing the mathematical representation of these networks [2]. However, as in many scientific papers, I will use each equivalent term interchangeably.

Most of the analysis is done using R programming language. The underlying X-Road data is gathered programmatically through an *application programming interface* (*API* here-on out) hosted by the Information System Authority of Republic of Estonia. In this case, the API is a public URL where parameters like date, time, desired columns of the data, and any constraints on values can be defined [5]. Once parameters are defined, the URL is executed and the response is a human- and machine-readable JSON (*JavaScript Object Notation*) file.

R enables installing custom packages to provide niche functionality, like API requests, and I'm using the **igraph** package to create network objects. *Network objects* are data types that can be manipulated, analysed, and plotted using Graph Theory and Network Science practices.

2.2 Graph Theory Background

A *graph* (G) is an object consisting of two sets called its vertex set (V) and its edge set (E), thus a graph is notated $G(V, E)$ [10].

Graphs can have multiple edges and loops. A *loop* is an edge that joins a vertex to itself. A *multiple edge* means there's two or more edges that connect the same nodes (and in the same direction, for directed graphs). If a graph doesn't have loops or multiple edges then it's called a *simple* graph, otherwise it's considered a *multigraph*. We will aggregate our data such that there are no multiple edges, but there are loops as the same X-Road member can have multiple sub-services that request data from each-other. So we still consider our graph technically a multigraph.

A *directed* graph is a graph which edges have orientations. The rigorous definition of a graph in the case of a directed multigraph (also called a *quiver*) must be expanded. A quiver is an ordered triple $G(V, E, \phi)$ where:

- V is the set of vertices/nodes.
- E is the set of edges/links.
- $\phi : E \longrightarrow \{(x, y) \,|\, (x, y) \in V^2\}$ is an *incidence function* mapping every edge to an ordered pair of vertices.

In practice, the incidence function is provided inside the R function that creates the network object and we just need to provide the set of nodes, and set of links. Since *set* is a specific mathematical term, I will be using alternatives such as *list*, *array*, *table*, or similar when speaking about the more practical applications of the node and link sets.

2.3 Network Science Background

2.3.1 Total number of nodes and links

A basic graph metric we cover is the number of nodes (N) of a graph. It represents the total number of unique members, and also called the size of the network. To distinguish the nodes, they've been labelled by their official unique member codes (each of which can be mapped to a human-readable member name).

Secondly, number of links (L) represents the total number of distinct interactions between the nodes. Since there's no multiple links (connecting same nodes in the same direction) as mentioned in Section 2.2, then links don't have to have unique names as they can be identified through the nodes they connect [2]. For example, the (70009770, 74000091) link connects from node 70009770 to node 74000091 (8-digit number is the standard form of ID for X-tee Member IDs/codes).

The *maximum number of links* is denoted L_{max}. For our type of network (directed with loops, but no multiple links in the same direction) the maximum number of links occurs when each node (N total) is directly connected to every other node ($N - 1$) and itself (loop, counts as 1 link), thus $L_{max} = N \times (N - 1 + 1) = N^2$. A network where $L = L_{max}$ is called a *complete* network (a *clique*).

2.3.2 Adjacency matrix of weighted networks

As mentioned in Section 2.2, we will aggregate our data such that there are no multiple edges in the same direction. This means we will sum all queries made between the same node pair and turn this sum into the weight

attribute of a single link in the given direction. Therefore each link has a *weight* that could be notated as w_{ij} for a link that connects from i to j.

Adjacency matrix is a way to represent the complete list of the links in a matrix form. The adjacency matrix of a network of N nodes has N rows and N columns. Every node connection can be represented by the weight between the nodes. If a link does not exist, the weight is simply 0. Thus the matrix can be defined by it's elements:

$$A_{ij} = w_{ij} \text{ for } i,\, j \in V$$

Since we have a directed network, w_{ij} could be different from w_{ji} and thus the adjacency matrix A_{ij} is likely to be asymmetric (i.e. $\exists\, i,\, j$ s.t. $A_{ij} \neq A_{ji}$).

2.3.3 Degree, Distributions, Strength

The *degree* of the i^{th} node in the network is denoted k_i and it describes how many link are connected to this specific node. In directed networks, k_i more precisely describes the node's *total* degree, which is actually comprised of its *incoming* degree k_i^{in} and its *outgoing* degree k_i^{out}:

$$k_i = k_i^{in} + k_i^{out}$$

The node with the highest degree, notated $k_{\max}$, is the most connected node, also called a *hub*.

In a directed network the total number of links, L, can be expressed as the sum of the node degrees:

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$

Thus we can find the third metric that describes the structure of the network - the *average degree*:

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

The average degree can be considered the expected value of the *degree distribution*, denoted p_k, of the graph:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

For a network of N nodes, the degree distribution is the normalized histogram given by $p_k = \frac{N_k}{N}$ where N_k is the number of nodes with degree k.

For weighted graphs, instead of summing number of incoming and outgoing nodes, it's in some cases more informative to sum incoming and outgoing link weights for each node, which is called the node *strength*. Similarly as above, maximum strength, average node strength and strength distribution are the weighted equivalent of the "degree" counterparts.

