Universiteit Leiden

ICT in Business and the Public Sector

Critical node analysis on supply chains: a network approach

Name: Corjan Meijerink
Student-no: 1520024

Date: 04/07/2019

1st supervisor: Prof. Dr. Stefan Pickl
2nd supervisor: Prof. Dr. Wolfgang Bein

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

In this thesis three theoretical chapters are provided. In these chapters, the fundamental concepts of supply chains and network analytics are extensively discussed. Another chapter is focussed on critical node analysis which links the two concepts together to show how a social network approach can be used to analyse supply chains. The goal of this method is to find the critical nodes in a real supply chain. Finding these nodes allows an organisation to optimise their supply chain, leading to a competitive advantage.

Given real data from a large car manufacturing company, an analysis is performed. Throughout this thesis, a theoretical framework is provided to show how social network analysis can be used as a method of analysis. With the use of several examples and sample calculations it is shown how critical nodes can be identified. This framework mainly relies on centrality measures and clustering. Complementary to the framework, basic statistics can be used to analyse the received data. By calculating median values and the median absolute deviation, actionable results are found. This is achieved by comparing idle and transport times with each other. By filtering on specific transportation methods / agents / paths, the critical nodes are found. Location 6 has a very high idle time but a very short transportation time, therefore this location is identified as the most critical. Furthermore, the shipping method is used for small distances which was not expected and trucks have a long transportation time which is not constant at all.

The main conclusions of this thesis are that a social network approach facilitates a great framework for supply chain analysis. When the data is structured as a real world network, great results can be found with the use of simple methods. The complementary analysis that relies on median values and the median absolute deviation have been proven very useful in this thesis. A solid foundation has been laid and based examples it has been shown how a social network analysis can be conducted on supply chains. Actionable results were found in the real data, mainly due to the complementary analysis. By improving the processing of the goods between the transportation methods at the critical nodes, greater supply chain efficiency can be achieved.

Acknowledgements

Thanks to Stefan Pickl for the intensive guidance. It was a wonderful experience to work together and learn more about supply chains, networks analysis and related fields. I felt very welcoming when invited to Munich where I was able to meet all the scientists working at the COMTESSA institute. A special thanks to Sorin Nistor for the regular skype sessions and aiding in the collection of data.

I am grateful to Wolfgang Bein for being the second supervisor.

To Arno Knobbe and other employees of the LIACS department. This thesis is the final assessment of my two year masters degree ICT in Business & Public Sector. I have however completed this degree in one year which meant that there were some logistical issues. He, amongst others, helped me with all the organisational tasks required to complete all courses within a year.

Last but by no means least, thanks to my loving girlfriend and my family for supporting me in every way possible.

I thank you all!

Contents

Chapter 1

Introduction

One of the major buzzwords in the 21^{st} century is Big Data: a concept describing how different data is nowadays than a decade earlier. Developed societies nowadays are capable of analysing every single aspect of an organisation, device or even a human. With the introduction of smart watches a person's health is constantly monitored, tracking cookies exploit knowledge about your activity on the web for a better advertising experience and organisations can monitor all their internal processes to control if all goes as planned. One of the processes that is often monitored is the supply chain.

With all the available data about supply chains it is crucial to find the most important, critical information. To find what this is, a method to analyse the data has to be introduced. By approaching the supply chain data as a network structure, methodologies based on network topology can be used to find the critical point in a supply chain. Once a supply chain is fully analysed and the critical nodes are found, management can use the obtained information and act accordingly. Some critical nodes might be intended while others are an indication about suboptimal processes. The interpretation of the critical nodes is therefore entirely up to the management itself.

1.1 The project

This thesis will lay the foundation for quick and easy methods that can be used to analyse supply chain data. The goal of this is to enable management to identify critical processes in the supply chain in order to take action accordingly on time. The main vision is to use methodologies provided by network science. With this, a general approach to the identification of critical nodes can be given. Based on these methodologies, a theoretical framework will be introduced throughout the thesis. With the help of multiple examples and sample calculations it will be shown how such a social network approach will work in the framework. Besides the network methodologies, basic statistics are used as a complementary analysis. By calculating the median and median absolute deviation values, another method is provided of finding the critical nodes opposed to the introduced framework. This thesis aims at using the social network framework to provide a comprehensible way of finding critical nodes in a supply chain. Using such a network approach, easy to understand visual reporting can be used to summarise the findings on management level. With these reports, management can take the required actions to leverage the competitive advantage created by their supply chain.

"How to find critical nodes in supply chains from a network perspective"

1.2 Thesis overview

In this chapter, the context of this thesis has been briefly introduced. Chapter 2 will elaborate on the necessary theories to understand supply chains in depth, along with the challenges supply chains face in the 21^{st} century. Chapter 3 introduces basic concepts about networks and it's topology. Together with these theories, methodologies are elaborated upon to comprehend how networks can be analysed. The final fundamental chapter is Chapter 4 where the concept of critical nodes is discussed. After these three chapters providing insight in the theories, Chapter 5 shows what practical steps are taken to analyse the used data. In Chapter 6, the results of the data are shown and interpreted. Chapter 7 summarizes the results into actionable steps and mentions aspects for further research. The final Chapter 8 concludes.

Chapter 2

Supply Chains

2.1 General introduction

An economy is identified by two factors: supply and demand. Customers have a certain need for products and the suppliers meet this desire by producing the goods. By doing this, a market is born and economic value is created by this interaction between suppliers and buyers. There are all sort of markets that facilitate trade. Ranging from, but not limited to, physical, non-physical and financial products. These previous examples are very broad and it is therefore not hard to think of a more specific market that lies within for example the physical market. The computer, food and car market are sample markets that lie within the physical markets. A market is merely a concept describing a 'place' of exchange that enables the allocation of resources between people.

Within each market there is a supply side identified by the producers and a demand side showing the wish of customers. In the early years of mankind this was no different. It may however seem strange to compare a modern economy with ancient trading but it illustrates an important point that is fundamental for understanding supply chains: we need others to work for us so one can specialize. In early ages this was achieved by people keeping cattle for milk, meat and skin while others would focus on agriculture. Economies were much simpler those days but even at that point a supply chain would exist. People would need clothes to wear and food to eat. Cooks would rely on the meat from farmers to provide the food and tailors would buy the skin from farmers to make clothes. This transaction of goods is a simple, but existing supply chain.

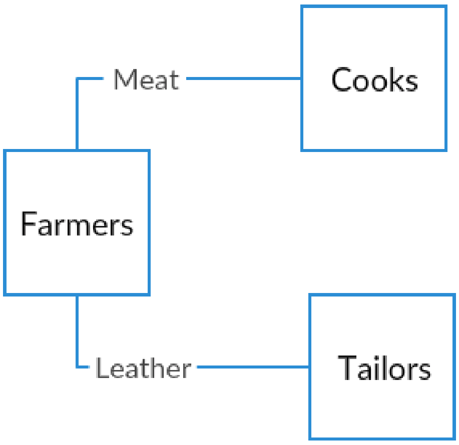


Figure 2.1: A basic supply chain.

Supply chains nowadays have become a bit more complicated but are still relying on the same, previously illustrated, principle. A company needs goods or services from another company that once again relies on someone else.

An example of modern products that show how supply chains have evolved over time are smartphones. A supply chain in this industry, is much more complex than the previous example. These organisations rely on manufacturing done by other companies as they cannot do everything themselves. Screens for the phones need to be produced together with processors and other electrical components. Once the final product is assembled, it needs to be sold worldwide and for this a global distribution network is required.

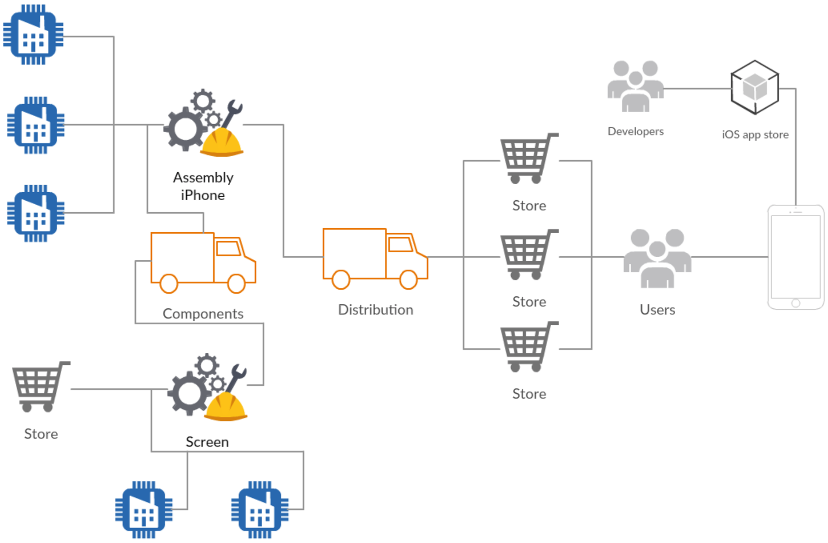


Figure 2.2: A simplified version of a smartphone supply chain

This supply chain is no different than that of the very simple other example when disregarding the complexity. The key issue with supply chains is however the complexity. The more complex the supply chain, the harder it is to manage it correctly. Obviously, a poorly managed supply chain will lead to organisational problems. Supply chain management (SCM) is focused on ensuring a correctly managed supply chain.

2.2 Supply Chain Management (SCM)

The supply chain management (SCM) concept is relatively new, making an appearance halfway the 1980's [1]. Theories that SCM rely on, such as the management of operations between multiple organisations [2] and systems integration [3], have already been introduced two decades earlier. Multiple definitions of SCM have been given throughout literature. Houlihan set the goal as one to "lower the total amount of resources required to provide the necessary level of customer service to a specific segment" [4]. Some researchers extend this statement a bit by stating that SCM should also create a competitive advantage over organisations [5]. By applying correct SCM an organisation can focus on their main business processes while confidently knowing that their dependency on other companies is well managed. The decision to rely on others allows an organisation to specialize which ultimately leads in a cost reduction as each involved party does what he can do best. By doing so, a competitive advantage is created [5].

The concept of competitive advantage was widely discussed by Porter [6] and summarised in the following value chain:

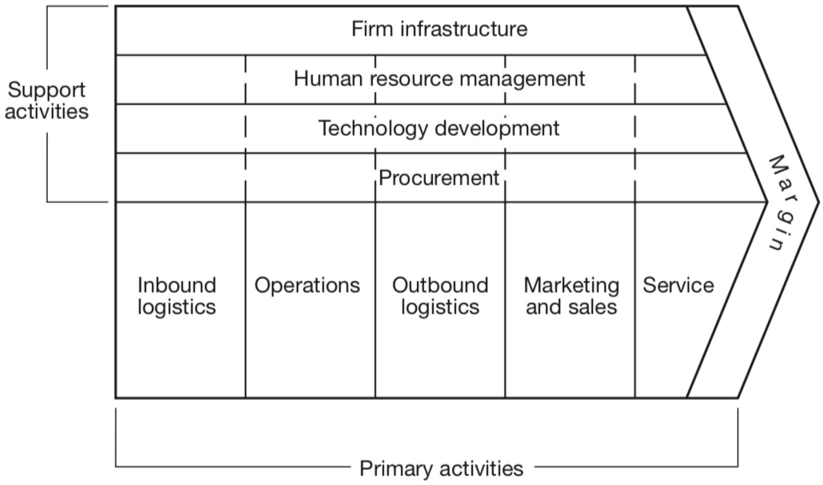


Figure 2.3: Porter's value chain [6]

This value chain consists of primary and support activities. The former are focused directly on the product or service created while the latter are processes that exist to support the primary one and enable correct execution. Porter stated that an organisation can gain a competitive advantage if it executes these activities more efficiently than their competitors.

The real question that arises, now that has been established that the goal of SCM is to gain competitive advantage, is whether one does actually gain a competitive advantage when applying SCM. This matter has been researched Li et al [7] by answering the following three hypotheses:

1. Firms with high levels of SCM practices will have high levels of organizational performance.

2. Firms with high levels of SCM practices will have high levels of competitive advantage.

3. The higher the level of competitive advantage, the higher the level of organizational performance.

By introducing a model to test the hypotheses, data was collected from 196 organisations leading to the validation of all three hypotheses. Concluding that higher levels of SCM do in fact lead to better organisational performance and a better competitive advantage, furthermore achieving higher levels of competitive advantage do also improve organisational performance.

To understand on what aspect each party can focus, two important different segments (upstream and downstream) can be identified within SCM [8]. The first, upstream, is relying on companies closer to the raw material sides (eg. the farmer for meat or leather) while the other is closer to the delivery at the consumer (eg. transportation to the customer). Both segments have their own challenges and the company relying on them has no experience with either segment meaning that trying to do any other process than their main business process will be incredibly difficult. A company can become vertically more integrated by doing upstream/downstream processes by themselves but this is generally considered more challenging opposed to integrating horizontally which simply means a company expands its operation in their current industry. [8]

The manner in which an organisation has an integrated supply chain cq is vertically integrated has been presented in a four stage model by Stevens' [9]. This model describes the four levels indicating how much a supply chain is integrated in an organisation.

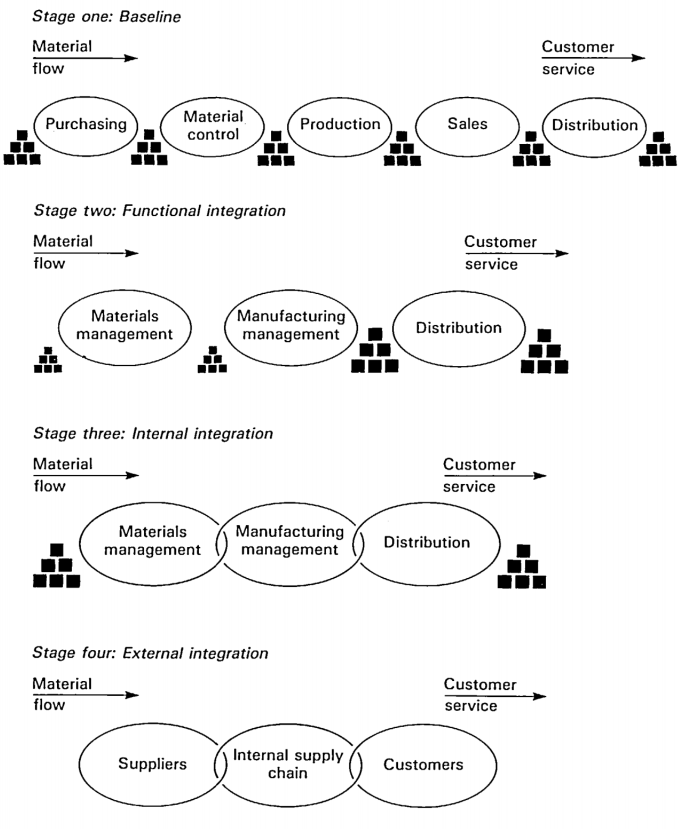


Figure 2.4: Stevens' four stage model [9].

Stage one: Baseline

This first stage is identified by the supply consisting of separate, independent, departments. In this stage, the planning is short term. Due to the present inefficiencies, the effectiveness of the supply chain as a whole is in danger. Characteristics for this stage are: staged inventory and incompatible control systems due to failed integration between the supply chain departments.

Stage two: Functional integration

Learning from inefficiencies in the first stage, supply chains at this level focus on cost reduction by creating business functions that each have an inventory buffer to cope with variations..The focus lies mainly on the inward flow of goods. However, the business still operates rather reactive meaning that they do not truly listen to their customers but just respond to the wishes of the one with the most money. Planning is not done very accurately yet either resulting in poor performance due to real customer demand being unclear.

Stage three: Internal integration

At this phase, it is recognized that focussing on the inward flow of goods is useless when it is not done efficiently. Therefore, now a strategy will be designed that utilises all the available organisational tools allowing the supply chain to be truly integrated within the organisation. At this point the organisation is switching from short term to medium term planning and available tools are used efficiently to focus on having an optimal flow of goods to the customer.

Stage four: External integration

Only at this point full supply chain integration is achieved. By having designated long term plans for all organisational levels with regard to implementing the supply chain true efficiency can be accomplished ensuring that the organisation has the best flow of goods to the customer. At this stage, the organisation has shifted from product oriented to a designated customer-orientation making sure that the customer’s requirements are adhered to and all departments are linked together smoothly rather than being treated as separate entities.

An organisation should continue improving their SCM up to the fourth stage to achieve the highest level of efficiency. By achieving a more mature stage than your competitors is what gives your organisation a competitive advantage over other companies within the industry.

Besides SCM, logistics management is also a relevant concept. This can be defined as *"The process of planning, implementing, and controlling the efficient, cost-effective flow and storage of raw materials, in-process inventory, finished goods, and related information from point-of-origin to point-of-consumption for the purpose of conforming to customer requirements."* [10]. The term 'Logistics' finds it origin in early 19th century warfare book where it was stated that *"Logistics is the art of well ordering the functionings of an army, of well combining the order of troops in columns, the times of their departure, their itinerary, the means of communication necessary to assure their arrival at a named point"* [11].

Given the theory of SCM and this definition of logistics management, it can be understood how they differ from each other. The latter is a subset of the former as within SCM, the management of logistics, is also included. SCM however includes much more than this due to globalisation. An organisation has to work together with a lot more companies for it to be to be able to operate efficiently. It can be said that SCM is focused more on managing the relations between multiple organisation ranging from upstream to downstream [8]. The previously definition about logistics management can therefore be extended by first giving a clear definition for a supply chain: *"A set of three or more entities (organizations or individuals) directly involved in the upstream and downstream flows of products, services, finances, and/or information from a to a customer, (and return)"* [12]. Following this definition along with previously explained theories, the following definition for SCM by The International Center for Competitive Excellence in 1994 will be used: *"Supply chain management is the integration of business processes from end user through original suppliers that provides products, services and information that add value for customers."* [13]. From the previously provided examples in Figures 2.1 and 2.2, it can therefore be concluded that the first (simple) example is merely logistics management whereas the more complex example is actually SCM as multiple entities are integrated into upstream and downstream flow.

The differences between logistics and SCM can be summarised in the table below:

Comparison	Logistics Management	SCM
Meaning	Moving goods around	Managing the relationship between multiple organisations in the supply chain
Objective	Customer satisfaction	Competitive advantage
Origin	1830 [11]	1980's [1]
Involved organisations	One	Multiple

2.3 Challenges

An organisation deals with a lot of different challenges on multiple organisational levels. These vary from operational and tactical levels up to the strategic level which all need to be tackled in order to apply high level SCM in your organisation [14].

Operational level
This level includes all the day-to-day decisions that a company has to make with regard to running the organisations. Examples are the planning of distribution and quotations.

Tactical level
Here decisions are made for the quarter or year. Purchasing and production decisions are tactical decisions.