2.3.4 Random Networks, Critical point

Random Network Model is a model where links are placed randomly between nodes. This aims to reproduce the complexity and messiness of real networks. In this model, the existence of a link is dependent on a chosen probability (chance of two nodes being connected). Higher probability results in more links on the network resulting in higher average node degree $\langle k \rangle$ and *vice versa* [7].

Under this model, it's been observed that there exists a *critical point* where one *giant component* (a large cluster) emerges. This critical point happens to be $\langle k \rangle = 1$. Every network with $0 < \langle k \rangle < 1$ is called *subcritical*, $1 < \langle k \rangle < \ln N$ *supercritical*, and $\langle k \rangle > \ln N$ *connected* network (every node is absorbed into the cluster) [15]. Real networks are often supercritical as they are not always fully connected but tend to have one large cluster, so this is one condition we will be checking on our data.

2.3.5 Sparseness

Real world networks, which we might expect to have in the X-Road monitoring data, are sparse, meaning there's a big variation in node degrees. *Sparse* graph is broadly defined as a situation where the number of links in a graph is considerably smaller than the maximum possible number of links:

$$L \ll L_{max}$$

Equivalently this means the average in- and out-degrees are considerably smaller than the total number of nodes:

From Section 2.3.3

$$\langle k^{in} \rangle = \langle k^{out} \rangle = \frac{L}{N} \implies L = N \times \langle k^{in} \rangle = N \times \langle k^{out} \rangle$$

and from Section 2.3.1

$$L_{max} = N^2$$

Thus inequality becomes

$$N \times \langle k^{in} \rangle = N \times \langle k^{out} \rangle \ll N^2$$

$$\langle k^{in} \rangle = \langle k^{out} \rangle \ll N$$

2.3.6 Power Law Degree Distribution

It's pretty common for the degree distribution of real world networks to follow the *power law distribution*. It's a rich-get-richer type of distribution that was also the basis of Pareto Principle, that describes a pattern of 20% of causes resulting in 80% of the effect [11]. The distribution is formulated by:

$$p_k \sim k^{-\gamma} \text{ where } \gamma > 0 \text{ is a constant}$$

The challenge is to figure out the best way to detect a power law distribution. For this, we'll actually plot the logarithm of degree distribution ($\ln p_k$) against the logarithm of network's degrees ($\ln k$). We should expect to fit a straight line as seen by formula:

$$p_k \sim k^{-\gamma} \implies \ln p_k \sim -\gamma \ln k$$

In case of a power law, the observations will result in a downward trending straight line, with slope $-\gamma$. We will see if our data holds up to this common characteristic of real networks.

2.3.7 Clustering Coefficient

The *clustering coefficient* captures the degree to which the neighbors of a given node link to each other, and hence are a common metric for measuring network's connectedness.

For a node i with degree k_i the *local* clustering coefficient is defined as

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

where L_i represents the number of links between the k_i neighbors of node i. Note that C_i is between 0 and 1

- $C_i = 0$ if none of the neighbors of node i link to each other
- $C_i = 1$ if the neighbors of node i form a complete graph, i.e. they all link to each other.

- C_i is the probability that two neighbors of a node link to each other

The *average* clustering coefficient $\langle C \rangle$ expands this idea to the whole graph:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

Similarly to every nodes clustering coefficient, the average clustering coefficient will also lie between $0 \leq \langle C \rangle \leq 1$. The closer the coefficient is to 1 the more connected the links are, on average.

Lastly, the *global clustering coefficient* (C_G), equivalently called *transitivity*, is an alternative metric to the average clustering coefficient which prioritises nodes with smaller degrees by observing triplets of nodes. Node triplets can be with either two links (considered *open*) or three links (considered *closed*). The global clustering coefficient is simply the ratio of closed triplets over the total number of triplets, both closed and open. Transitivity can be applied to both undirected and directed networks [9].

2.3.8 Clustering and Communities

Often it's insightful to also find out whether there are any similar groups apparent in the network. *Clustering* means automatically splitting the network into modular structures, also called *communities* or compartments, of related members (conventionally understood to be large sub-graphs with high internal densities). Clustering algorithms can provide a scalable way to identify functionally important or closely related classes of nodes from interaction data alone [8].

There are different programmed algorithms to reach an arbitrary number of clusters. The function I'm using calculates the *optimal* community structure for a graph, in terms of maximal modularity score. The general gist of the modularity score approach is to split the network into a number of partitions and calculate a score that reflects every partition's internal connectivity. The partitions with highest scores will be chosen as designated communities [3].

There exists also *greedy* algorithms that have reduced complexity and hence fit better for use in big networks ($N > 100$). However as the name might suggest, greedy algorithms optimise for short-term gain and hence the result will often not be optimal.

3 Results

3.1 Understanding the data

After the X-tee data is downloaded from the open data API provided by the Information System Authority of Estonia, it is turned into a network object where nodes are the members of X-tee and links are the directed and weighted data queries between the members.

Both nodes and links have extra attributes as well. Nodes are identified by their official member codes (for Estonia, it's usually an 8-digit number). Besides a human-readable name in Estonian, every member is given a class and an instance attribute as well, e.g. an Estonian government institution will be labelled "GOV" and "EE" respectively, a Finnish commercial organisations will be labelled "COM" and "FI" respectively *etc.*

Edges have two attributes - weight and ln(weight), where "weight" really means the total query volume between two nodes. The logarithmic edge weight attribute is used mostly for visual purposes as the difference between the smallest and the largest edge weights is multiple orders of magnitude. As mentioned in Section 2.2, loops are possible as one member could have multiple subsystems that query data from each other.

Since the querying activity on X-tee is high volume (millions of transactions per day), for computing purposes, we'll be observing 1-hour chunks worth of activity data at a time.

It's best to start with a visualisation of the network. Since we're dealing with around 600 daily active members on the data exchange infrastructure, it makes sense to visually show only top n most active members. Figure 1 shows 1-hour worth of X-Road activity made by the top 20 most active members on a sample working day.

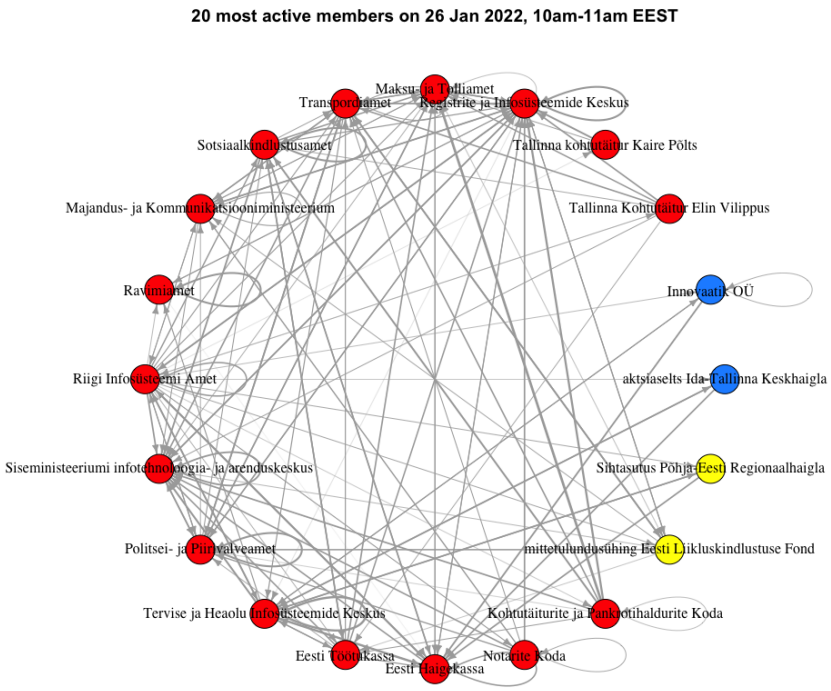

Figure 1: Visual representation of 20 most active members on Jan 26, 2022 from 10am-11am EEST displayed with their member names (in Estonian) and colored by their member class. Red - Government; Blue - Commercial; Yellow - Non-Government Organisation

We can immediately see how much the data infrastructure is dominated by the public sector. Also noteworthy is that from the commercial and NGO members (blue and yellow), 3 out of 4 operate in the healthcare sector. Hence we can expect the healthcare sector to be a very significant community on the network.

In Figure 1, each link's thickness is in proportion to the logarithm of the number of queries performed. This means that node pairs with thicker links are more actively querying information between each other, in link's direction. However, to better understand the difference in magnitude of query volume we ought to plot a heat-map:

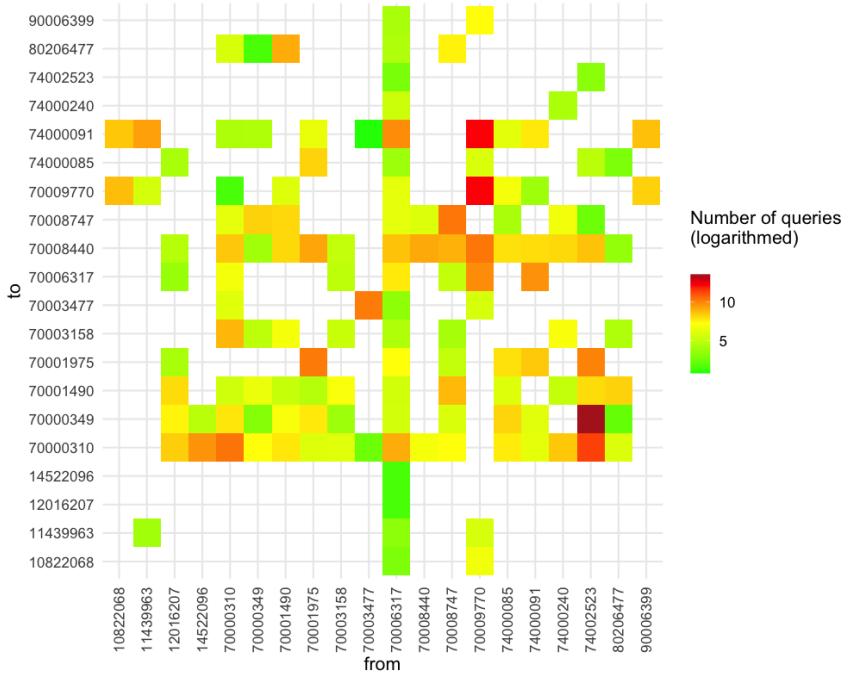

Figure 2: A heat-map for the same 20 most active members on Jan 26, 2022 from 10am-11am EEST displayed with their member codes. Horizontal axis displays outward direction and the vertical axis displays inward direction activity. Scale: $0.2 < \ln w_{ij} < 13.7$, absence of color tile indicates no link ($w_{ij} = 0$).