Strategic level
These include the decisions that have an effect on the organisation that last for a long time. Outsourcing, partnering and plant location are on a strategic level.

Each of these levels have their own challenges with regard to SCM. Consider product design as an example, choosing a specific design for your product can have a huge impact on your supply chain. It may be possible that one specific design increases the inventory holding costs while another design reduces the lead time of manufacturing. When organisations have to make such a strategic decision it is important to take the supply chain in consideration. If your design is very complicated and you have to rely on a lot of different suppliers, what would happen if one supplier fails to deliver? If customer demand suddenly drops, can you still work together with the same suppliers or do you have to buy a minimum quantity? These sorts of questions affect the decision that management will make. If you cannot fully rely on all the links, it might be a safe decision to limit the complexity of your supply chain. Some mentioned risks can be mitigated by establishing supply contracts in which all scenarios are documented and agreed upon by both parties ensuring a healthy cooperation. However, sometimes the mere fact that one organisation cooperates with another is already a risk. The issue of strategic partnering is one of sharing information. By integrating your supply chain with other organisations you have to share information but what has to be shared to make cooperation successful? All these challenges have to be overcome for SCM to have a positive outcome.

Another major challenge with regard to SCM is a phenomenon first observed by Forrester in 1961: the bullwhip effect [15]. The observation stated that supply chains have inefficiencies due to increasing fluctuations in the upstream part of the supply chain. Meaning that predictions about inventory closely to the end consumer can be done rather precise but the further away from the consumer one tries to predict inventory, the more off the prediction will be. In research by Lee et al. the following main causes of the bullwhip effect were examined [16]:

- Demand forecast updating
- Order batching
- Price fluctuations
- Rationing and shortage gaming

In the same research, methods to counteract the causes are given as the bullwhip effect is a fundamental threat to effective SCM. The conclusion given by Lee et at. is that *"the choice for companies is clear: either let the bullwhip effect paralyze you or find a way to conquer it"* [16].

Next to the bullwhip effect one other important challenge for SCM is risk pooling [14]. This effect is a statistical suggestion that variability and uncertainty can be reduced by aggregation. In the context of supply chains this means that variation in demand can be reduced by aggregating demand. If your demand is aggregated over different locations, there is a higher chance that variations in one location are mitigated due to similar, opposite, variations in another.

With all the above mentioned challenges for SCM it is relevant to show that the risks are real and it is crucial to mitigate these in order to implement SCM effectively. Chopra et al. broke down the supply chain based on a case study exploring many supply chain related risks [17]. Two examples provided at the beginning directly show the importance of SCM. The first example is that of Nokia. The company was one of the major customers of a chip plant that faced supply issues after a fire. Due to their multiple-supplier strategy, Nokia could efficiently switch to other suppliers and continue production without too much damage. Ericsson was however another major customer of the same chip plant but due to their single-sourcing strategy they could not switch to a new supplier and suffered greatly. Learning from this event, they eventually switched their strategy, showing the challenge of risk pooling in practice. With a critical node analysis, Ericsson might have been able to detect this risk beforehand, preventing an impact on this scale.

2.4 Developments

Now that the supply chain concept has been introduced it is good to identify how supply chains have evolved over the more recent years. Understanding the recent changes of supply chains contributes to the comprehension of why critical node analysis is a relevant issue that management should be able to address. In previous sections, it has been clearly shown why SCM is important and what the risks are of not implementing this effectively. Christopher illustrated how the competitive environment is changing [1]. According to this book, the four main changes are:

- New rules of competition: main competitive advantage can be achieved by having high levels of SCM
- Globalisation: supply chains will increase in size, making SCM more complex
- Downward pressure on price: organisations have to become cheaper and cheaper causing them to only focus on their main business process and needing to outsource

the rest making the supply chain more complex and inevitably, SCM will be more difficult

- Customers control: the position of customers is becoming more important, forcing organisations to become customer oriented as in the fourth phase of Stevens' model [9].

These four developments all have to same effect: SCM will become more and more important over the next years. With the help of critical node analysis, higher levels of SCM can be achieved ensuring continuing existence of the organisation in the global market.

2.5 Methods of analysis

Multiple methods exist to analyse supply chains [18]. These can be:

- Deterministic analytical models: in which the variables are known and specified
- Stochastic analytical models: where at least one of the variables unknown, and is assumed to follow a particular probability distribution
- Economic models: a game theory approach
- Simulation models: evaluate effects of different strategies

There are five main strategies when using simulation models [18] of which one is to reduce time delays at each stage of the supply chain. Another strategy aims at improving decision rules within the supply chain [19]. When decision rules are optimised within the supply chain and greater efficiency is achieved, higher levels of integration can be achieved as defined in Stevens' four stage model [9].

The eventual goal of this thesis is to analyse the performance of a supply chain. Given the strategies related to simulation models, it would be best to apply simulation techniques. One of the methods to simulate the supply chain is with the help of graph theory [20]. Using a graph related approach, netchains are an introduced concept. A netchain describes how all parties involved in a supply chain are connected with each other. Not only the horizontal relationships between multiple suppliers can be drawn but also the vertical ones between suppliers and buyers.

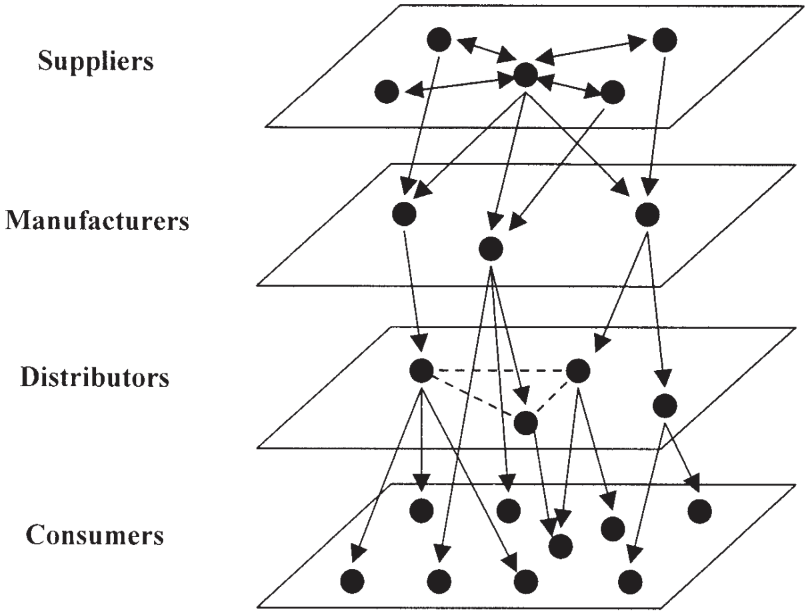


Figure 2.5: A generic netchain [20].

With netchains as depicted in Figure 2.5, the desired analysis of supply chains can be achieved. By using a layered structured, the performance of every single aspect within the supply chain can be measured. This is exactly what will be done in this thesis. Now that it has been shown that with graph theory a detailed analysis of supply chains can be conducted, the next chapter will elaborate all the relevant methodologies within graph theory that support such an analysis.

Chapter 3

Network Topology

3.1 Theories

After establishing what supply chains are and how they work, another important concept has to be introduced: networks.

In 1741, Leonhard Euler published 'Solutio problematis ad geometriam situs pertinentis' [21] in which the first foundation for graph theory was accomplished. In this paper Euler presented the (now famous) 'Seven Bridges of Königsberg' problem. The problem he was presented with was that of verifying if it was possible to walk through the entire city of Königsberg (now Kaliningrad, Russia) while only crossing each of the seven bridges exactly once. For this issue there was not yet a known solution and therefore Euler is credited significant for his contribution.

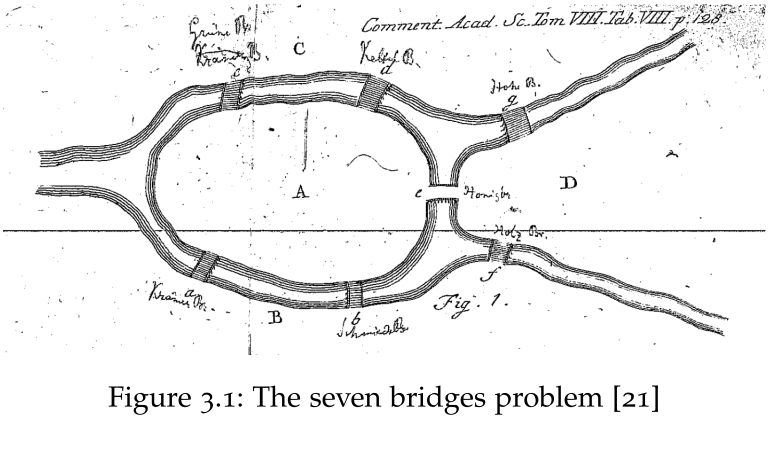


Figure 3.1: The seven bridges problem [21]

Even though modern terms used for describing networks were not used in this paper, it is still regarded to be the first problem description concerning a real network.
In the seven bridges problem, there were four sections of land which had to be visited. The bridge is called an edge and the land is a node. The combination of these two form a graph which can mathematically be denoted as $G = \{V, E\}$ where G is the graph, V is a set containing all the nodes and E is a set containing sets of edges. This can be elaborated with a simple example where the graph below can be considered:

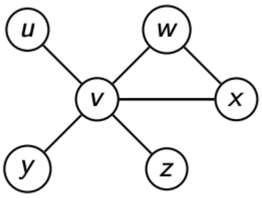


Figure 3.2: A simple graph

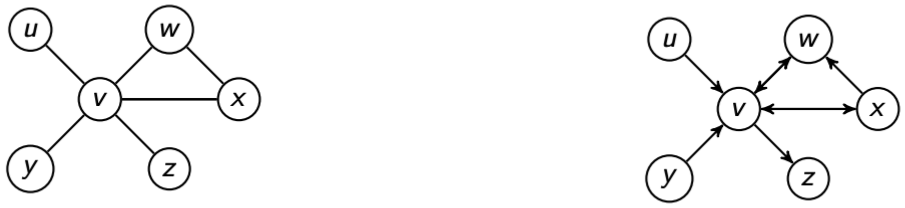
In this graph $G = \{V, E\}$ with:

- $V = \{u, v, w, x, y, z\}$
- $E = \{\{u, v\}, \{w, v\}, \{v, x\}, \{x, w\}, \{y, v\}, \{v, z\}\}$

The amount of nodes (n) is de amount of elements in V meaning that $n = 9$ and similar, the amount of edges (from E) gives $m = 6$. The rather simplistic approach in describing graphs will be of great use in later chapter when later explained methodologies will be used to analyse supply chains.

There are different types of graphs as the directionality, link type, metadata, mode and temporariness change from network to network [22].

Directionality
The directionality of a graph refers to the presence or absence of a specified direction for the edges. Given $E = \{\{a, b\}, \{b, a\}\}$, E is commutative in an undirected graph but not in a directed graph. When using the earlier provided simple graph, the difference between these two types of directionality can be clearly seen.



(a) An undirected graph (b) A directed graph

Figure 3.3: Sample graphs

Link type

There are three types of links for a given graph G:

- Weighted: each edge has a specific weight assigned allowing a specific link between nodes to have more impact than another. This can be the case with a collaboration graph, each edge between nodes indicates how often the people (nodes) have worked together. The higher the weight of an edge, the higher the frequency of collaboration between them.

- Unweighed: every edge in the network is treated equally. A friendship network can be unweighed as a link could indicate the mere fact two people are friends.

- Signed: the edges have a positive or negative effect of equal strength. Within a friendship graph, a signed version can show who you do and do not like.

Metadata

To add extra information to graphs it is possible to add metadata to the network. This can be done by annotating nodes and/or edges.

Mode

A graph can have one, two or more modes. A mode is defined as the amount of categories that a node can have.

- One-mode (homogenic): every node has the same meaning. In the friendship graph this means that every node simply represents a person

- Multi-mode (heterogenic): there are two (or more) categories that a node can possibly belong to, but only one at a time. A two-mode friendship graph could consist of nodes indicating the institutions through which friends have met with edges to nodes representing the individuals.

Temporariness
A graph can be either static or dynamic. With a static graph, nodes/edges do not have timestamps while with a dynamic (temporal) graph they do. With the existence of timestamps in a graph, it is possible to do an analysis on the data at a specific point of time with the exact representation of the network at the selected time.

3.2 Metrics

For a given graph G it has been shown that there are a lot of different types. Merely counting the amount of nodes and edges within a certain network does not provide a detailed analysis. Making a distinction on other properties as defined in the previous section will also not contribute a lot to a comparison of different graphs. To facilitate an analysis between graphs, certain metrics have been established throughout the literature [22].

3.2.1 Density, Degree & Distance

The density of a graph indicates how much edges are present compared to the theoretical maximum. If the density is high, the graph is called dense as opposed to sparse when there are relatively few edges. Whenever a graph consists of the maximum amount of edges, it is called a complete graph. The maximum amount of edges (m_{max}) for both undirected and directed graphs can be defined:

- Undirected: $m_{max} = \frac{1}{2}n(n-1)$
- Directed: $m_{max} = n(n-1)$

This number can be understood easily as $n(n-1)$ represents that every node is connected with each other node. The commutative property of edges in undirected graphs results in m_{max} being twice as small. A graph is sparse whenever $m \ll m_{max}$. The density is the relation between the number of edges and the theoretical maximum. The density for graph G:

$$den(G) = \frac{m}{m_{max}} \tag{3.1}$$

For undirected graphs:

$$den(G) = \frac{|E|}{2|V|(|V|-1)} \tag{3.2}$$

and directed graphs:

$$den(G) = \frac{|E|}{|V|(|V|-1)} \tag{3.3}$$

Whenever a graph is dense, the average amount of nodes with a high degree will also be high. The degree of a node is the amount of connected edges to that node. Given the previous undirected graph in Figure 3.3a, the $deg(v) = 5$. For directed graphs, two variations on degree exist. The indegree indicates how many edges point to the node while the outdegree shows the amount of outward pointing edges. For the directed graph in Figure 3.3b the $indeg(v) = 4$ and $outdeg(v) = 3$.

Consider the same undirected graph G. A path is any possible walk over de edges. An example path $p = (v, w, x, v, y)$. The length of the path equals $|p| - 1$ which results to 4 in this case. A simple path is one in which there are no edges traversed twice [23]. Another important (perhaps the most important) aspect of path traversal is finding the shortest path. The shortest path from u to x is $sp = (u, v, x)$, the distance

$$d(u, x) = |sp| - 1 \tag{3.4}$$

Finding the shortest path in a large network is a very challenging exercise even with widely accepted algorithms such as Dijkstra [24] and Floyd-Warshall [25] but due to their complexity of $\mathcal{O}(mn)$ per node and thus $\mathcal{O}(mn^2)$ for the entire network it becomes a rather expensive operation to run on a large network. The average distance of a graph can be computed by

$$\bar{d} = \frac{1}{n(n-1)} \sum_{v,w \in V} d(v, w) \tag{3.5}$$

For the average distance calculation, the shortest path has to be found first. This can be used to compare graphs and provides additional information together with density and degree statistics. These more complex metrics help to establish methodologies to generate a good comparison. However, there are still more metrics to be discussed. Based on distance, two more metrics can be introduced: diameter and eccentricity.

Eccentricity
Given a node u, the eccentricity of that node equals the length of the longest shortest path from u to any other node $v \in V$.

$$e(u) = max_{v \in V}\ d(u, v) \tag{3.6}$$

Diameter
The diameter is the maximum possible distance between two nodes while taking the shortest paths between them.

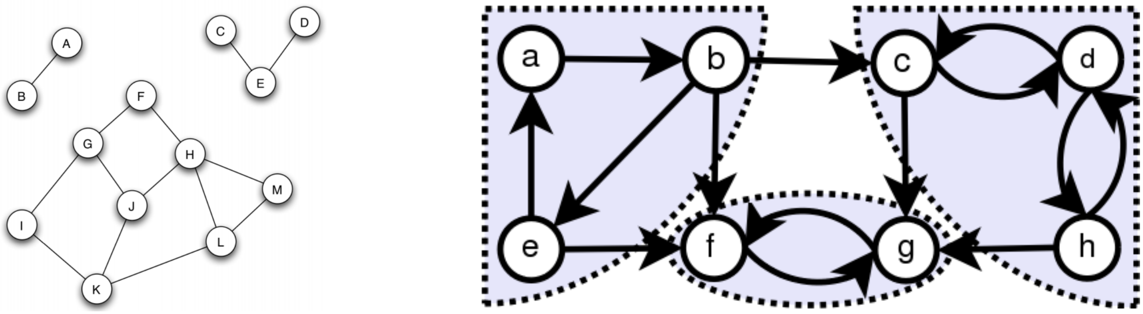
$$D(G) = max_{u,v \in V}\ d(u, v) \tag{3.7}$$

Using the introduced concept of eccentricity, the diameter can also be defined as follows:

$$D(G) = max_{u \in V}\ e(u) \tag{3.8}$$

3.2.2 Components & Communities

If the distance between two nodes x and y is infinite, this means that components exist. For an undirected graph, there are 'islands' in the network as can be seen in Figure 3.4a [26] while in a directed graph, there is simply no traversal possible between the two given nodes while obeying the direction of the edges. In Figure 3.4b there is no path between the nodes g and a. It can also been seen that there are three components clearly identified.



(a) An undirected graph with components

(b) A directed graph with components

Figure 3.4: Graphs with components

A connected component is a subset of nodes ('island') where there is a possible path from each node to any other node even though this does not have to be directly as in a connected graph. The component with the most nodes is called the giant component [27].

3.2.3 Clustering coëfficient

Consider the graph G in Figure 3.5.

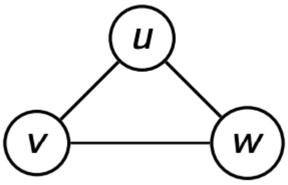


Figure 3.5: A triangle

For a node v, the amount of triangles is

$$\delta(v) = |\{\{u,w\} \in E : \{v,u\} \in E : \{v,w\} \in E\}| \tag{3.9}$$

but as a triangle contains three nodes, summing over all the nodes results in counting three triangles. Therefore, when calculating all the triangles in a graph G

$$\delta(G) = \frac{1}{3}\sum_{v \in V} \delta(v) \tag{3.10}$$

Another relevant concept to be understood before introducing the actual clustering coëfficient in a triplet Y [28]. This is for a node v a path of length two where the node v is at the center and the amount of triples for a node can be calculated by

$$v(v) = (\frac{deg(v)}{2}) \tag{3.11}$$

For an entire graph G, the number of triples equals

$$v(G) = \sum_{v \in V} v(v) \tag{3.12}$$

The clustering coëfficient follows from the now defined triangles and triples [29]. A distinction can be made between an overall indication of the clustering in a network and the degree of which a single node is embedded.

Local clustering coëfficient

Given a single node v, this value can be calculated by

$$C(v) = \frac{\delta(v)}{v(v)} \tag{3.13}$$

Global clustering coëfficient

For the entire graph G, the clustering coëfficient C can be calculated by using V': the set of nodes $v \in V$ with $deg(v) \geq 2$ and the local clustering coëfficient as defined in equation 3.13.

$$C(G) = \frac{1}{|V'|} \sum_{v \in V'} C(v) \tag{3.14}$$

It has been shown Newman et al. that this is the Transitivity of a graph [30] and is defined as

$$T(G) = \frac{3 * \delta(G)}{v(G)} \tag{3.15}$$

3.2.4 Centrality measures

There are three main measures of centrality that will be introduced. The idea of all these measures is creating a method that identifies the importance of a single node and more importantly compare it to other nodes in the same network [31]. Given a centrality measure C_M, all nodes $v \in V$ will return a value $C_M(v) \in [0;1]$. Given two nodes v and w, if $C_M(v) > C_M(w)$ it implies that node v is more important than w. The only exception is closeness centrality, here a smaller value indicates a higher level of importance.

The measures that will be discussed are degree, closeness and betweenness centrality as introduced by Freeman [32].

Degree centrality
The degree has previously been defined in this section. The centrality for this metric equals:

$$C_d(v) = \frac{deg(v)}{n-1} \tag{3.16}$$

This local measure indicates how many adjacent nodes are connected to a specific node v taking all the other $(n-1)$ nodes into account. As a variant on this measure, directed graphs also have the indegree and outdegree centrality measure.

Closeness centrality
This is the average length of all shortest paths between one specific node and all the other nodes in the graph. The lower the closeness centrality, the more close the node is to the rest of all the nodes.

$$C_c(v) = (\frac{1}{n-1} \sum_{w \in V} d(v,w))^{-1} \tag{3.17}$$

This is the reciprocal of the farness [33]

$$C_c(v) = \frac{1}{\sum_{w \in V} \; d(w,v)} \tag{3.18}$$

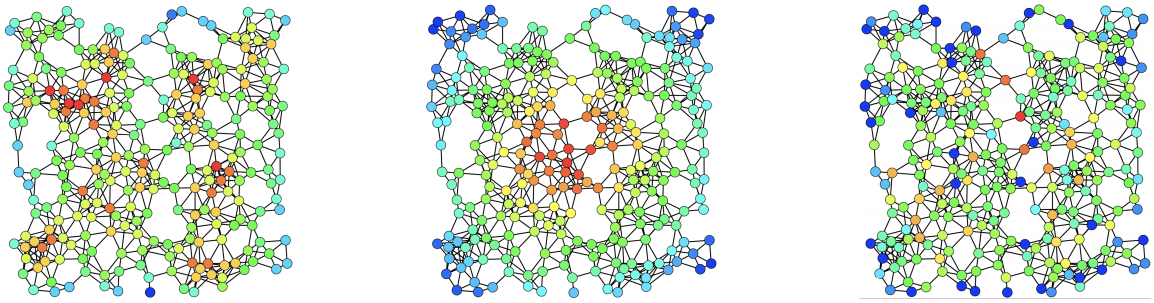
Betweenness centrality

To find nodes that link components together the betweenness centrality measure can be used. This shows how many shortest paths pass through a specific node.

$$C_b(u) = \sum_{v,w \in V} = \frac{\sigma_u(v,w)}{\sigma(v,w)} \tag{3.19}$$

Where $\sigma_u(v,w)$ is the amount of shortest paths from v to w through u and $\sigma(v,w)$ is simply all the shortest paths between these nodes. Furthermore: $v \neq w, u \neq v, u \neq w$.

The previously three discussed centrality measures are summarized in Figure 3.6. Warmer colours indicate a higher ranking for the relevant centrality measure.



(a) Degree (b) Closeness (c) Betweenness

Figure 3.6: Centrality measures compared [22]

To further elaborate the concept of these centrality measures, consider the popular tv serie *Game of Thrones*. Based on the thrid novel (*A Storm of Swords*) of the original written series, Beveridge and Shan [34] composed centrality measures. Whenever two characters in this novel were mentioned within 15 words of one another, a relationship between them was formed. The network formed is summarized below in Figure 3.7.

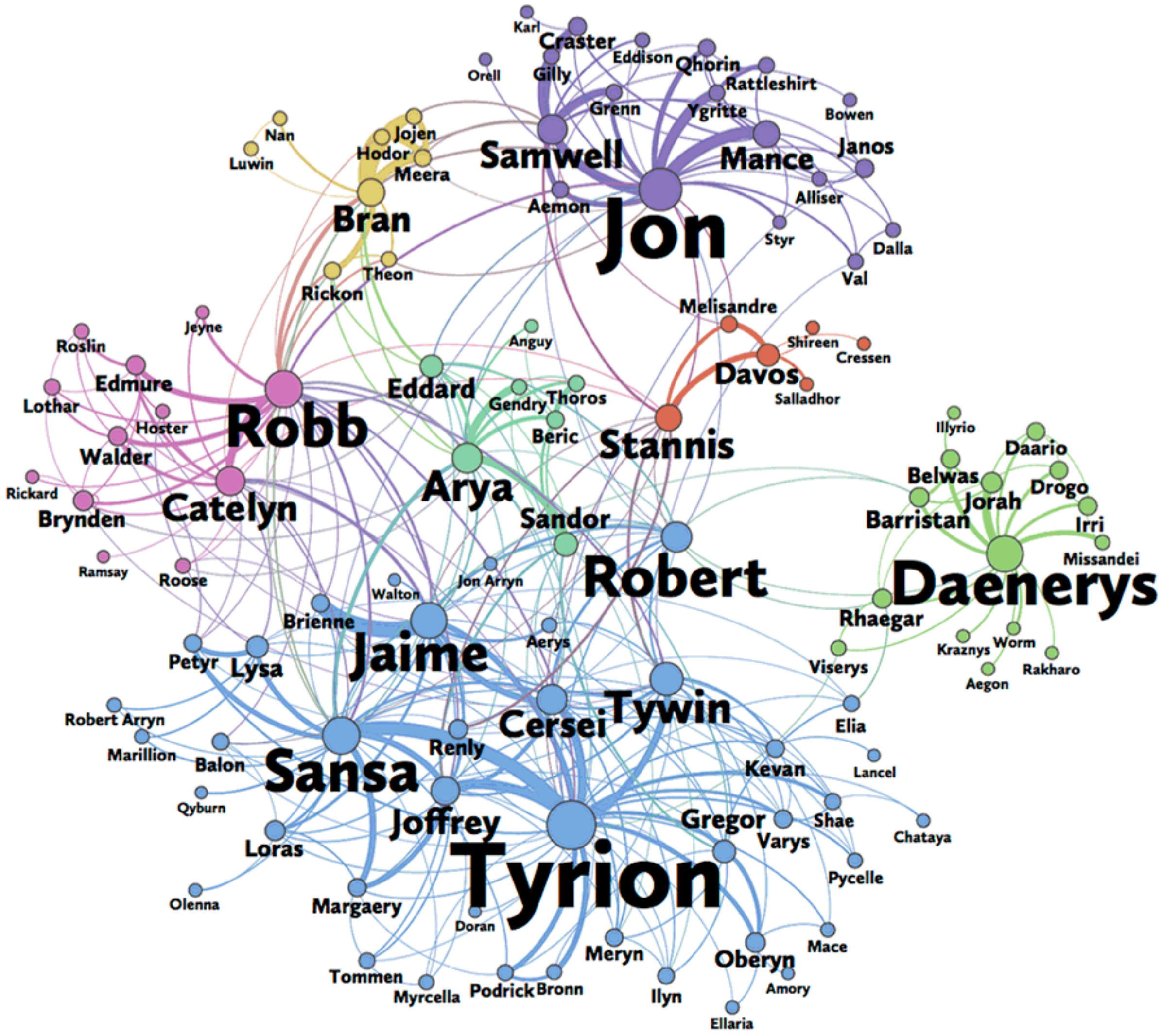


Figure 3.7: Social network based on *A Storm of Swords* [34]

In this network there are 107 nodes with in total 353 weighted edges. The higher the weight, the more frequent those two names are mentioned within the required 15 words and the thicker an edge will be. Furthermore, the colours indicate the community that a node belongs to. In this figure the node size is determined by the PageRank value and label size by betweenness centrality. The results for all centrality measures are shown in Figure 3.8.

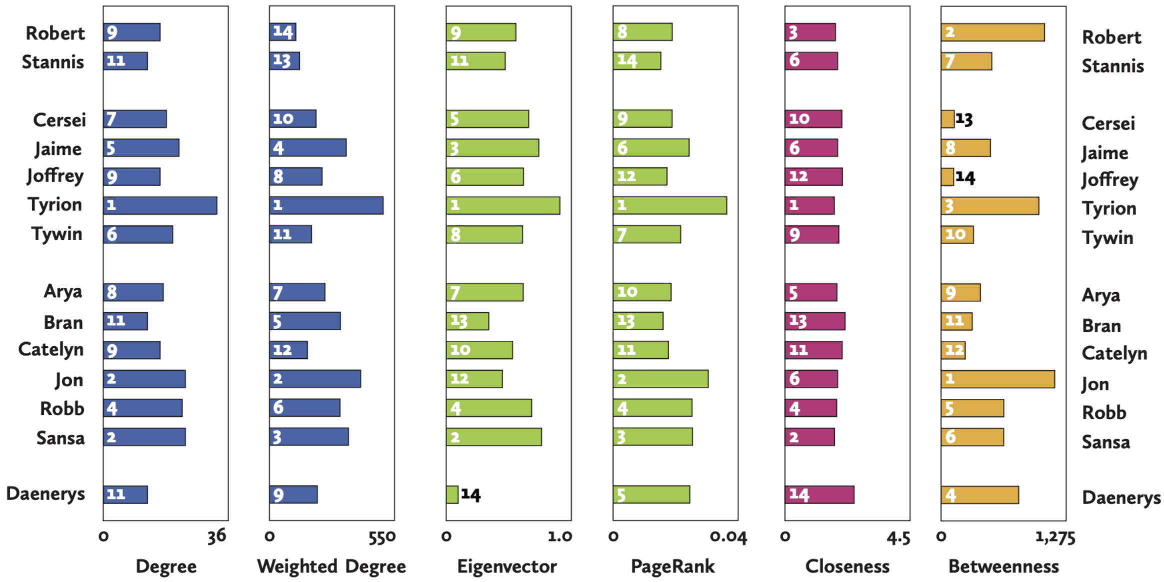


Figure 3.8: Results of computed centrality measures [34]

The numbers in the bars indicate the ranking. The difference between degree and weighted degree centrality is that the former indicates the number of connections to other characters while the latter shows the amount of interactions with different characters. Tyrion has the highest degree (weighted & unweighted) and closeness ranking. Showing that he has most connections to others and most interactions with them while also being most centrally located in the entire network. Jon has however the highest betweenness ranking which indicates that he links most communities together. It can therefore be concluded that Tyrion and Jon are the most influential characters in this novel based on network centrality measures.

3.3 Sample calculations

Given the methodologies as mentioned in this chapter a example can be provided. All the theories and metrics will be applied to the example as shown below in Figure3.9. Calculations are made using a Python program which is the first script in the appendix.

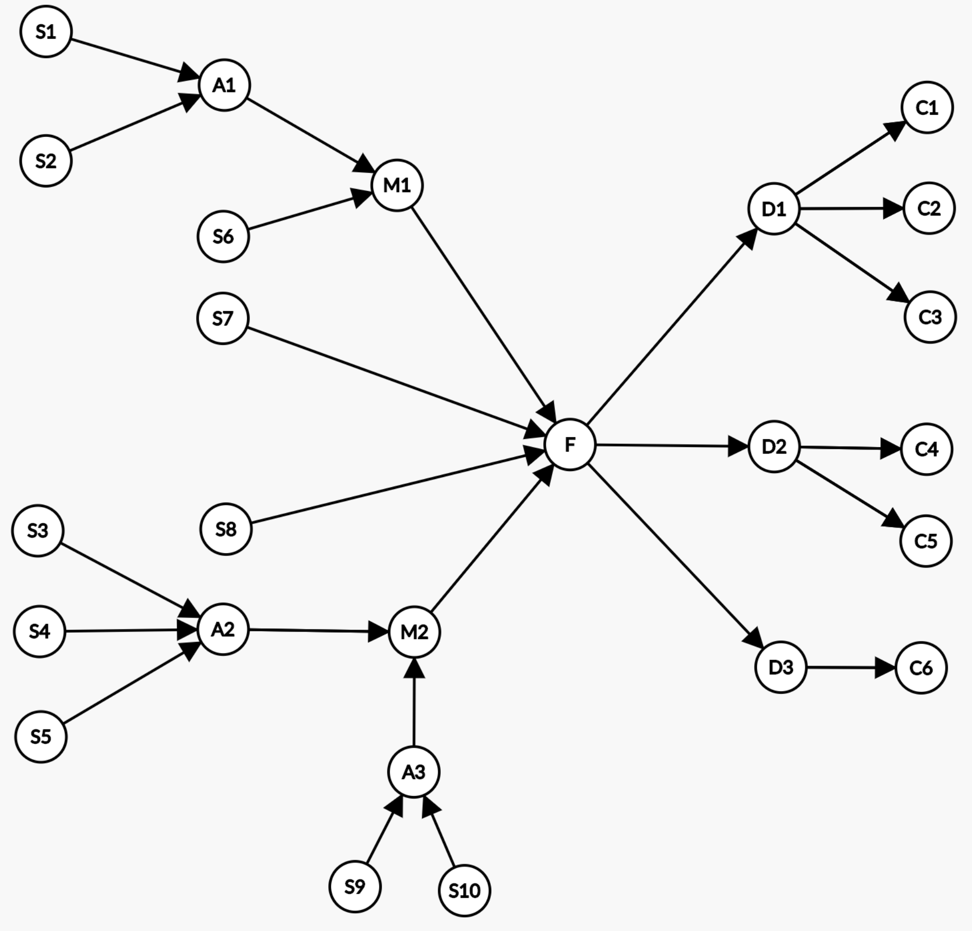


Figure 3.9: Simple supply chain network

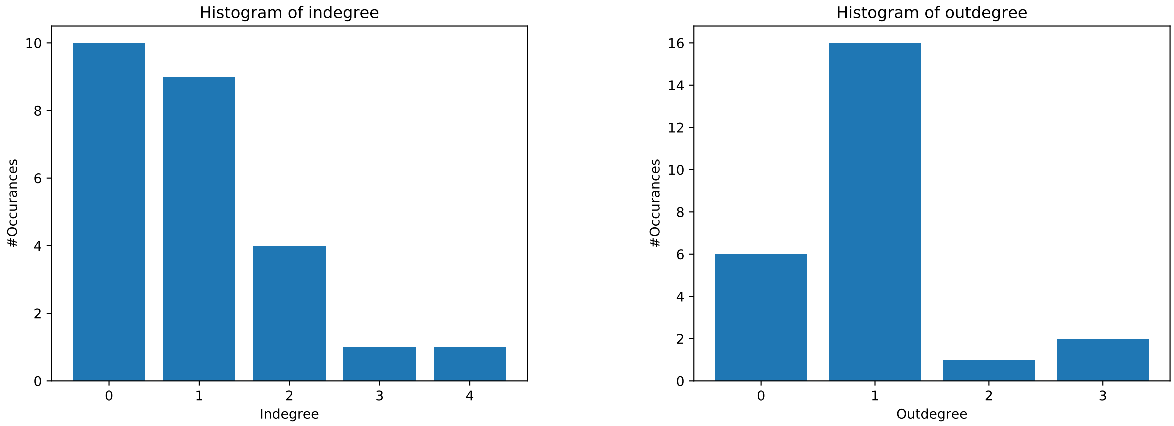
In this network there are different type of nodes that (partly) represent the netchain theory as described in Section 2.5.

- S: Suppliers, delivering equipment to other parties
- A: Assemblers, put together small pieces of equipment
- M: Manufacturers, create larger parts of the product
- F: Factory, where everything is put together
- D: Distributors, regional center where the products are sold
- C: Customers, the final destination of the product

The network contains 25 nodes and 24 edges and has a density of 0.04. This means that the graph is not strongly connected as can be seen when looking at Figure 3.9 and comparing the number of nodes / edges. As previously mentioned, for a graph to be strongly connected the number of edges needs to be much greater than the amount of nodes.

Degree centrality

As the graph in this example is a directed graph, a distinction can be made between the indegree and outdegree.



(a) Indegree distribution

(b) Outdegree distribution

Figure 3.10: Degree distributions

In Figure 3.10 the distributions of indegree (3.10a) and outdegree (3.10b) are shown. The top 5, highest ranking nodes are:

Node	Indegree
F	4
A2	3
A1	2
M1	2
M2	2

Node	Outdegree
F	3
D1	3
D2	2
S1	1
A1	1

Table 3.1: Top ranking nodes (indegree and outdegree)

The results as shown in Table 3.1 show that the factory has the highest degree value for both indegree and outdegree. It makes sense that assemblers and manufacturers have a higher indegree value (as they produce) while the distributors have a higher outdegree value due to their distribution role.

Closeness & betweenness centrality

Besides the degree measures the closeness and betweenness values can be calculated for the provided example. The top 5 for both measures are provided in Table 3.2.

Node	Closeness
F	0.284
D1	0.218
D3	0.218
D2	0.218
C1	0.182

Node	Betweenness
F	0.245
M2	0.127
D1	0.087
M1	0.0725
A2	0.0598

Table 3.2: Top ranking nodes (closeness and betweenness)

The closer the values are to 1, the more close/between the node is.

Distance

With regard to the distance, a distribution can be plotted showing how often a specific path length can be found.