The heat-map in Figure 2 is essentially a visualisation of the adjacency matrix as defined in Section 2.3.2. Color shows logarithmic query volumes (logarithmic weight, $\ln w_{ij}$). As we can see, it is indeed asymmetric. There's only a very few nodes with dark-red tiles meaning a small number of members (even among the 20 most active) perform dominant proportion of total queries which might hint at a power law distribution described in Section 2.3.6. So next we're going to more rigorously test if the X-tee activity resembles a real world network.

3.2 Real Network Characteristics

As mentioned in Section 2.3.4, it's common for real world networks to be supercritical i.e. $1 < \langle k \rangle < \ln N$. To find out whether this applies for X-tee data, we have tabled the changes in the average node degree $\langle k \rangle$ and the

natural logarithm of total number of active nodes (members) $\ln N$ in Table 1.

Network	Date	Time	N	L	$\langle k \rangle$	$\ln N$
X-tee	26 Jan 2022	8am-9am	658	1603	2.44	6.49
X-tee	26 Jan 2022	9am-10am	658	1829	2.78	6.49
X-tee	26 Jan 2022	10am-11am	659	1672	2.54	6.49
X-tee	26 Jan 2022	11am-12pm	656	1654	2.52	6.49
X-tee	26 Jan 2022	12pm-1pm	657	1812	2.76	6.49
X-tee	26 Jan 2022	1pm-2pm	658	1665	2.53	6.49
X-tee	26 Jan 2022	2pm-3pm	659	1694	2.57	6.49
X-tee	26 Jan 2022	3pm-4pm	658	1804	2.74	6.49
X-tee	26 Jan 2022	4pm-5pm	659	1628	2.47	6.49
X-tee	26 Jan 2022	5pm-6pm	658	1512	2.31	6.49
X-tee	26 Jan 2022	6pm-7pm	659	1635	2.48	6.49
X-tee	26 Jan 2022	7pm-8pm	658	1417	2.15	6.49

Table 1: Overview of X-tee with key structure metrics on Jan 26, 2022 from 8AM-8PM EEST

In our case $1 < \langle k \rangle < \ln N$ throughout the day meaning our graph is in supercritical, meaning the network shows one large dominating cluster (giant component). This gives us confidence that our network has the characteristics common to real world networks.

The network is also sparse throughout the day because $L_{max} = N^2 \approx 430\,000$ so $L \ll L_{max}$, as defined in Section 2.3.5.

Having currently passed all checks for common characteristic of real networks, we can next take a look at the degree distribution and see whether it follows a power law $p_k \sim k^{-\gamma}$, as proposed in Section 2.3.6. If indeed we have a real world network then we should be able to fit a straight line onto a plot of $\ln p_k$ against $\ln k$. The fitted line should have a slope of $-\gamma$. Since we're dealing with a directed graph, we have to observe the in-degree distribution separate to the out-degree distribution.

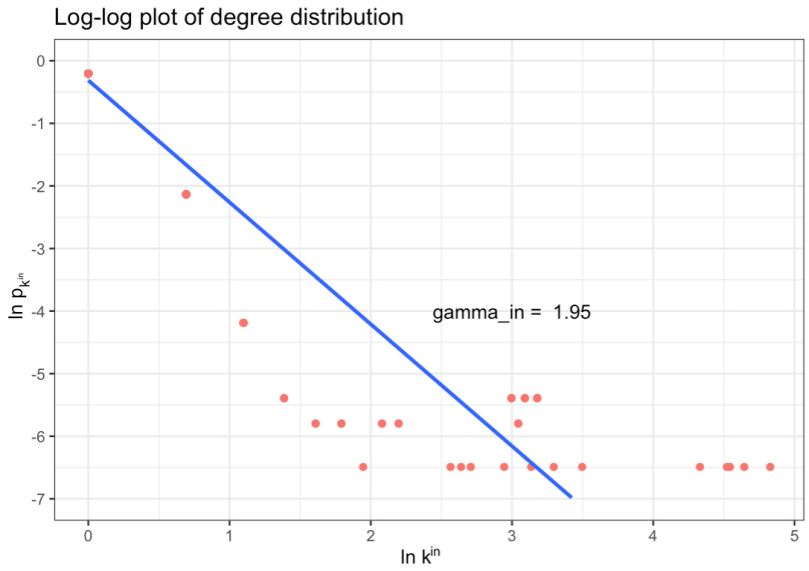

Figure 3: Log Distribution against Log of Incoming Degrees fitted with a straight line of slope $-\gamma_{\text{in}} = -1.95$