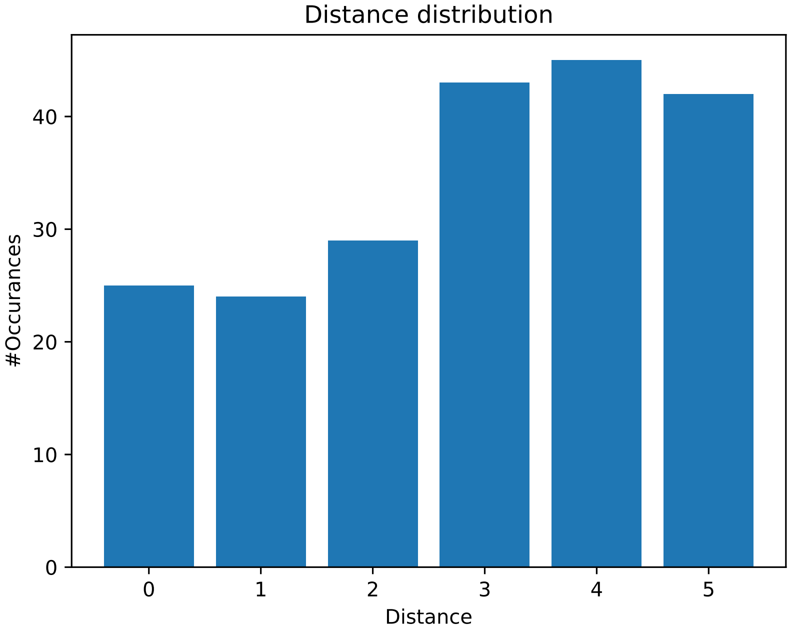


Figure 3.11: Distance distribution

The maximum path length possible is 5. Due to the fact that the graph is directed, the computation of the diameter or eccentricity values is not directly possible as the graph is not strongly connected. Therefore the values are infinite as not all nodes are reachable from every other node. In this network, there are no components. Due to the absence of triangles, the clustering coëfficient is zero.

Given the results to the provided example in Figure 3.9 a interpretation needs to be made. What node is the critical node, where are the bottlenecks and what optimisation is possible? The next chapter will provide insights in how this can be identified.

Chapter 4

Critical Node Analysis

In the previous two chapters the fundamental concepts of this thesis have been established. Now that supply chains and network topology have been made clear, it this necessary to link these two. This enables the derivation of what critical node analysis is.

4.1 Theory

A supply chain is a linear relationship between multiple organisations or processes within an organisation that flows from upstream operations to downstream delivery [35]. The relationships between the multiple parts of a supply chain is simply a large network mapping dependencies.

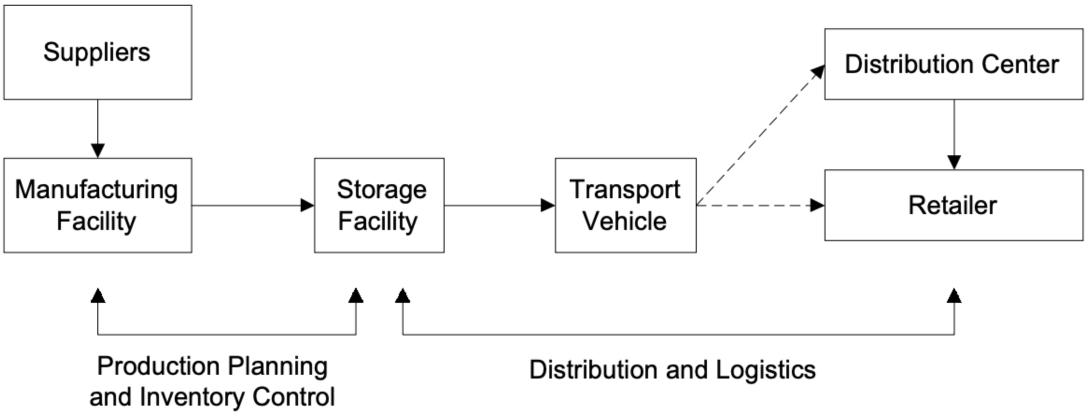


Figure 4.1: The supply chain structure [18]

For an organisation it is crucial to know what processes are essential to their operational performance as often huge losses are suffered due to inefficiencies in supply chains [36].

There are huge possibilities for using network topology methodologies to analyse supply chains [37]. Multiple metrics have already been established which rely on methods introduced in the previous chapter. Craighead et al. have shown methods relying on centrality and density [38] to identify critical nodes. When analysing supply chains, a distinction can be made on three different levels [39] with each level having their own relevant methodologies:

- Node-level: such as clustering coëfficient & centrality measures (degree, closeness and betweenness)
- Network-level: density & clustering
- Link-level: including flow type, multiplexity & tie strength

Using the introduced network topology methodologies in Chapter 3 it is therefore possible to make a detailed analysis for the first two levels. Furthermore it is useful to understand the structure of a typical supply chain network.

A supply chain network typically has the same structure as a (small) real-world network meaning that there are three important characteristics [40]:

- A short characteristic path length
- A high clustering coëfficient
- The presence of a power law connectivity distribution

The first property stating that there is a short path length, is small-world problem as identified by Milgram [41] in 1967. Meaning that two random firms in a supply chain network need relatively few steps to reach each other. A high clustering coëfficient means that there is a high probability that two nodes connected to the same other, are also connected to each other.

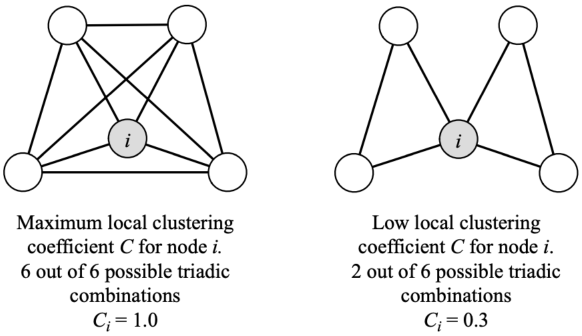

i
i
Maximum local clustering coefficient C for node i.
6 out of 6 possible triadic combinations
$C_i = 1.0$
Low local clustering coefficient C for node i.
2 out of 6 possible triadic combinations
$C_i = 0.3$

Figure 4.2: Clustering coëfficient [40]

Especially in supply chains the presence or absence of these triangles is important [40] as it shows how multiple companies within a supply chain collaborate or merely compete [42].

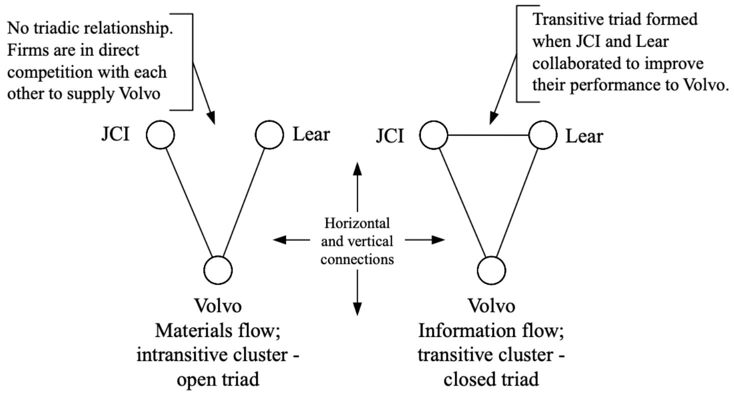

No triadic relationship. Firms are in direct competition with each other to supply Volvo
Transitive triad formed when JCI and Lear collaborated to improve their performance to Volvo.
JCI
Lear
JCI
Lear
Horizontal and vertical connections
Volvo
Materials flow; intransitive cluster - open triad
Volvo
Information flow; transitive cluster - closed triad

Figure 4.3: Volvo case [42]

When analysing the supply chain of Volvo as done by Dubois et al. in Figure 4.3 a change is clearly noted. Initially two organisations were competing to supply to Volvo but eventually they started to cooperate which could be seen by the extra link between the nodes of these two organisations. The effect on the network-level is that the clustering coëfficient increased. The final real-world network property is that of a power law connectivity distribution.

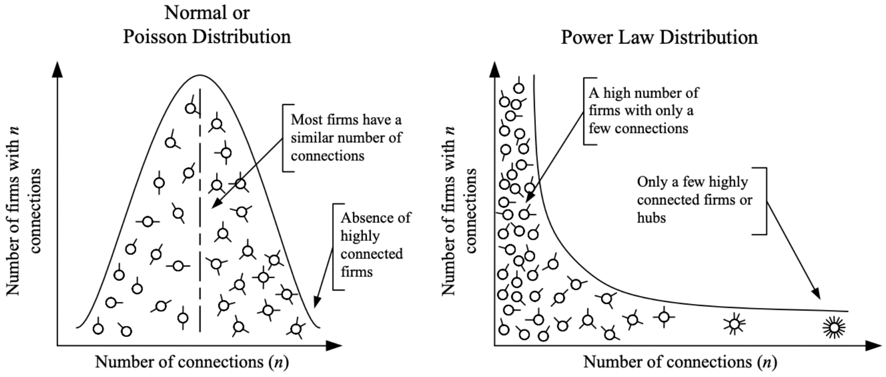


Figure 4.4: Power law connectivity distribution [40]

As shown in Figure 4.4 there are only a few nodes is a supply chain that have a lot of connections to other organisations or processes. These nodes are the hubs in a supply chain.

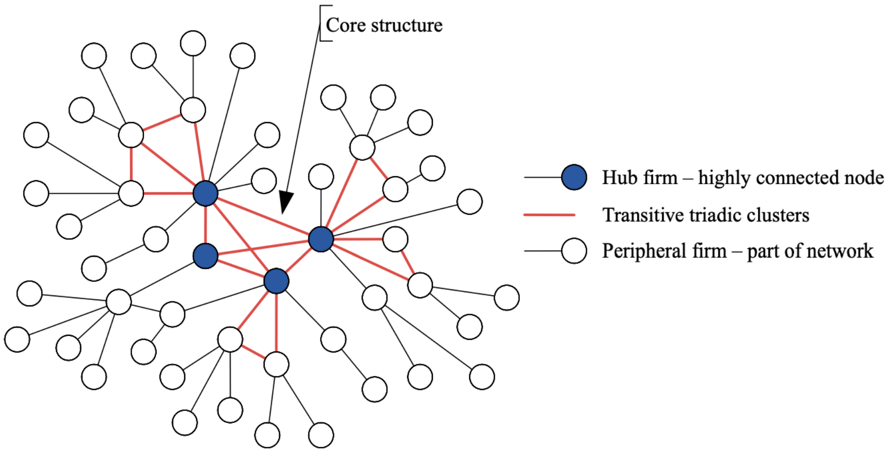

Core structure
Hub firm – highly connected node
Transitive triadic clusters
Peripheral firm – part of network

Figure 4.5: Supply chain with hubs [40]

The hub has a central position in a supply chain and is therefore often a critical node. This because when all the related peripheral firms have no relationship with another hub, the entire supply chain can become jammed. An example of this has already been given in Section 2.3 with the case study of Chopra et al. [17]. A multi-sourcing strategy is a way to limit the impact when a critical node such as a hub fails. All the supplying peripheral firms are still able to distribute their goods to another hub.

The concept of hubs can be illustrated very well with airport connectivity. All over the world there are a lot of cities and villages where one is more remotely located than the other. Yet we all wish to travel all over the world with the help of airplanes. It makes no sense to provide direct flights from a small remote village to for example Amsterdam as this is not cost effective at all. To make flying affordable, it is optimal to transport all people in a region to a specific airport and fly with all of them at once to the destination further away. These airports where the long distance flights depart from are the so called hubs. In the USA, the flights operated by the airliner *Delta* are shown in Figure 4.6.

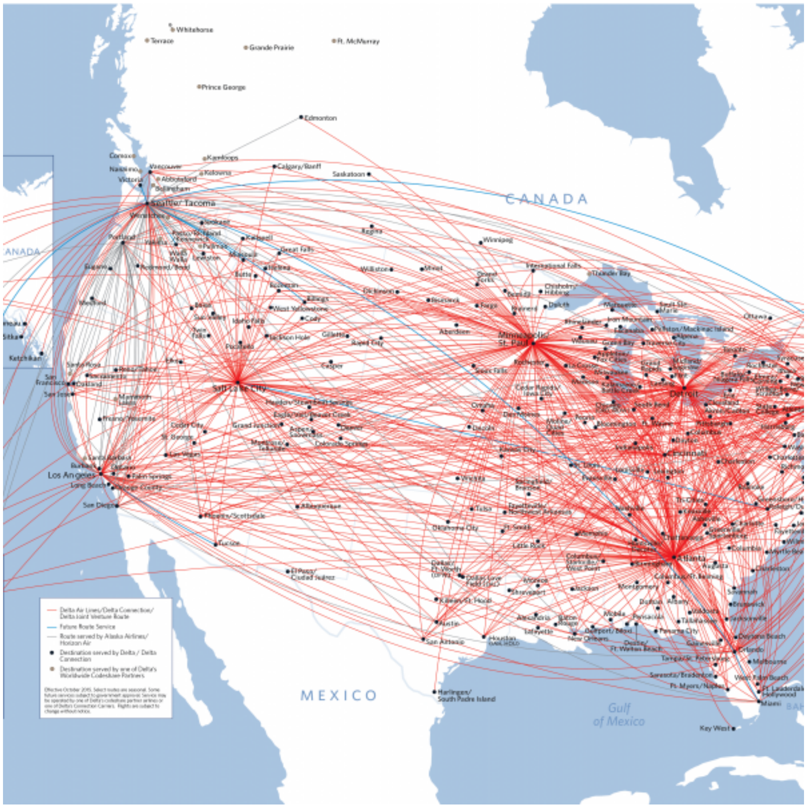


Figure 4.6: Flights operated by Delta in the USA (2015) [43]

No analysis is needed to conclude what hubs are present. By simply looking at the the network, a visual analysis shows that there are 4 clear hubs that are connected with each other and all other smaller airfields are connected to the hub closest by. The main hubs are: Detroit, Atlanta, Minneapolis and Salt Lake City. In 2011, the average amount of airports required to travel from any random airport to another was 3.24 [44]. This number shows the presence of connecting flights achieved by hubs. If the average shortest path length would be closer to 1, it would imply that there are more direct flights. The fact that the value lies above 3 shows that indeed flying from one small airport to another small one is achieved by passing through another airport which likely is the hub.

With these three real-world properties explained, Hearnshaw et al. [40] propose three main statements for supply chain networks:

- Efficient supply chain systems demonstrate a short characteristic path length for all connection types
- Efficient supply chain systems demonstrate a high clustering coefficient for their flows of information
- The connectivity distribution of efficient supply chain systems follow a power law for all connection types as indicated by the presence of hub firms

Given this theory, a sample analysis can be made.

4.2 Analysis

The netchains were introduced in Section 2.5 and sample calculations with network metrics were provided in Section 3.3. In this chapter it has been introduced how these two can be linked. With the results calculated using metrics as shown, a approach is used in analysing supply chains [37].

When obeying the structure of a netchain as described by Lazzarini et al., there are four layers: Suppliers, Manufacturers, Distributors and Customers [20]. The main difference with the example from Section 3.3 is that there are no assemblers or a main factory. However, these two parties can simply be considered a manufacturer. Consider the network shown below.

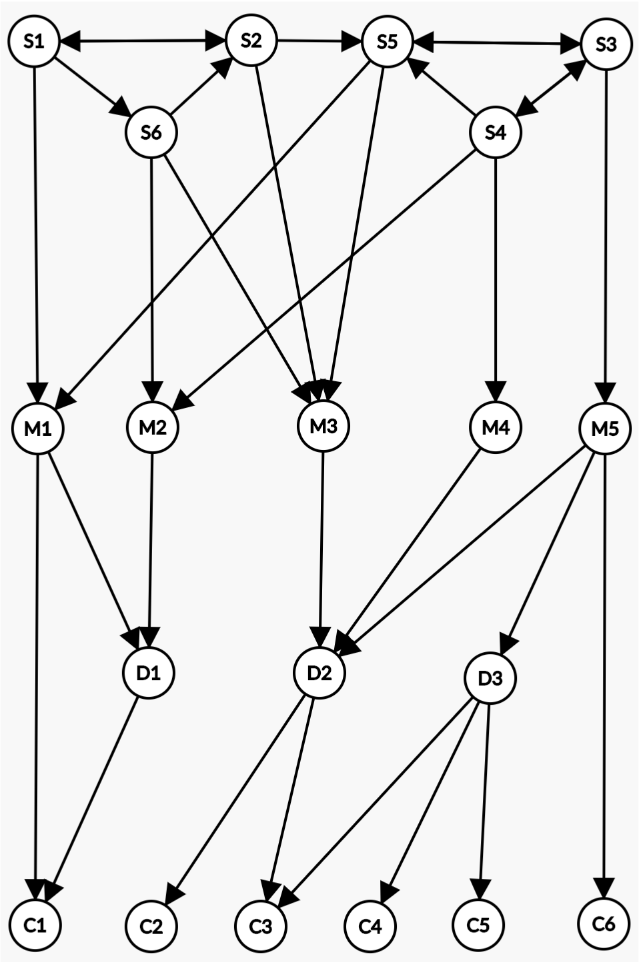


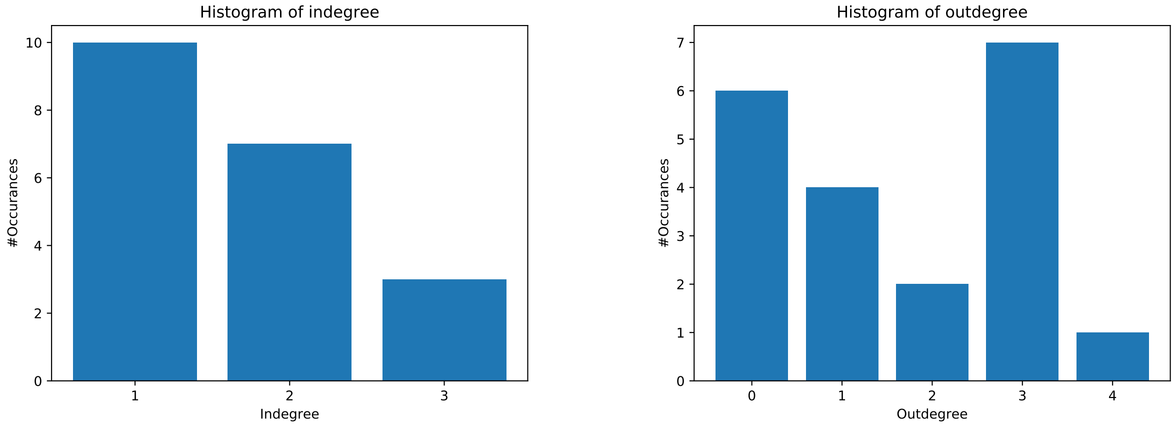
Figure 4.7: Netchain with 4 layers

The example in Figure 4.7 differs is quite some aspects from the previous example in Section 3.3. Calculations are made using a Python program which is the first script in the appendix. Now the number of edges (33) is actually significantly larger than the amount of nodes (20). This causes the density to increase to 0.09 which is slightly more connected than before. The top 5 highest ranking nodes with respect to the same metrics are:

Node	Indegree
S5	3
M3	3
D2	3
S2	2
S3	2

Node	Outdegree
S4	4
S1	3
S2	3
S5	3
S3	3

Table 4.1: Top ranking nodes (indegree and outdegree)



(a) Indegree distribution

(b) Outdegree distribution

Figure 4.8: Degree distributions

Node	Closeness
D2	0.266
C3	0.236
C1	0.213
M3	0.211
D1	0.211

Node	Betweenness
S3	0.0994
S5	0.0936
M5	0.0789
S2	0.0643
D2	0.0497

Table 4.2: Top ranking nodes (closeness and betweenness)

From the numbers it can be concluded that S4 (a supplier) is the most important node when focussing on degree measures. As this node supplies to most parties (4) a failure here would have a high impact. With regard to the indegree measure it can be seen that S5, M3 and D2 rely on 3 other nodes meaning that a failure in one of their supplying sources would cause issues for their operations. When looking at the closeness/betweenness values for this network, there are no remarkable findings as values lie either very close to each other or are very small.

There were a few characteristics of efficient supply chains mentioned in Section 4.1. This example indeed has a short characteristic path length, there are a lot of short paths from suppliers to customers.

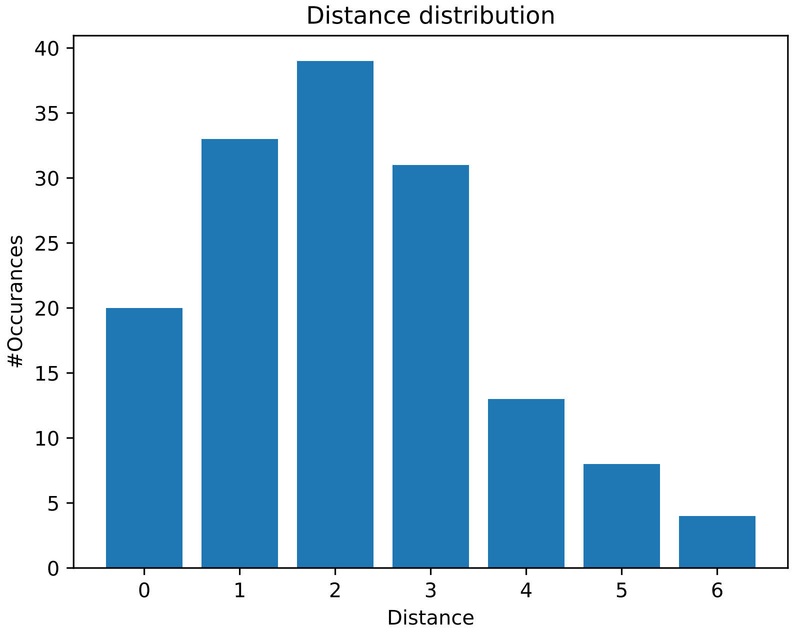
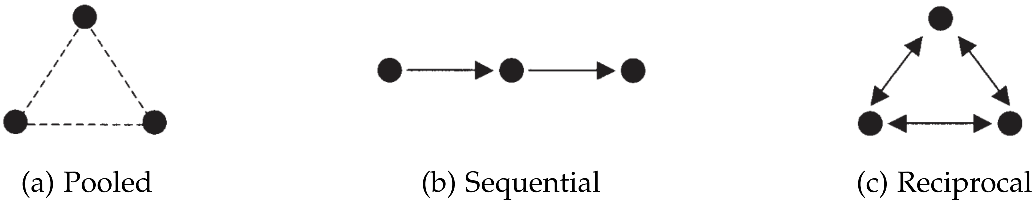


Figure 4.9: Distance distribution

The distance distribution ans shown in Figure 4.9 shows the typical power law distribution. It is not as skewed yet due to nature of being a small example.

Another indication of efficient supply as mentioned was the sharing of information as indicated by the clustering coëfficient within information flows (in the suppliers layer). Three types of information sharing can be identified [20]:


(a) Pooled

(b) Sequential

(c) Reciprocal

Figure 4.10: Types of interdependence

In pooled interdependence, the agents are loosely meaning that they all operate independently with limited knowledge sharing. In a sequential process, the (knowledge) output of one node is the direct input of another. In a reciprocal interdependence, agent's are relying on each other through direct cooperation and information sharing. Within the provided example all different types can be found. Pooled relationships have no connection, sequential once are a linear relationship with no clustering and reciprocal interdependence have a existing clustering coëfficient. The non-zero clustering coëfficients for the network in Figure 4.7 are:

Node	Clustering coëfficient
C1	0.5
S6	0.25
S3	0.25
S4	0.22
S2	0.22
S1	0.2
S5	0.18
D1	0.17
M3	0.17
M1	0.08

Table 4.3: Nodes with non-zero clustering coëfficient

WIthin the supplier layer, there are a lot of triangles present as shown by the clustering coëfficients in Table 4.3. This reciprocal interdependence indicates information sharing which is a characteristic of efficient supply chains.

The remainder of this thesis will apply the introduced concepts to a real dataset. The results will be interpreted to analyse what the critical nodes are. The goal is to identify the critical points in such a way that management knows where the supply chain can be optimised to increase their competitive advantage.

Chapter 5

Real data & Methodology

5.1 Used data

Given the nature of this thesis, the aim was to get a large dataset with real world properties as previously discussed and described in the netchain structure [20]. This would enable a good analysis based on network methodologies such as those mentioned in Chapter 3. The received data did not entirely match this real world requirement. The implications of this are that the analysis will rely less on the specific methodologies as described, but more on general statistics as will be discussed later.

The data used for this thesis comes from a real dataset from a large car manufacturer containing all transportation processes of all vehicles produced in Europe. The specific data that will be analysed is that of one transportation route meaning that all vehicles within this sample all have the same origin factory, dealership destination and visited nodes (hence the limitations to using the specified methodologies). The used data is anonymous and contains nearly 20.000 entries from 1.600 unique vehicles, describing how the vehicle was transported from start to end through multiple locations. With 38 columns the dataset is very detailed but many entries do not contribute to this analysis and therefore only the following 10 columns will be used:

Column	Data
Vehicle No.	Integer: The identification number of each vehicle
Planned Arrival	Date: The date that the vehicle should have arrived
Arrival Day	Date: The date that the vehicle has arrived
Shipping Agent	String: The agent carrying out that transportation
Mean of Transport	String: The method of transportation used
Action	String: Indication of whether the vehicle is in holding or in transit
From	String: Current location where the vehicle departs from
From Timestamp	Datetime: Precise time when vehicle departs current location
To	String: Future location where the vehicle will arrive next
To Timestamp	Datetime: Precise time when vehicle arrives in the next location

5.2 Methodology and limitations

Now that the data is defined, a structure can be determined. The loaded data will be stored as a network where each node represents a location in the transportation process and the edges between them is a transaction from one location to another. Using the theories as described in Section 3.1 the created network can be summarised in Figure 5.1 and as follows:

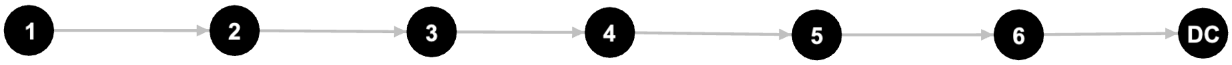


Figure 5.1: The generated network

Theory	Implementation
Directionality	Directed
Link type	Unweighed
Metadata	Nodes: Idle times
	Edges: Duration, Departure, Vehicle ID, Transport agent and method
Mode	Homogenic
Temporariness	Static

In Figure 5.1, 7 nodes are plotted with a edge between them. The vehicles go through 6 transportation locations before reaching the final dealership centre (DC). With this visualisation, it becomes even more evident that the described methodologies in Chapter 3 provide little insights. Locations 2 till 6 have the same indegree as outdegree as by definition as all vehicles pass all locations in the the same (defined) order. Location 1 and the DC have respectively an indegree or outdegree of 0 while the other degree measure is similar to that of the other nodes. Furthermore, the longest shortest path (diameter) is 6, only one component is present, no triangles exist and centrality measures are pointless. It is merely a sequential interdependence [20] which has limited methods of analysis.

Made clear that new ways of finding the critical nodes are needed, this analysis will rely on more standard statistical methods. More specific, four aspects will be researched:

1. Compare amount of vehicles that arrive early, on time or late
2. Find what transportation method takes longest and is most unreliable
3. Check whether transportation time is the most crucial process or that idle times (the time that vehicles are not in transit) makes up the longest part of the total delivery process
4. If idle times are significant, research where they are the highest

The methods that will be used to research these questions range from simple histograms and visual analysis to calculating median and median absolute deviation values.

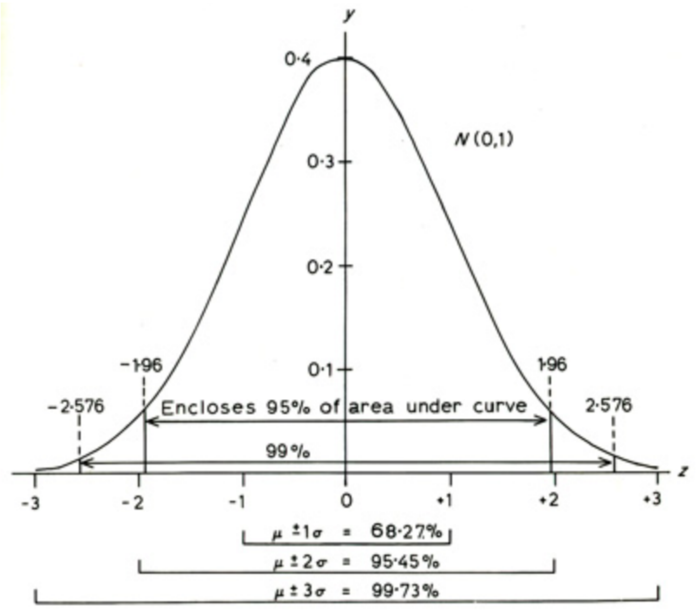


Figure 5.2: Gaussian distribution [45]

When for example calculating the mean value of all transportations carried out by shipping agent x, the mean value ($\bar{x}$) gives a good indication on how fast (or slow) they usually deliver. When the standard deviation (σ) is also calculated it can be said with 95% confidence [45] that this agent will deliver within $\bar{x} \pm 2 * \sigma$. Whenever an agent (or another researched aspect) has a high standard deviation, it can be said that the performance is not very well. The lower standard deviation, the higher the reliability of that process.

However, this method of analysis only works for normally distributed data as shown in Figure 5.2. Because the data worked with is more often skewed than not, the median will be calculated instead together with the median absolute deviation (MAD) [46] to provide insights in the performance. Even though a precise confidence cannot be provided, the MAD still allows a good analysis. The lower MAD value, the higher the reliability of that process. This type of exploratory search will be capable of identifying critical nodes.
Please note: duration this method of analysis, any mentioning of average in the results section (Chapter 6) implies the median value.

Regardless of the findings to the proposed questions, multiple expectations are present:

1. Data is skewed rather than normally distributed.

2. The transportation method by ship will take longest as it makes no sense to use ships for short distances.

3. The longer the transportation time is, the longer the idle times will be for that location. When transport takes long, you want to transport a larger bulk of goods. Ships are able to take a lot of vehicles at once so logically this method is only used for long distances (see assumption 2). To fill up a ship you need a larger batch of vehicles. The first one in that batch needs to wait relatively long before the last vehicle in the batch arrives. This drives up the average idle time and creates a high standard deviation.

4. The lower amount of vehicles per transportation batch, the shorter the idle times will be for that location. This reasoning follows from assumption 3.

If these assumptions hold, will be evaluated after the results to the experiment are shown.

Chapter 6

Analysis

In this chapter, the results to the questions as mentioned in Section 5.2 will be discussed in the same order. Each of the following subsections will provide an answer to one of the questions providing a clear guideline through the analysis. Calculations are made using a Python program which is the second script in the appendix.

6.1 General performance

The first step in finding the critical nodes is identifying how the overall transportation process performs. This was done by comparing the dates that the vehicles were expected to arrived with the actual arriving dates. Based on this, a vehicle transportation can be classified in one of the following four categories:

- Early: the vehicle arrived earlier than planned
- On time: the planned and actual date are identical
- Late: the vehicle arrived later than planned
- Failed: there is no information about when the vehicle arrived

With regard to the last category (failed) it is important to note that these transportations are discarded for further analysis as that data is incomplete. The findings are summarised in Figure 6.1

Status	Amount
Early	1370
On time	0
Late	1
Failed	288

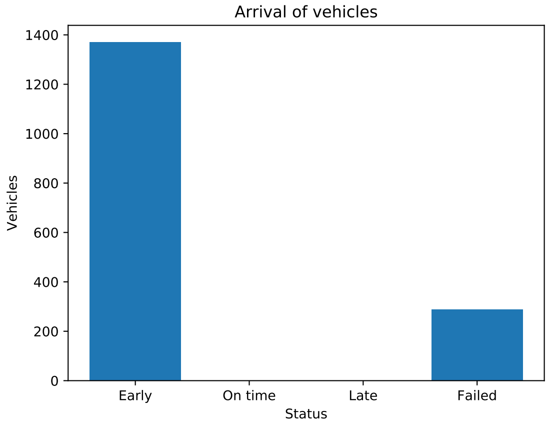


Figure 6.1: Arrival status of transportations

The findings here are remarkable. Practically all vehicles arrive early and the remainder (17%) has no available data concerning date the date of arrival. However, it is safe to say that the 288 vehicles did arrive at some point because they are not simply lost in the transportation process. Unfortunately it cannot be determined whether they arrived early, late or on time. Furthermore, it is quite an accomplishment that all other vehicles that did arrive, were early. The question this raises is: how early?

	Median	MAD	Min	Max
Days early	8	1	1	12

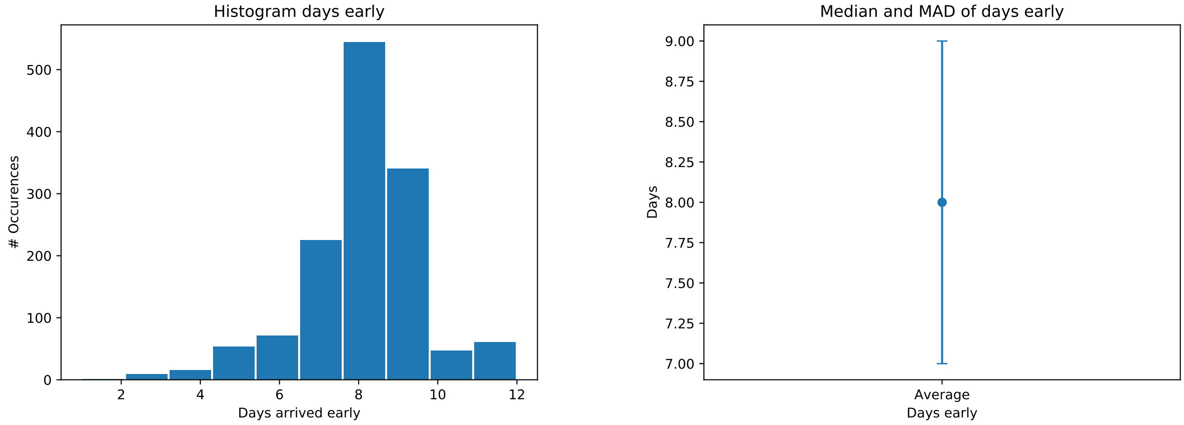


Figure 6.2: Amount of days early

As shown in Figure 6.2 the average vehicle (that arrives) will arrive 8 days early.

6.2 Transportation time

How long does the average transportation process take?

	Median	MAD	Min	Max
Transportation process (hours)	42	23.217	0	454.283

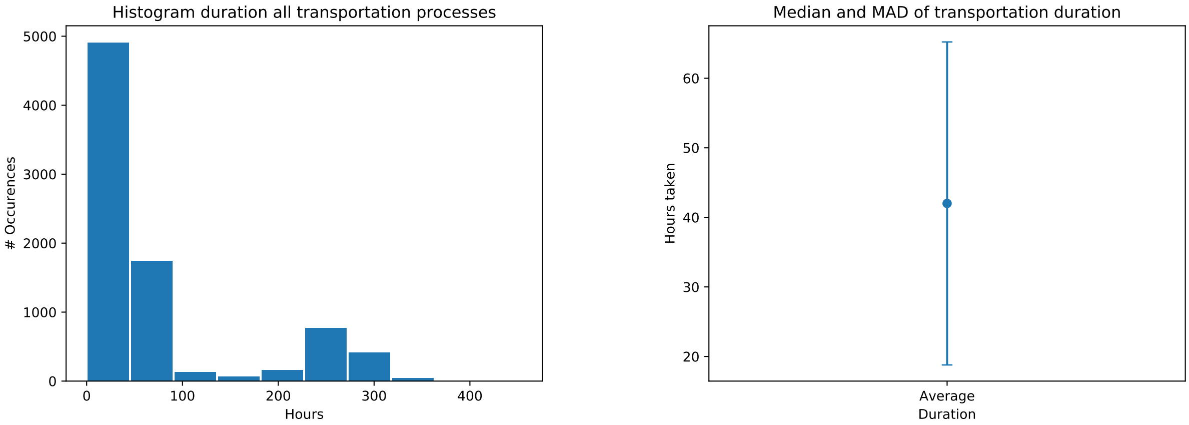


Figure 6.3: Duration of transportation process in hours

The average value of a transportation process (edge between two nodes) is 42 hours. The MAD is more than 50% of the median value which indicates a high variability in the duration of transportation processes. Further analysis of all the processes is needed in which the performance of each shipping agent, location (node) and transportation method will be filtered. There are specific relationships betweens locations, agents and methods as agents often only operate on one location with the same transportation method. The relationships between these filters for this dataset are summarized below in Table 6.1.

Method	Agent	Location	Agent	Location	Agent
Train	Agent 1	Location 1	Agent 1	Location 1	Train
Ship	Agent 2	Location 2	Agent 2	Location 2	Ship
Own Axis	Agent 3	Location 3	Agent 2 & 6	Location 3	Ship
Ship	Agent 4	Location 4	Agent 3	Location 4	Own Axis
Truck	Agent 5	Location 5	Agent 4	Location 5	Ship
Ship	Agent 6	Location 6	Agent 5	Location 6	Truck

Table 6.1: The relationships between agents, methods and locations

To make the visualisation of the results more appealing, results will be merged based on the relationships as just explained. Table 6.2 summarizes the relationships. Each row in the column indicates a unique result.