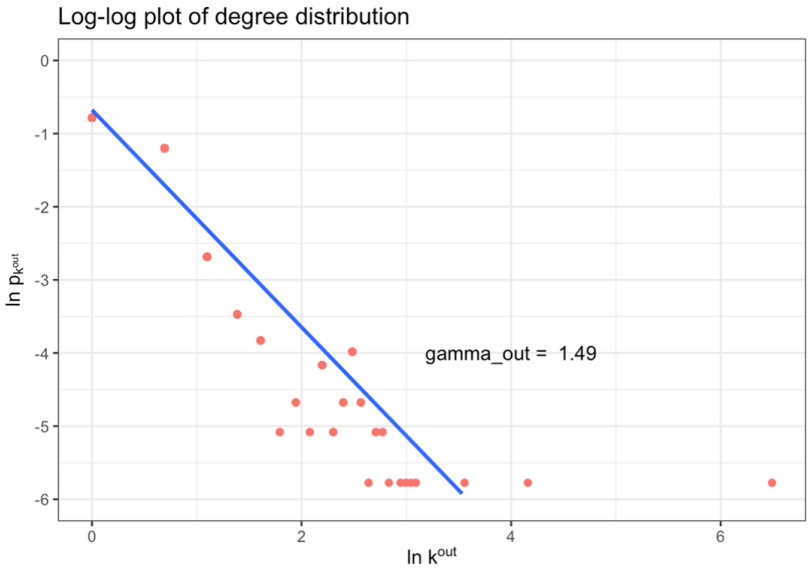

Figure 4: Log Distribution against Log of Outgoing Degrees fitted with a straight line of slope $-\gamma_{\text{out}} = -1.49$

Both plots are based on the 10am-11am data shown in Table 1. For the incoming degrees, the distribution doesn't seem to fit a line. It could potentially be considered a transition between two lines, a steep one for small $\langle k \rangle$ ($\ln\langle k \rangle < 2$) and a shallow one for larger $\langle k \rangle$ ($\ln\langle k \rangle > 2$). For out-degrees in Figure 4, the data reasonably fits a straight line, meaning that at least the out-degrees do seem to follow a power law with constant $\gamma = 1.49$.

Overall, similar patterns of degrees, sparseness, and distributions hold for other analysed days as well. Hence these metrics can give reasonable insights to the X-Road activity, which seem to share many of the commonly occurring patterns of real world networks.

For X-tee specifically, we found out that typically there's just under 660 members active on the network at each hour throughout the day. Each member, on average, is receiving data queries from 2-3 members and sending data queries to 2-3 members. And the receiving and requesting members are likely not the same as hinted with the asymmetric adjacency matrix in Section 3.1.

3.3 Hubs

As briefly introduced in Section 2.3.3, hubs are nodes with the highest degree in a network. Since the X-Road network is directed we can observe the hubs (in other words the most connected members) by finding the node with largest in- and out-degrees. As an example in Table 2, we can see that the in-degree hub changes throughout the day.

Hub status in Table 2 mostly fluctuates between Ministry of Internal Affairs IT- and Development Center, which provides IT solutions for a lot of different internal activities from Police e-services to passport- and ID verifications [12], the Tax and Customs Board, and during the daytime Estonia's national Health Insurance Fund.

Time	k^{in}_{max}	Hub Member Name (EN)
12am-1am	82	Ministry of Internal Affairs IT- and Development Center
1am-2am	71	Ministry of Internal Affairs IT- and Development Center
3am-4am	64	Estonian Tax and Customs Board
4am-5am	78	Ministry of Internal Affairs IT- and Development Center
5am-6am	58	Ministry of Internal Affairs IT- and Development Center
6am-7am	60	Estonian Tax and Customs Board
7am-8am	72	Ministry of Internal Affairs IT- and Development Center
8am-9am	75	Ministry of Internal Affairs IT- and Development Center
8am-9am	123	Estonian Health Insurance Fund
9am-10am	127	Estonian Health Insurance Fund
10am-11am	125	Estonian Health Insurance Fund
11am-12pm	121	Estonian Health Insurance Fund
12pm-1pm	117	Estonian Health Insurance Fund
1pm-2pm	119	Estonian Health Insurance Fund
2pm-3pm	125	Estonian Health Insurance Fund
3pm-4pm	115	Estonian Health Insurance Fund
4pm-5pm	110	Estonian Health Insurance Fund
5pm-6pm	87	Ministry of Internal Affairs IT- and Development Center
6pm-7pm	94	Ministry of Internal Affairs IT- and Development Center
7pm-8pm	80	Ministry of Internal Affairs IT- and Development Center
8pm-9pm	81	Ministry of Internal Affairs IT- and Development Center
9pm-10pm	83	Ministry of Internal Affairs IT- and Development Center
10pm-11pm	71	Estonian Tax and Customs Board
11pm-12am	73	Estonian Tax and Customs Board

Table 2: X-tee hubs throughout Jan 26, 2022 by largest incoming-degree

When observing the out-degree hubs then throughout the day the Estonian Information Systems Authority (ISA) is the biggest hub. It's likely that the ISA is launching automated queries to test other member's IT systems as the ISA is responsible for the up-keeping of X-tee [4].

Similarly, we can find nodes with the highest node strength. This will show nodes with the highest in- and outward query volume, the weighted equivalent of hubs. We find that volume-wise, and throughout the day, the Tax and Customs Board (TCB) is the member with highest volume of incoming queries with a total amount of around 10 million per day. Biggest out-going node strength, in other words the node who does the most data querying is the Estonian Chamber of Bailiffs and Trustees in Bankruptcy (ECBTB) with also just below 10 million data queries sent per day. And the data activity between the TCB and ECBTB seem to overlap to some extent from looking at the thick link from "Kohtutäiturite ja Pankrotihaldurite Koda" (ECTBTB) to "Maksu- ja Tolliamet" (TCB) on Figure 1.