Agent	Method	Location
Agent 1	Train	Location 1
Agent 2		
Agent 3	Own Axis	Location 4
Agent 4		Location 5
Agent 5	Truck	Location 6
Agent 6		
	Ship	
		Location 2
		Location 3

Table 6.2: Summary of identical results

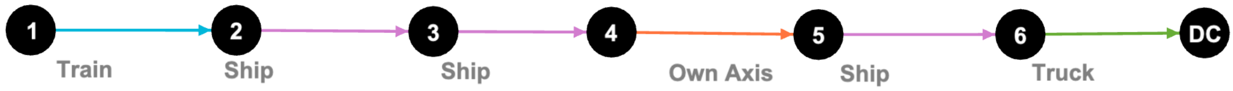
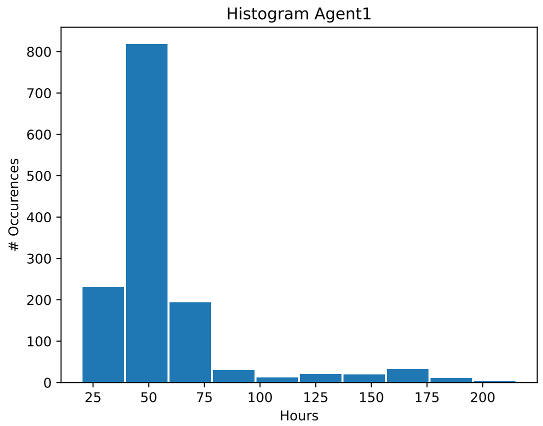
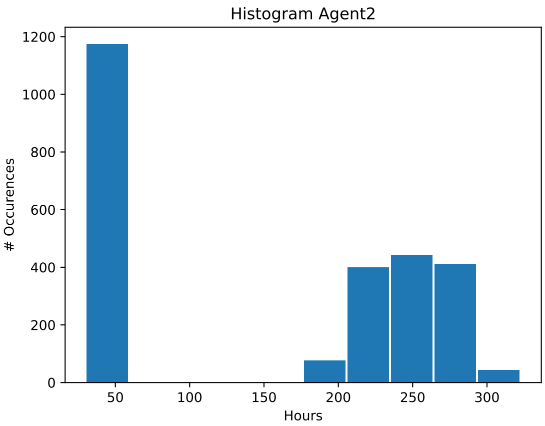


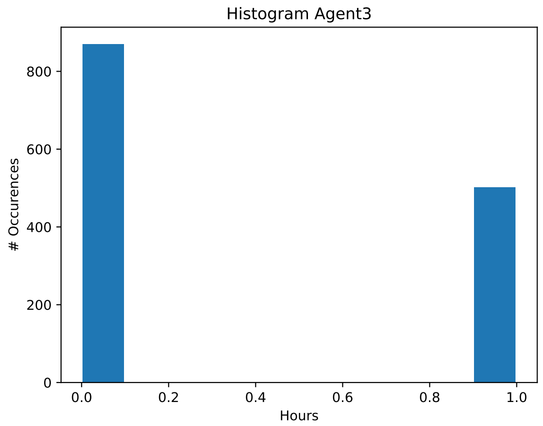
Figure 6.4: Graph with transportation methods

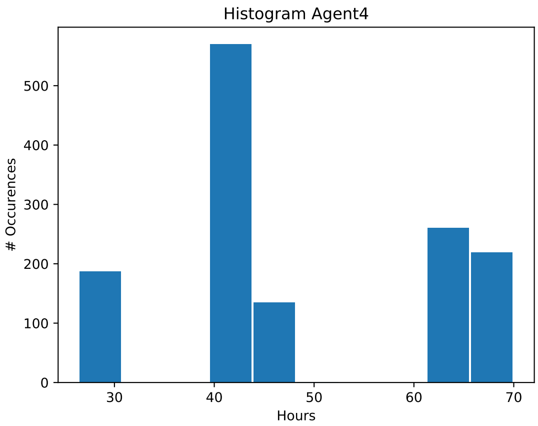
Based on these relationships (as visualised in Figure 6.4), some plots will not be shown twice or more in the analysis. The following subsections show the results per agent / method / location and interprets these.

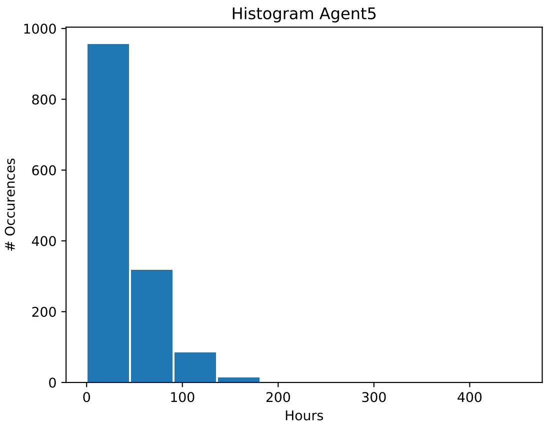
6.2.1 Agent performance

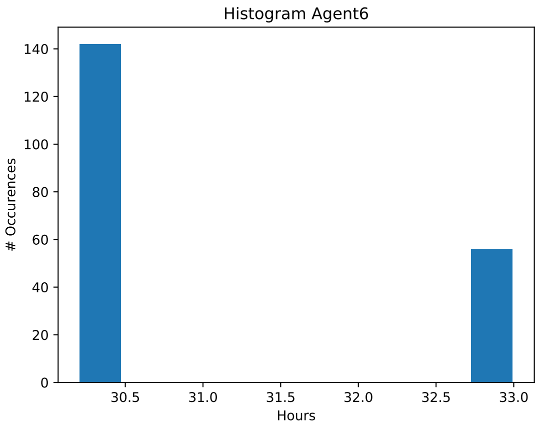












Which can be summarised as follows:

Agent	Median	MAD	Min	Max
Agent1	48.992	6.408	19.850	215.283
Agent2	208.550	74.033	30.050	322.717
Agent3	0	0	0	1
Agent4	42	5	26.417	70
Agent5	22.033	19.367	0	454.283
Agent6	30.200	0	30.200	33

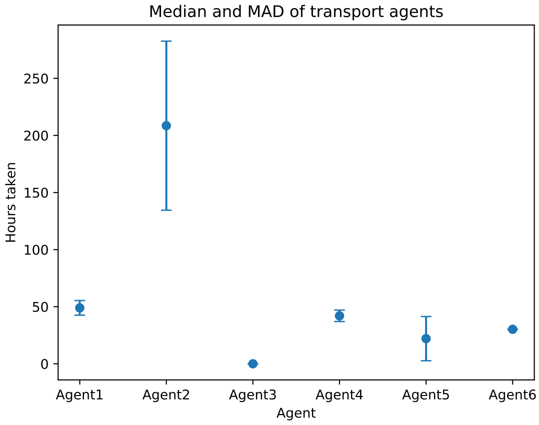
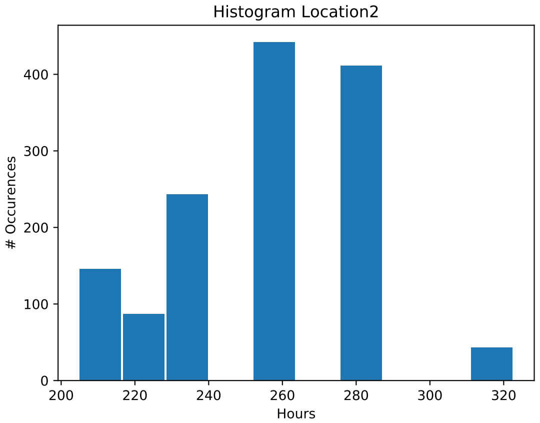


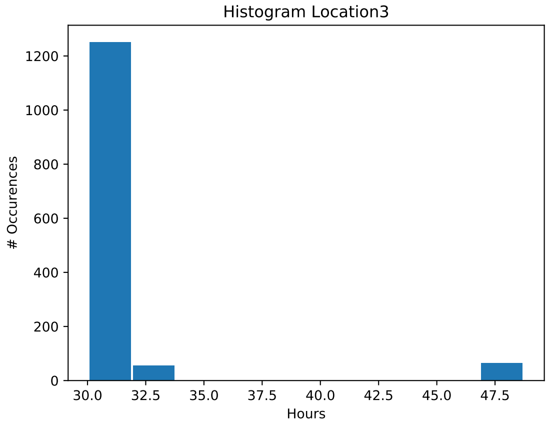
Figure 6.8: Performance of agents

The first assumption in Section 5.2 stated that the data is not normally distributed but rather skewed. This was the assumption that laid the foundation of using the median and MAD as an analysis method. After showing the histograms for agents, it can be concluded that this assumption was right. It even sometimes occurs that the data is not even skewed but rather consists of two (or more) peaks. However, the data is definitely never normally distributed as was expected. Further histograms shown throughout this section will be in line with this conclusion.

As seen in Figure 6.8, Agent2 has a very high MAD value. To find the cause for this, the different locations need to be analysed as this agent operates on Location2 & Location3 as can be seen in Table 6.1. After Agent2, the performance by Agent5 is least consistent.

6.2.2 Location performance





See Table 6.2 to find histograms of the other locations. Summarised:

Location	Median	MAD	Min	Max
Location1	48.992	6.408	19.850	215.2
Location2	255.933	24.883	204.733	322.7
Location3	30.150	0.017	30.050	48.73
Location4	0	0	0	1
Location5	42	5	26.417	70
Location6	22.033	19.367	0	454.2

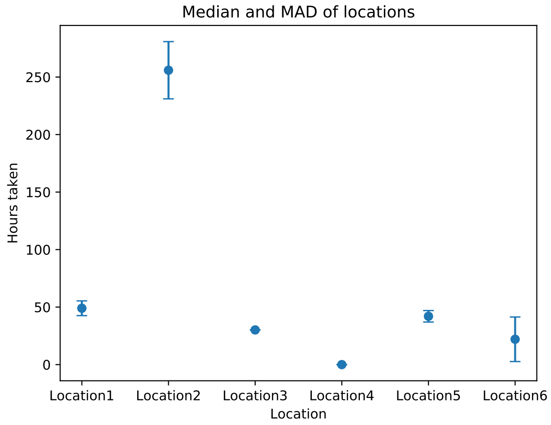


Figure 6.10: Performance of locations

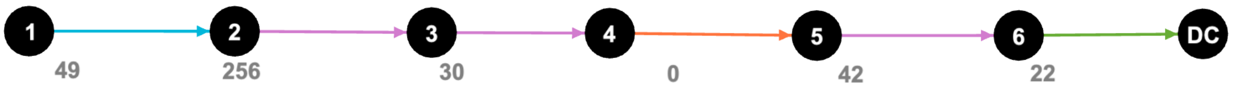
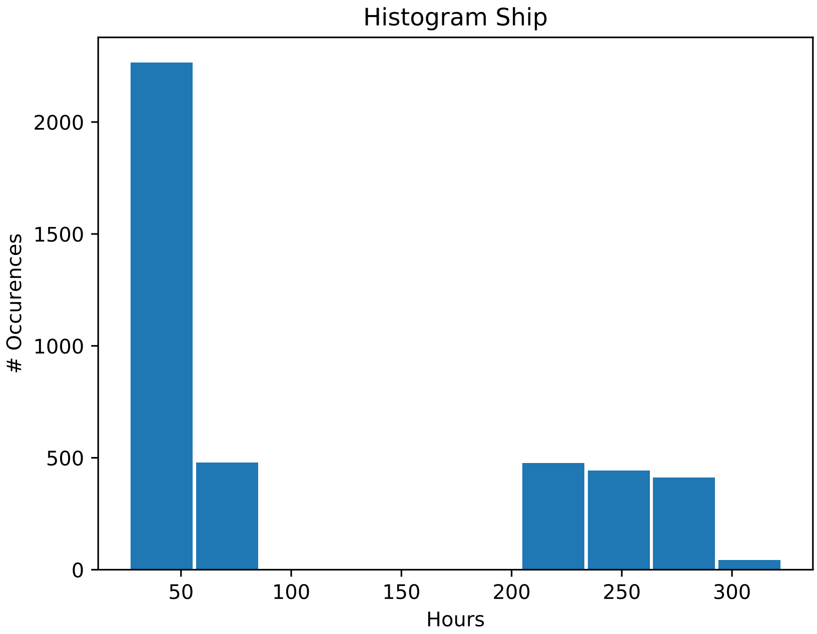


Figure 6.11: Graph with transportation methods and median transportation time

With these results, it can be understood why the performance of Agent2 is not consistent. Location 2 and 3 have very different durations. When looking at the MAD values, Location 2 and 6 are relatively high. This is in line with the conclusions drawn about the agent performance as Location2 is operated by Agent2 and Location6 by Agent5.

6.2.3 Transport method performance



See Table 6.2 to find histograms of the other transportation methods. Summarised:

Method	Median	MAD	Min	Max
Train	48.992	6.408	19.850	215.283
Ship	42	11.95	26.417	322.717
Own	0	0	0	1
Truck	22.033	19.367	0	454.283

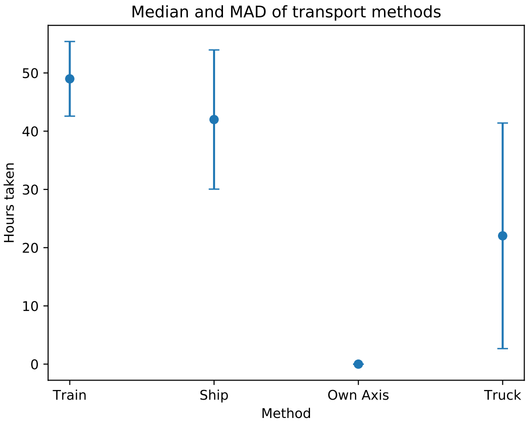
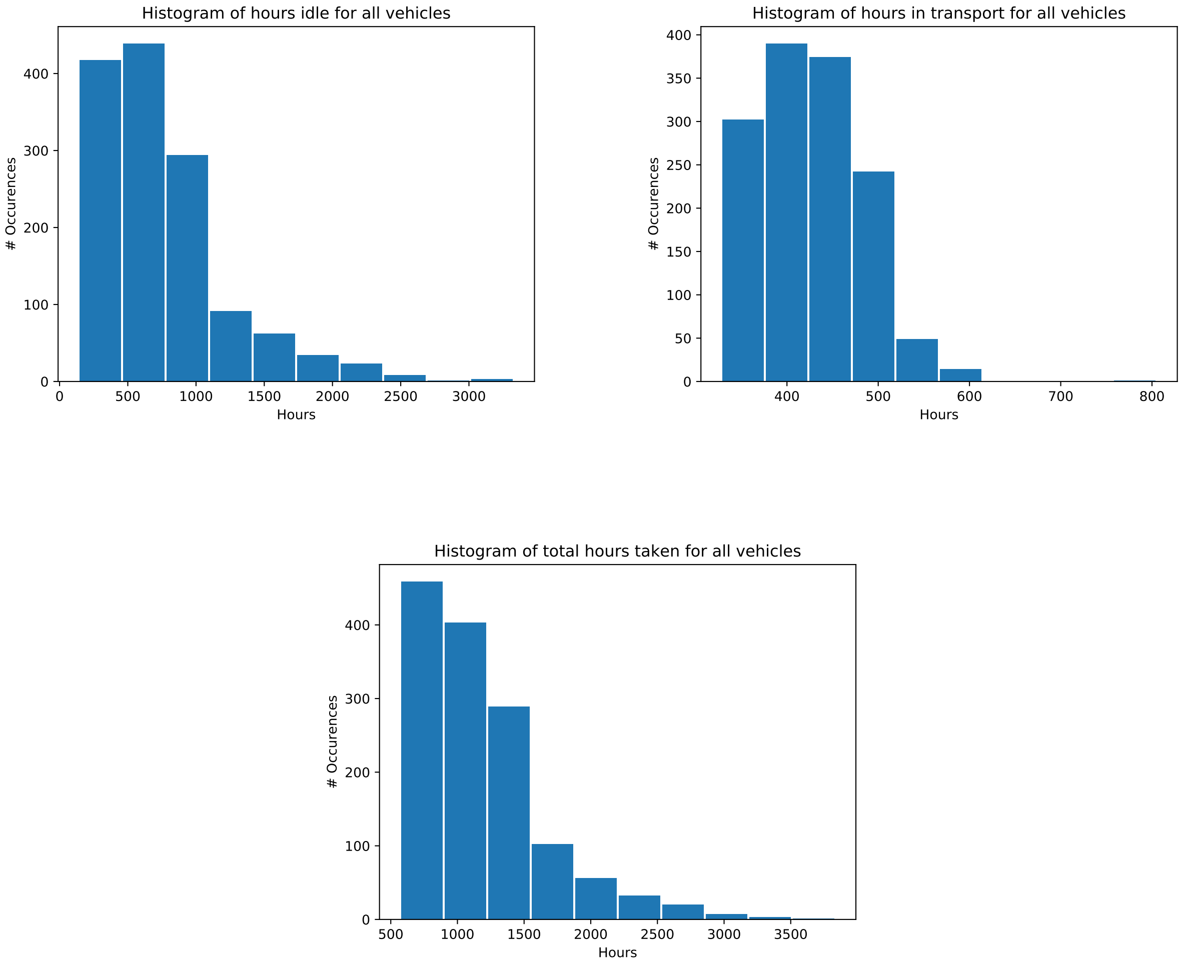


Figure 6.12: Performance of methods

The second assumption that shipping methods would only be used for longer distances is not true. In Figure 6.12, it can be seen that trains on average take longer. Locations 2, 3 & 5 are using ships (Table 6.1) and especially at Location 3 and 5 the shipping process goes relatively fast (Figure 6.10), this was not expected. The MAD of trucks is very high compared to other transportation methods.

6.3 Transportation vs idle time

Now that is has been analysed how the transportations processes perform it can be identified whether these processes actually take longest or that the idle time of vehicles take up the majority of the total transportation time.



Process	Median	MAD	Min	Max
Total	1065.283	240.717	569.65	3834.45
Transport	423.058	42.342	328.25	805.20
Idle	624.775	212.617	139.85	3329.475

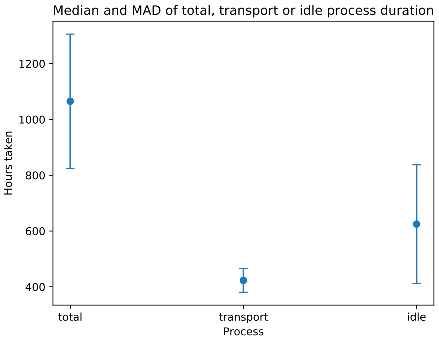
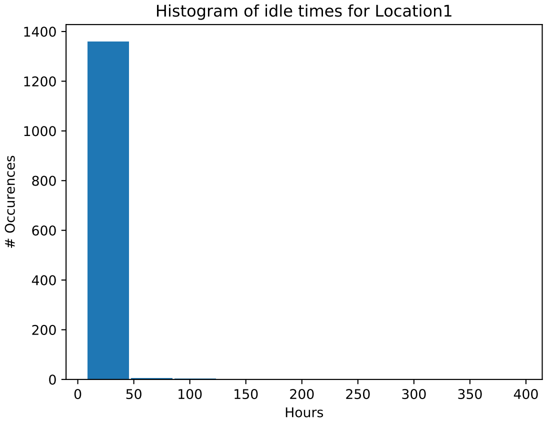
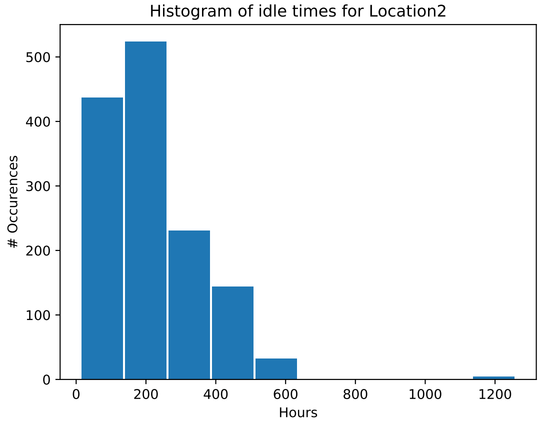


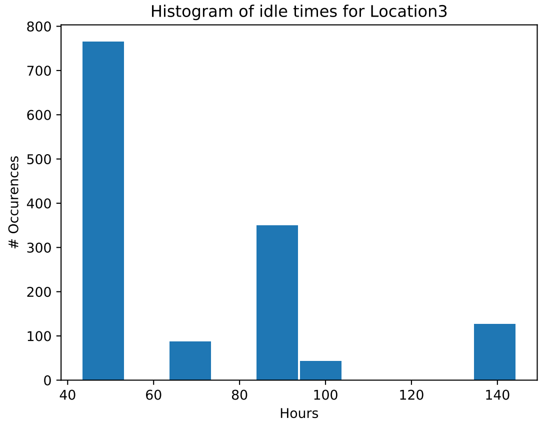
Figure 6.15: Transportation vs idle time

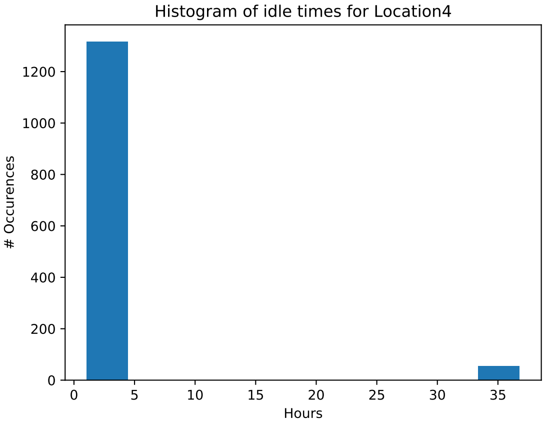
Based on Figure 6.15, it can concluded that the idle times are more significant than the transportations times. Knowing this fact, it can be researched where idle times are highest.

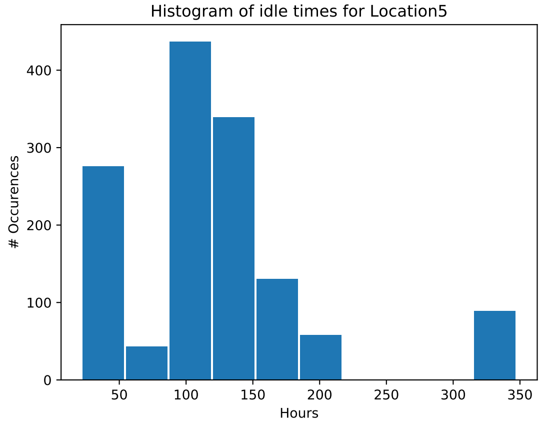
6.4 Idle time

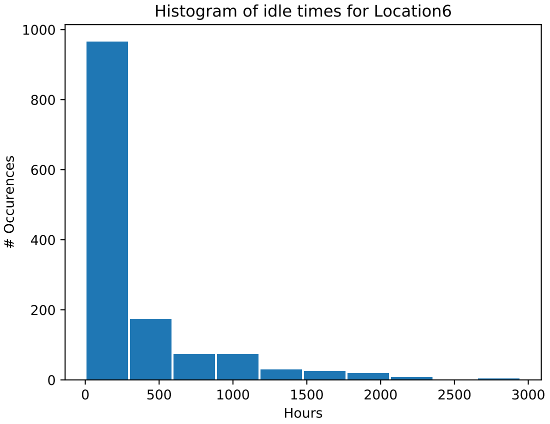








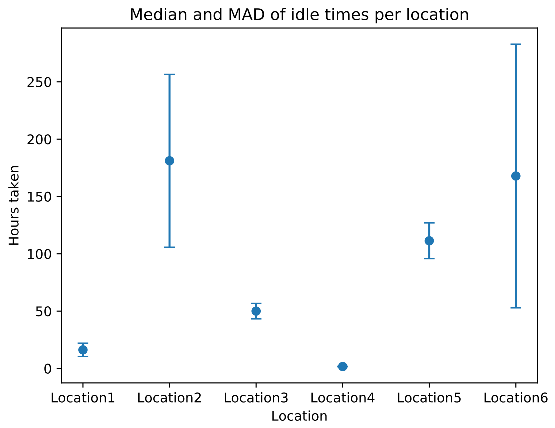


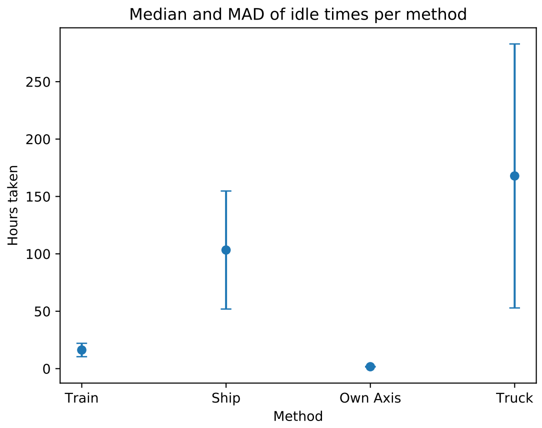


With these histograms and the relationships presented in Table 6.1, the following summary can be generated:

Location	Median	MAD	Min	Max
Location1	16.233	5.833	7.883	396.183
Location2	181.150	75.383	12.950	1259.317
Location3	50.017	6.767	43.250	144.483
Location4	1.683	0.217	0.967	36.883
Location5	111.367	15.583	21.783	347.550
Location6	167.867	115	2.600	2949

Method	Median	MAD	Min	Max
Train	16.233	5.833	7.883	396.183
Ship	103.333	51.417	12.950	1259.317
Own Axis	1.683	0.217	0.967	36.883
Truck	167.867	115	2.600	2949





In Figure 6.20, the idle times are shown graphically. Within a node, the identifier is shown followed by a “-” and the value of the idle time.

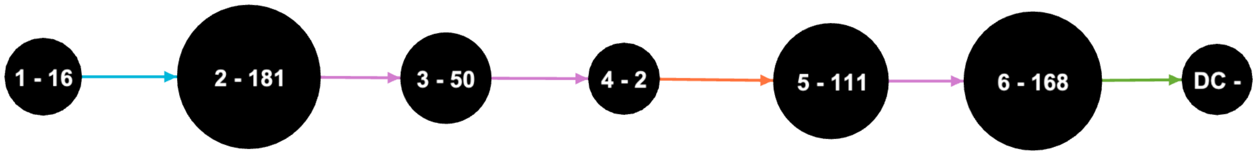


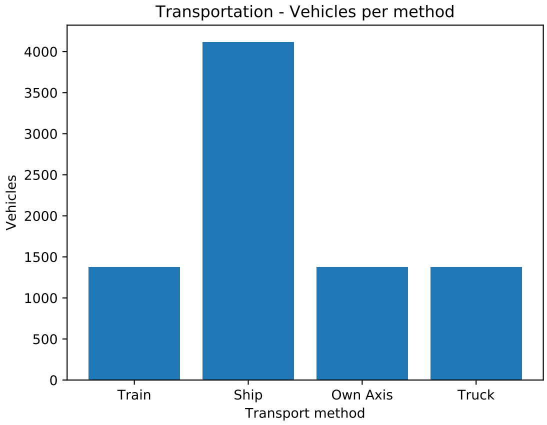
Figure 6.20: Graph with idle time (node size = idle time)

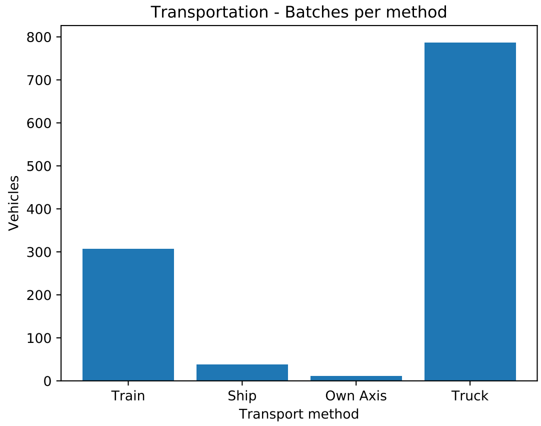
The average idle time for Location2 is highest with also a high MAD value. This can be caused due to the fact that ships transport large bulks of vehicles (see assumption in Section 5.2) and the ships need to wait until a large batch of cars is available to transport. That would imply that the first car in the batch has a long idle time before the final car for the batch arrives. The drives up the median and MAD value. However, this would not explain why the median value at Location5 is so high. At Location4 the vehicles arrive by ship and transported by own axis to the next ship which should be a smooth transition ensuring low idle times. This is not the case as the median value is relatively high and a significant (though not impressive) MAD value exists. Following the previous reasoning, it is logical that the median and MAD values of the own axis idle times are so low. This process is taking the goods of a ship and as all vehicles arrive at exactly the same moment, this can be done instantaneous.

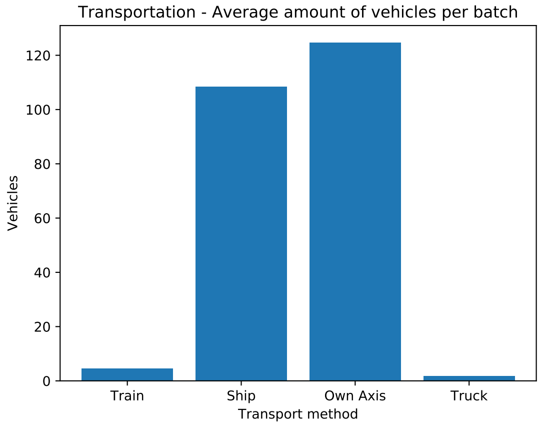
The outstanding value is Location6 / Truck as these have very long idle time and high MAD value. Previously it was concluded that trucks have a relatively low transportation time (Figure 6.12) and the fact that the idle time is high conflicts the third assumption in Section 5.2. The contrary is true for Location1 / Train as this method with relatively long transportation time has very short idle times.

When looking at the amount of batches per transportation method and the average amount of vehicle per batch, more insights can be gathered in possible causes.

Method	Amount	Batches	Avg/batch
Train	1372	307	4.5
Ship	4116	38	108.3
Own Axis	1372	11	124.7
Truck	1372	787	1.7







The fourth assumption was that the lower the amount of vehicles per batch, the shorter the idle times would be. This is now also contradicted as it has been established that trucks have a long idle time while the average amount of vehicles per batch is lowest of all methods. An interesting result is furthermore that the own axis transportation method has so few batches. This could be explained by the fact that the transportation time is also extremely low. Possibly this method is just an internal administrative duty that transfers all vehicles from one ship to another. That ships carry most vehicles per batch is as expected. Figure 6.23 shows the total summary. As before, the node shows the identifier together with the value of the idle time. In this figure, the transportation time between a node and the next one is displayed in red at the origin node.

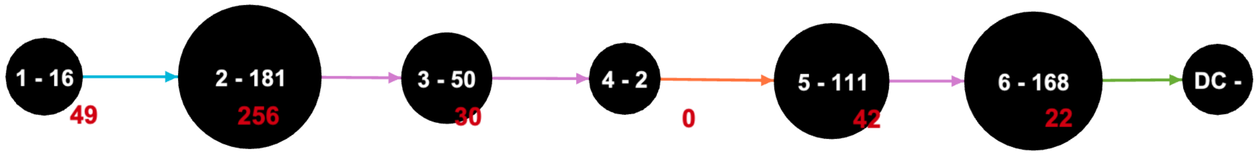


Figure 6.23: Graphical summary of the results (node size = idle time)

From this analysis it can be concluded that idle times are the most important factor. At Location 6 the idle time is very high and inconsistent. After unloading the ships, the cars are apparently held in the harbour as inventory while ideally they would immediately be transported further to be sold. Further discussion of the results is in the next chapter.

Chapter 7

Recommendations and further research

Based on the given results the following recommendations can be given to the car manufacturer concerning their transportation process on this route:

- With 17% of the data not having an entered timestamp of arrival this should be improved. Even if the vehicles arrived late, it is crucial to log this to measure performance.
- The shipping method is used three times in the same transportation route. Two of them have a very short duration, alternatives to shipping should be considered.
- Idle times play a crucial role in improving overall performance therefore the specific cause for idle times should be investigated. Is it due to imperfect transitions between previous transportation method, order batching or external reasons?
  - If transitions are imperfect, these should be improved. By aligning the transportation processes in such a way that for example when a ship arrives, the vehicles are directly loaded on the other so the vehicles have reduced idle times.
  - When order batching takes place, higher idle times are a logical consequence. This would however not explain why Location5 and Location6 have high idle times.
  - External reasons can be for example border customs. If all vehicles have to be checked by customs it can be explained why idle times are high. Being an internationally well known organisation might make it possible to reduce idle

times at borders by cooperating with governments.

- Transportations starting in location 2 and 6 can be performing better based on the idle times. Further evaluate agents carrying out these transportations to improve their operations.
- High stock levels are indicated by the high idle times. Inventory is the same as waste because the vehicles will not be sold when not on the move. Reducing stock levels can improve operational performance. Especially the idle time at Location6 is worrying.

For further research this method of analysis can be used for other transportation routes as well. At this point, the analysis is only limited to providing insights in the data but manual interpretations are needed. Results now indicate that idle times are very high at a certain node but this might be a totally logic decision from a managerial perspective. With the help of extra data that indicate the costs of certain decision it can be made possible to make automated optimisations. Now a conclusion is that ships are used for short distances where it was expected that other transportation methods would be used. With financial information it can be shown whether this is a sensible cost-based decision. Also, doing a time series analysis on this network might already indicate that idle times are decreasing and some improvements were already made.

Best results can however be generated if the data contains all transportation processes with possibly even the transportation of materials to the factory before assembly. This will be a large hub and spoke network as introduced in Chapter 4 enabling the analysis methods as shown in Chapter 3. As mentioned before, this data consists of merely sequential interdependence. The entire output of one node is the input of the next. This data reflects one specific transportation route for a specific factory, this factory does however have a lot of suppliers itself.

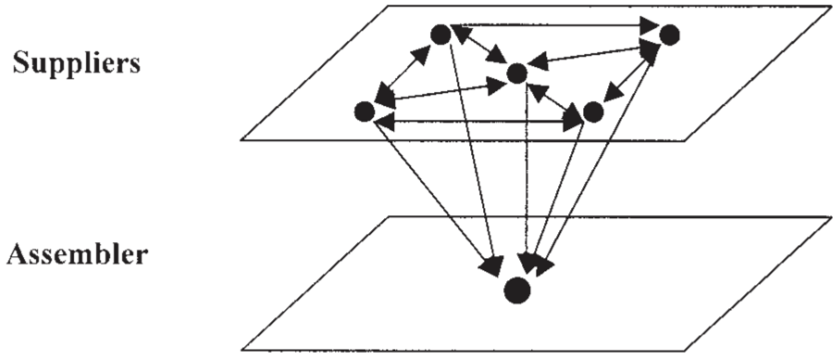


Figure 7.1: Buyer-supplier relationship [20]

The relationship between the factory and all the suppliers is depicted in Figure 7.1. Building cars and selling them is a team effort with multiple parties involved form all over the world. The data as received only contained data from one specific transportation route. It would therefore be very interesting to add all other transportation routes together with supply chain data from the suppliers to the factory.

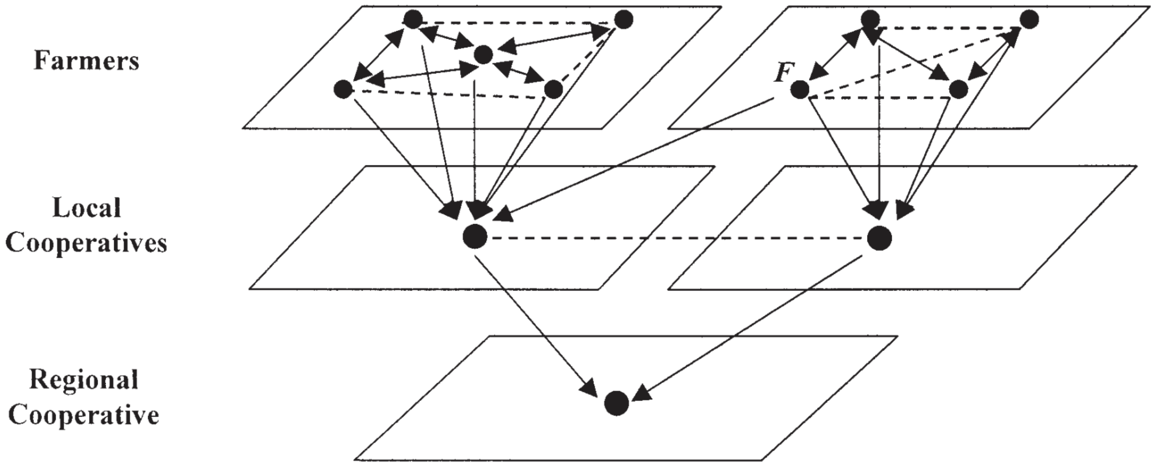


Figure 7.2: A macrohierarchy [20]

Like the macrohierarchy as shown in Figure 7.2, the production and distribution of cars is not very different. Better said, the structure will be much like a hourglass similar to the example in Section 3.3 (Figure 3.9). Parts are manufactured all over the world, brought together for assembly of smaller parts and eventual production of the car in central factories. After finishing the production of cars, the products are transported to distribution centre's all over the world to be sold at dealerships. The former process represents the top part of the hourglass (narrowing the amount of involved parties to the central factories) and the latter represents the bottom half (using more parties to sell over the world). This in combination with the netchain structure [20] allows a very detailed analysis. How such a structure can be analysed has been shown in Section 4.2.

The main recommendation for further research is therefore the use of a more complete (non sequential) dataset that follows the netchain structure. This thesis delivered a framework that can be used in such a way.