3.4 Average and Global Clustering coefficients

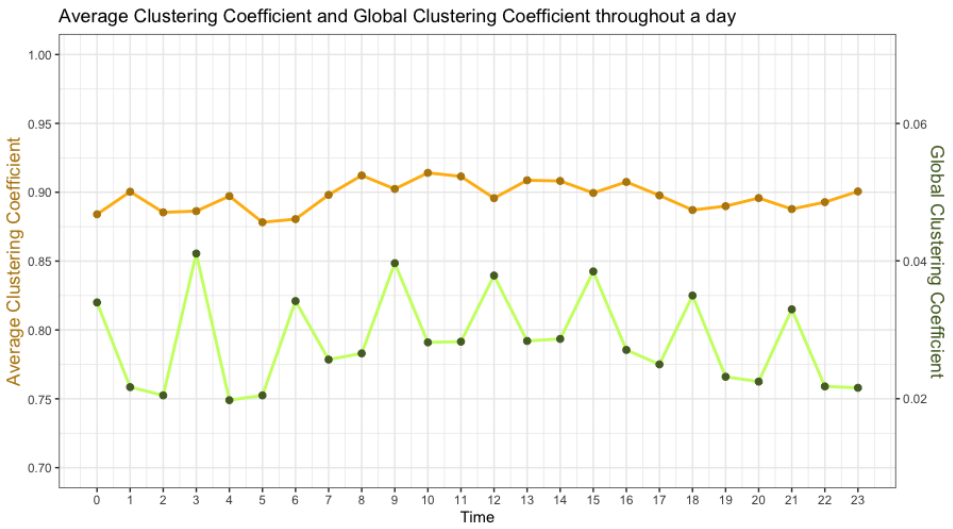

Figure 5: Average and Global Clustering coefficients on a dual y-axis depicted throughout Jan 26, 2022, x-axis shows hours since midnight

The clustering coefficient introduced in Section 2.3.7 shows how diversely the nodes in the network are connected. The average clustering coefficient

stays stable around 0.9, meaning that on average every member's neighbor is locally very well connected. Global clustering coefficient is a connectedness metric where more emphasis is put on so-called closed connections hence the small coefficient hints at a lot of one-way querying and centralised data sources.

These results show that smaller members query data from well connected, higher degree, members. This makes sense, to the extent that a lot of relatively small private commercial organisations depend on well connected government institutions for data flow. But many government organisations, with the exception of ISA, don't have a lot of reason to queries the smaller degree members.

3.5 Communities

Lastly, I'm going to observe communities. As introduced in Section 2.3.8, community clustering algorithms use elaborate methods to decide on optimal partitioning of the network in order to outline important and closely related groups of members (nodes).

Since clustering algorithms are computationally heavy, I ran clustering only for the top 50 most active members of the infrastructure for each 1-hour slot. An example of which can be see in Figure 6.

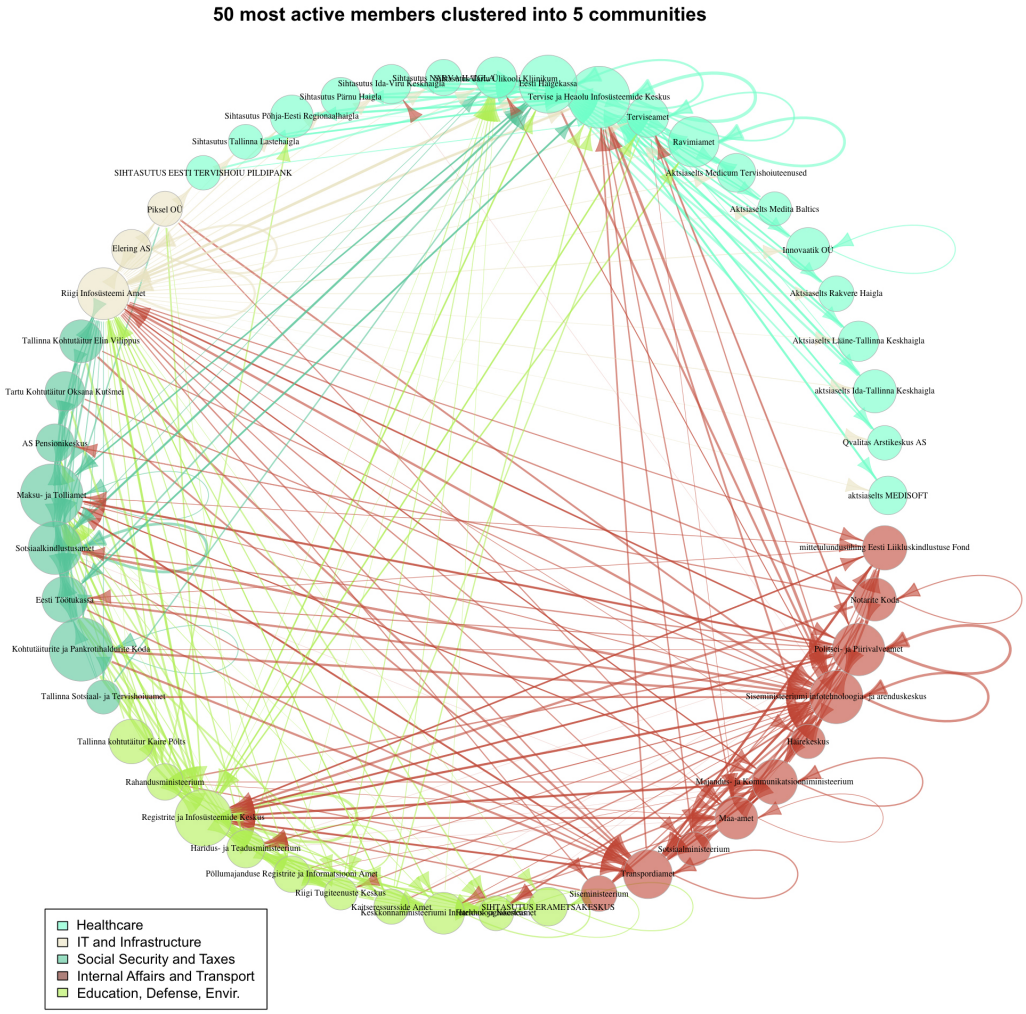

Figure 6: 50 most active members grouped into 5 optimal communities, Jan 26, 2022, 10am-11am data, each node is sized in proportion to it's strength, legend identified by hand

As predicted in Section 3.1, healthcare and medical sector do play a significant role next to many government-led communities. In healthcare, the communication is more concentrated internally, followed by communications with the Social Security and Taxes community. This connection shows the function of free/insured public healthcare for citizens as medical institutions

are in a constant need to query social information such as employment and tax data.

Throughout the day, the Internal Affairs and Transport sector seem to be the community that communicates with the outside the most. Under this community we have organisations who provide digital services of national importance, such as live data for emergency services, passport and ID-card verifications, and radio and communications. In fact, the Estonian Police is the second strongest node in this community, second to the Ministry of Internal Affairs IT- and Development Center.

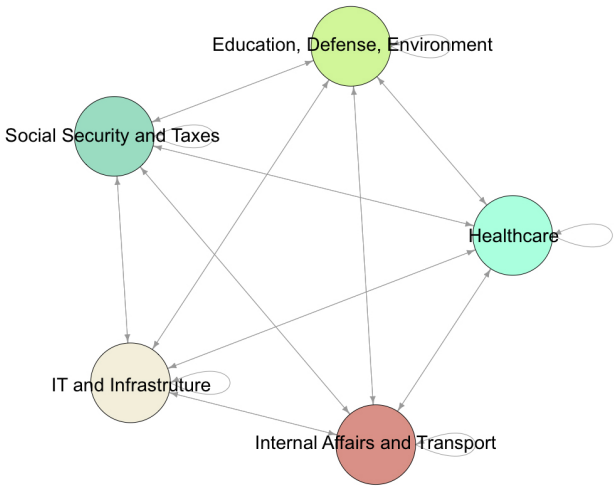

Figure 7: Interactions between the 5 optimal communities leads to a complete graph. Further analysis could make link thickness proportional to query volume.

Overall, it's a very promising result that clustering algorithms grouped similar organisations into the same communities impressively accurately. This shows that there is indeed enough information for network analysis methods to imply the sector of an organisation based on simply their transactional activity data without the core data itself. So this is a topic that should definitely be expanded further.

4 Conclusions and Outlook

4.1 Conclusions

Over the course of the project, I analysed over 30 million data queries made on the Estonian instance of X-Road, the national data exchange infrastructure in Estonia. We reached many interesting conclusions thanks to Network Science analysis methods.

When approaching the data from a network perspective we found that X-tee network shares many attributes with other real world networks. It's a rather sparse network, with a giant component, and parts of the network follow a power law distribution. It's overall not very connected ($L \approx 1670 \ll L_{max} \approx 430\,000$), a lot of nodes do one-way queries (global cluster coefficient $C_G \approx 0.03$), but it's nodes in general are well connected (average cluster coefficient $\langle C \rangle \approx 0.89$). Assuming other X-Road networks follow similar patterns, this would lay a good foundation for further modelling of such networks. For example, some prediction models assume the underlying data follows a power law distribution [14], and we've shown that the outgoing degree distribution of the X-tee network follows indeed that.

We deduced that the public sector is the backbone of our data exchange infrastructure because the network's biggest members, both in connectedness and volume, are governmental organisations. Most diverse querying at night was made by the Tax and Customs authorities and IT services, and during daytime it was the healthcare-related services that had the biggest appetite for data. In terms of query volume, the Tax and Customs authority's services are queried the most. The biggest volume of queries sent/requested is from the Estonian Chamber of Bailiffs and Trustees in Bankruptcy, also contributing for a big portion of the Tax and Customs Board's incoming queries. It's seems evening and nighttime is a prime time for mass data queries from government organisations on people and companies given the activity of Tax Authority and Bankruptcy Bailiffs mentioned previously. Similarly during daytime, it's the service sector that flourished, most notably healthcare, from the Health Insurance Fund to actual hospitals showing up in the top 20 most active X-tee members.

Remarkably, we found that modelling X-tee transaction data as a network enabled us to group members into communities fairly accurately. Most clustering attempts resulted in 5 communities: Healthcare, IT and Infrastructure, Social Security and Taxes, Internal Affairs and Transport, and lastly Education, Defense, Environment. This shows a potential of further insight on the data transaction flows between communities.

With this, I've achieved my research objective to gather more insight on the dynamic X-tee network from Network Science perspective. I represented the activity data as a network, fitted potential models for its distribution, gave metrics on connectedness, outlined most active members, and clustered 5 distinct (and realistic) communities. The aim for the results was to add value to the already existing infrastructure insights [6] which we did by new visualisations and metrics that could be applied to real-time data.

4.2 Outlook

This project has lots of opportunity for expansion. I've just scratched the tip and there's more to do from all perspectives – in connectedness metrics, distribution modelling, and analysis of the data flows.

Connectedness could be explored with concepts like reciprocity, betweenness and other similar metrics. Distribution models could be expanded as brought up in Section 3.2, for example fitting the in-degree distribution with two linear lines instead of one. And analysis of the data flow could be expanded in multiple ways. Firstly, the flow changes between the defined communities over time could quantify a standard data transaction procedures we should expect on such data infrastructures. And secondly, the original data has query packet sizes that could be taken account into query volume to clear the model. E.g. in this project I assumed 2 queries to be higher volume than 1 query, but what if the 1 query is the size 100KB and the 2 queries in total make 50KB – then the volume of 2 queries is actually smaller than the volume of 1 query and the inference on network structure and dynamics could differ.

Another interesting direction to take is to consider the fact that only 3% of requests on X-tee are submitted by citizens [6], everything else is automated. When queries could be distinguished into types (single vs. automated) and analysed the behavior of these types, then new insights mechanisms on actual citizen impact on the network could be built.

Lastly, the ultimate goal for this project would be to have a live overview of X-tee based on (near) real-time data to draw conclusions on behaviour- and follow metrics live. This could have the potential to spot anomalies independent from the infrastructure members and overall help improve the decision making and monitoring of any X-Road (or similar) system.

References

[1] Nadeem Akhtar, Hira Javed, and Geetanjali Sengar. "Analysis of Facebook Social Network". In: Sept. 2013, pp. 451–454. DOI: 10.1109/CICN.2013.99.

[2] Albert-László Barabási. *Network Science*. URL: http://networksciencebook.com/. (accessed: 26.04.2022).

[3] Ulrik Brandes et al. "On Modularity Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 20 (2008), pp. 172–188.

[4] Republic of Estonia Information System Authority. *Data Exchange Layer X-tee*. URL: https://www.ria.ee/en/state-information-system/x-tee.html. (accessed: 14.05.2022).

[5] Republic of Estonia Information System Authority. *X-Road v6 monitor project - Open Data Module, API documentation*. URL: https://github.com/ria-ee/X-Road-opmonitor/blob/master/docs/opendata/user_guide/ug_opendata_api.md. (accessed: 02.05.2022).

[6] Republic of Estonia Information System Authority. *X-Tee Factsheet EE*. URL: https://www.x-tee.ee/factsheets/EE/#eng. (accessed: 14.05.2022).

[7] E. N. Gilbert. "Random graphs". In: *The Annals of Mathematical Statistics* 30 (1959), pp. 1141–1144.

[8] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset. "Performance of modularity maximization in practical contexts". In: *PHYSICAL REVIEW* 81 (2010).

[9] Introduction to Graph Theory. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[10] Introduction to Graph Theory. *Trudeau, Richard J.* New York: Dover Pub., 1993. ISBN: 9780486678702.

[11] John Hagel. *The Power of Power Laws*. URL: https://edgeperspectives.typepad.com/edge_perspectives/2007/05/the_power_of_po.html. (accessed: 29.04.2022).

[12] Ministry of Internal Affairs IT- and Development Center. *About us (in Estonian language)*. URL: https://www.smit.ee/et/siseministeeriumi-infotehnoloogia-ja-arenduskeskus. (accessed: 08.05.2022).

[13] Petteri Kivimäki. Private Communication in Slack. 2022.

[14] Eric D. Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data with R*. Use R! Springer New York Heidelberg Dordrecht London, 2014. ISBN: 9781493909827.

[15] R. Solomonoff and A. Rapoport. "Connectivity of random nets". In: *Bulletin of Mathematical Biology* 13 (1951), pp. 107–117.

[16] Nordic Institute For Interoperability Solutions. *Case Study: Iceland joins the Nordic interoperability league with Straumurinn*. URL: `https://x-road.global/iceland-joins-the-nordic-interoperability-league-with-straumurinn`. (accessed: 14.05.2022).

[17] Nordic Institute For Interoperability Solutions. *Integrity and interoperability – the perfect match for Argentina's public service*. URL: `https://x-road.global/integrity-and-interoperability-the-perfect-match-for-argentinas-public-service`. (accessed: 14.05.2022).

[18] Nordic Institute For Interoperability Solutions. *X-Road® Data Exchange Layer*. URL: `https://x-road.global/`. (accessed: 14.05.2022).