Chapter 8

Conclusions

It can be concluded that this thesis has a laid a solid theoretical basis concerning supply chains. A network based analysis method has been introduced with metrics capable of identifying critical nodes within such a supply chain and solid instructions are provided for future work. If the data represents a real world hub and spoke network, the introduced framework can be tested. In addition to this, the used median and MDA analysis was capable of generating interesting results:

- 17% of the data did not have a recorded arrival date
- Average transportation time is not constant, a high MAD value exists
- Shipping methods are used for small distances which was not expected
- Trains are a method of transportation taking relatively long which also was not expected
- Idle times are more important than transportation times
- It was expected that transport methods with long duration and large amount of vehicles per batch would have long idle times, the opposite was remarkably true

All these result were found with using simple statistical analysis, imagine the possibilities when the desired network analysis could have been conducted. In this thesis, sample calculations of network metrics were given together with a critical node analysis based on another example using these metrics. The examples aid in the explanation of the framework that eventually can be used. With the written literature review and extensive examples, the network methodologies have been shown as much as possible. The used statistical analysis has been done from a network approach as all data was stored in a network structure using Python. Calculating the median and MAD values is a very good complementary asset to the network metrics and is a good addition to future research.

The results as found in Chapter 6, were already summarised in Figure 6.23. In this figure, the numbers presented were crucial in understanding the summary. The node size was equal to the idle time but a more friendly summary can also be provided as shown in Figure 8.1.

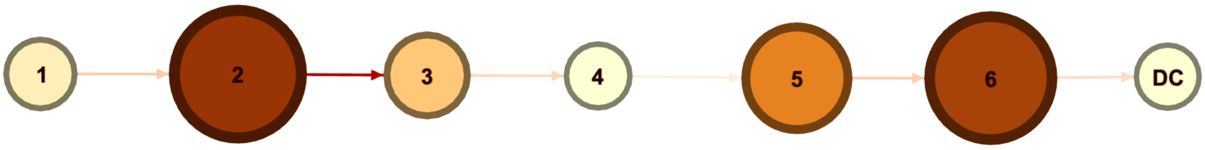


Figure 8.1: Summary of the results

Here the node size is still equal to the idle time but instead of showing a lot of numbers, colours are used to indicate the intensity of the underlying values. The warmer a colour is, the higher the value. Based on this, it can more easily been seen that node 2 and 6 have a high idle time and the path between node 2 and 3 takes relatively long compared to the rest of the transportation paths. The main difference between Figure 6.23 and Figure 8.1 is that the first can be considered an in-depth summary whereas the second is the management summary. Management would not be interested in the exact details at the first instance, rather they would just want to know where the issue is. After identifying where the problem occurs in the supply chain, the exact details can be found in another summary. If it is then still concluded that there is an issue, the individual graphs can be consulted to find the exact cause. With this type of visual reporting, fast conclusions can be drawn.

In Section 4.1, an example has already been provided related to the airline industry. The foundation laid in this thesis is not restricted to supply chains for a specific organisation but can also be used to analyse for example aviation networks. Just as in the summary above, an entire aviation network can be analysed and displayed in similar fashion. This enables an easy comparison of small airfields and large hubs. Using visual reporting, a very easy method is provided for management to interpret the performance of any network like structure. The results as discussed in this thesis were filtered on transportation method / agent / location. For an aviation network, custom filtering can also be used to compare the performance of different airliners. Nodes would represent airports and edges an existing flight between them, a homogenic network. Directionality can be used to filter the direction of the flight. The link type would be unweighted as every single flight would have their own edge. Metadata would contain information about departure time, flight duration, arrival time and airliner. With this information stored in the metadata, a temporal analysis can also be conducted so that the performance change over time can be researched. Using the visual techniques as discussed, the evolution of the aviation network can be beautifully shown while also providing a visual management summary of the relative performance. With the individual performance graphs for each filter, detailed results can be provided just as in the results from Chapter 6. By giving this introduction to aviation networks it has been shown how the theory from this thesis can be used for more than a traditional (organisational) supply chain analysis.

Concluding: well defined and actionable results were found regardless of the limitations. By identifying the critical nodes, management can verify whether these are deliberate decisions or that improvements can be made. By reducing stock and creating a better aligned transportation network with less idle times, a competitive advantage is achieved due to a higher level of integration as proposed by Steven's four stage model. A solid introduction has been provided showing how larger network structured supply chains can be analysed. The relevant theory is however not limited to just supply chains. In fact any network like structure can be analysed from an organisational perspective using the provided methodologies.

Bibliography

[1] R. K. Oliver and M. D. Webber, *Supply-chain management: logistics catches up with strategy*. M. L. Christopher (Ed.), Logistics: The strategic issues (pp. 63–75). London: Chapman & Hall., 1982.

[2] L. P. Bucklin, *A theory of distribution channel structure*. Berkeley, 1966.

[3] J. W. Forrester, *Principles of systems / by Jay W. Forrester*. Wright-Allen Press Cambridge, Mass, 1969.

[4] J. B. Hoolihan, *International Supply Chain Management*. International Journal of Physical Distribution and Materials Management, 1985.

[5] M. Cooper, "International supply chain management: implications for the bottom line," *Proceesings of the Society of Logistics Engineering*, 1993.

[6] M. E. Porter, "Competitive advantage: creating and sustaining superior performance," *The Free press*, 1985.

[7] S. Li, B. Ragu-Nathan, T. Ragu-Nathan, and S. S. Rao, "The impact of supply chain management practices on competitive advantage and organizational performance," *Omega*, vol. 34, no. 2, pp. 107–124, 2006.

[8] M. Christopher, *Logistics and Supply Chain Management*. Financial Times Series, Pearson Education, 2013.

[9] G. C. Stevens, "Integrating the supply chain," *International Journal of Physical Distribution & Materials Management*, vol. 19, no. 8, pp. 3–8, 1989.

[10] O. Brook, *Council of Logistics Management. Annual Conference, 1986, Anaheim, Calif. Proceedings*. Council of Logistics Management, 1986.

[11] A. Jomini, "Tableau analytique des principales combinaisons de la guerre, et de leurs rapports avec la politique des etats, pour servir d'introduction au traité des grandes opérations militaires," 1830.

[12] J. T. Mentzer, W. DeWitt, J. S. Keebler, S. Min, N. W. Nix, C. D. Smith, and Z. G. Zacharia, "Defining supply chain management," *Journal of Business Logistics*, vol. 22, no. 2, pp. 1–25, 2001.

[13] D. Lambert, "The international center for competitive excellence," 1994.

[14] D. Simchi-Levi, P. Kaminsky, E. Simchi-Levi, and R. Shankar, *Designing and managing the supply chain: concepts, strategies and case studies*. Tata McGraw-Hill Education, 2008.

[15] J. W. Forrester, "Industrial dynamics," *M.I.T. Press*, 1961.

[16] H. L. Lee, V. Padmanabhan, and S. Whang, "The bullwhip effect in supply chains," *Sloan management review*, vol. 38, no. 3, pp. 93–103, 1997.

[17] S. Chopra and M. S. Sodhi, "Managing risk to avoid supply-chain breakdown: by understanding the variety and interconnectedness of supply-chain risks, managers can tailor balanced, effective risk-reduction strategies for their companies," *MIT Sloan management review*, vol. 46, no. 1, pp. 53–62, 2004.

[18] B. M. Beamon, "Supply chain design and analysis:: Models and methods," *International journal of production economics*, vol. 55, no. 3, pp. 281–294, 1998.

[19] J. Wikner, D. R. Towill, and M. Naim, "Smoothing supply chain dynamics," *International Journal of Production Economics*, vol. 22, no. 3, pp. 231–248, 1991.

[20] S. Lazzarini, F. Chaddad, and M. Cook, "Integrating supply chain and network analyses: the study of netchains," *Journal on chain and network science*, vol. 1, no. 1, pp. 7–22, 2001.

[21] L. Euler, "Solutio problematis ad geometriam situs pertinentis," *Commentarii academiae scientiarum Petropolitanae*, pp. 128–140, 1741.

[22] F. Takes, "Social network analysis for computer scientists," 2018.

[23] J. A. Bondy, U. S. R. Murty, *et al.*, *Graph theory with applications*, vol. 290. Citeseer, 1976.

[24] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 8, pp. 269–271, 1959.

[25] R. W. Floyd, "Algorithm 97: Shortest path," *Commun. ACM*, vol. 5, p. 345, June 1962.

[26] D. Easley, J. Kleinberg, *et al.*, *Networks, crowds, and markets*, vol. 8. Cambridge University Press Cambridge, 2010.

[27] S. Janson, D. E. Knuth, T. Łuczak, and B. Pittel, "The birth of the giant component," *Random Structures & Algorithms*, vol. 4, no. 3, pp. 233–358, 1993.

[28] T. Schank and D. Wagner, *Approximating clustering-coefficient and transitivity*. Universität Karlsruhe, Fakultät für Informatik, 2004.

[29] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[30] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2566–2572, 2002.

[31] S. P. Borgatti, "Centrality and network flow," *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.

[32] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215 – 239, 1978.

[33] A. Bavelas, "Communication patterns in task-oriented groups," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.

[34] A. Beveridge and J. Shan, "Network of thrones," *Math Horizons*, vol. 23, no. 4, pp. 18–22, 2016.

[35] A. Cox, J. Sanderson, and G. Watson, "Supply chains and power regimes: toward an analytic framework for managing extended networks of buyer and supplier relationships," *Journal of supply chain management*, vol. 37, no. 1, pp. 28–35, 2001.

[36] K. J. Mizgier, M. P. Jüttner, and S. M. Wagner, "Bottleneck identification in supply chain networks," *International Journal of Production Research*, vol. 51, no. 5, pp. 1477–1490, 2013.

[37] S. P. Borgatti and X. Li, "On social network analysis in a supply chain context," *Journal of Supply Chain Management*, vol. 45, no. 2, pp. 5–22, 2009.

[38] C. W. Craighead, J. Blackhurst, M. J. Rungtusanatham, and R. B. Handfield, "The severity of supply chain disruptions: design characteristics and mitigation capabilities," *Decision Sciences*, vol. 38, no. 1, pp. 131–156, 2007.

[39] M. A. Bellamy and R. C. Basole, "Network analysis of supply chain systems: A systematic review and future research," *Systems Engineering*, vol. 16, no. 2, pp. 235–249, 2013.

[40] E. J. Hearnshaw and M. M. Wilson, "A complex network approach to supply chain network theory," *International Journal of Operations & Production Management*, vol. 33, no. 4, pp. 442–469, 2013.

[41] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[42] A. Dubois and P. Fredriksson, "Cooperating and competing in supply networks: Making sense of a triadic sourcing strategy," *Journal of Purchasing and Supply Management*, vol. 14, no. 3, pp. 170–179, 2008.

[43] D. A. Lines, "Route map for u.s., canada," 10 2015.

[44] D. P. Cheung and M. H. Gunes, "A complex network analysis of the united states air transportation," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 699–701, IEEE Computer Society, 2012.

[45] R. Till, *Statistical methods for the earth scientist: an introduction*. Macmillan International Higher Education, 1974.

[46] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical association*, vol. 88, no. 424, pp. 1273–1283, 1993.

Script for sample calculations

```
import networkx as nx
import numpy as np
import matplotlib.pyplot as plt
from texttable import Texttable as tt
from decimal import Decimal


def getDistribution(data):
    returnDict = {}
    for i in data:
        if i in returnDict.keys():
            returnDict[i] += 1
        else:
            returnDict[i] = 1

    return returnDict


def plotHistogram(data, label, title):
    plt.title(title)
    plt.xlabel(label)
    plt.ylabel("#Occurances")
    xval = list(data.keys())
    yval = list(data.values())
    plt.xticks(np.arange(min(xval), max(xval) + 1, 1.0))
    plt.bar(xval, yval)
    plt.savefig(label + ".pdf")
    plt.close()
    print("Computed␣" + title)


def sortDic(d):
    return sorted(d.items(), key=lambda x: -x[1])


def distanceDistribution(graph):
    distanceDistribution = {}
    dd = dict(nx.shortest_path_length(graph))
    for source in dd:
        for target in dd[source]:
            pathLength = dd[source][target]
            if pathLength in distanceDistribution.keys():
                distanceDistribution[pathLength] += 1
            else:
                distanceDistribution[pathLength] = 1

    plotHistogram(distanceDistribution, "Distance", "Distance distribution")


def createTable(header, data, title):
    counter = 1
    print(title)
    table = tt()
    table.set_cols_dtype(['a', 't', 't'])
    table.add_row(header)
    for row in data:
        value = list(row)
        if value[1] < 1:
            value[1] = '%.2E' % Decimal(value[1])
        value.insert(0, counter)
        table.add_row(value)
        counter += 1
    print(table.draw())


G = nx.DiGraph()

edges = [("S1", "A1"), ("S2", "A1"), ("S6", "M1"), ("A1", "M1"), ("S4", "A2"),
         ("S3", "A2"), ("S5", "A2"), ("M1", "F"), ("M2", "F"), ("S8", "F"),
         ("F", "D1"), ("F", "D3"), ("F", "D2"), ("D1", "C1"), ("D1", "C3"),
         ("D1", "C2"), ("D3", "C6"), ("D2", "C4"), ("D2", "C5"), ("S7", "F"),
         ("A2", "M2"), ("S9", "A3"), ("S10", "A3"), ("A3", "M2")]

edges2 = [("S1", "S2"), ("S2", "S1"), ("S2", "S5"), ("S5", "S3"), ("S3", "S5"),
          ("S4", "S5"), ("S1", "M1"), ("S5", "M1"), ("S4", "M4"), ("S5", "M3"),
          ("S2", "M3"), ("S3", "M5"), ("S6", "M3"), ("S6", "M2"), ("M1", "C1"),
          ("M1", "D1"), ("M2", "D1"), ("M3", "D2"), ("M4", "D2"), ("M5", "D2"),
          ("M5", "D3"), ("M5", "C6"), ("D1", "C1"), ("D2", "C2"), ("D2", "C3"),
          ("D3", "C3"), ("D3", "C4"), ("D3", "C5"), ("S4", "M2"), ("S1", "S6"),
          ("S6", "S2"), ("S4", "S3"), ("S3", "S4")]

G.add_edges_from(edges2)

n = G.number_of_nodes()
m = G.number_of_edges()

print("Number␣of␣nodes␣are:␣" + str(n))
print("Number␣of␣edges␣are:␣" + str(m))

density = nx.density(G)
print("The␣density␣is:␣" + str(density))

indegree = dict(G.in_degree())
indegreeDistribution = getDistribution(list(indegree.values()))
plotHistogram(indegreeDistribution, "Indegree", "Histogram␣of␣indegree")

outdegree = dict(G.out_degree())
outdegreeDistribution = getDistribution(list(outdegree.values()))
plotHistogram(outdegreeDistribution, "Outdegree", "Histogram␣of␣outdegree")

distanceDistribution(G)

top20in = sortDic(indegree)[:5]
createTable(["Ranking", "Node", "Indegree"], top20in,
            "Top␣5␣nodes␣based␣on␣indegree")
top20out = sortDic(outdegree)[:5]
createTable(["Ranking", "Node", "Outdegree"], top20out,
            "Top␣5␣nodes␣based␣on␣outdegree")
top20close = sortDic(dict(nx.closeness_centrality(G)))[:5]
createTable(["Ranking", "Node", "Closeness"], top20close,
            "Top␣5␣nodes␣based␣on␣closeness␣centrality")

top20between = sortDic(dict(nx.betweenness_centrality(G)))[:5]
createTable(["Ranking", "Node", "Betweenness"], top20between,
            "Top␣5␣nodes␣based␣on␣betweenness␣centrality")

print(sorted(nx.clustering(G).items(), key=lambda kv: kv[1]))
```

Script for real data analysis

```
import numpy as np
import networkx as nx
import csv
import matplotlib.pyplot as plt
import datetime
from texttable import Texttable as tt


def load(filename):
    sc = {}
    dataset = open(filename + ".csv", "r", encoding="utf-8")
    lines = dataset.readlines()[2:]  # skip first two rows
    discard = []

    for line in lines:
        entry = line.split(";")
        code = entry[0]
        if not entry[5] or not entry[7]:
            discard.append(code)
            # It might happen that start or end times are not well defined
            # Then we have to remove that vehicle entry
        else:
            start = datetime.datetime.strptime(entry[5], '%d/%m/%Y␣%H:%M')
            end = datetime.datetime.strptime(entry[7], '%d/%m/%Y␣%H:%M')
            if entry[3] and entry[35]:
                planned = datetime.datetime.strptime(entry[3], '%d/%m/%Y')
                actual = datetime.datetime.strptime(entry[35], '%d/%m/%Y')
                if actual < planned:
                    key = "Early"
                    if code not in arrivals["Diff"]:
                        difference = planned - actual
                        arrivals["Diff"][code] = difference.days
                elif actual == planned:
                    key = "Precise"
                else:
                    key = "Late"
                if code not in arrivals[key]:
                    arrivals[key].append(code)
        delta = end - start
        duration = delta.total_seconds() / 3600  # duration in hours of a process
        node1 = entry[4]
        node2 = entry[6]
        if code not in sc:
            sc[code] = {"paths": [], "durations": [], "dates": [],
                        "agents": [], "methods": [], "standing": {}}
            node1 = code

        if entry[8] == "Transporting":
            sc[code]["paths"].append([node1, node2])
            sc[code]["durations"].append(duration)
            sc[code]["dates"].append(start)
            agent = entry[36]
            method = entry[37].rstrip()  # strip the newline from string
            if agent not in possibleAgents:
                possibleAgents.append(agent)

            if method not in possibleMethods:
                possibleMethods.append(method)

            sc[code]["agents"].append(agent)
            sc[code]["methods"].append(method)
        else:
            sc[code]["standing"][node2] = duration

dataset.close()

discard = list(dict.fromkeys(discard))
if len(discard):
    for key in discard:
        del sc[key]

plot_bar(["Early", "On␣time", "Late", "Failed"],
         [len(arrivals["Early"]), len(arrivals["Precise"]),
          len(arrivals["Late"]), len(discard)],
         "Arrival␣of␣vehicles", "Status")

table(["Status", "Amount"],
      [["Early", len(arrivals["Early"])],
       ["On␣time", len(arrivals["Precise"])],
           ["Late", len(arrivals["Late"])],
           ["Failed", len(discard)]],
          "Arrival␣status␣of␣vehicles")

    histogram(list(arrivals["Diff"].values()), "Days␣arrived␣early",
              "Histogram␣days␣early")
    data = {"Average": list(arrivals["Diff"].values())}
    plot_median_mad(["Average"], data,
                    "Median␣and␣MAD␣of␣days␣early", "Days␣early", ylabel="Days")

    return sc


def process(sc):
    graph = nx.MultiDiGraph()

    with open(filename + "-trimmed.csv", 'w') as file:
        writer = csv.writer(file)
        writer.writerows([["vehicle", "source", "target", "duration"]])
        for code in sc:
            for i in range(len(sc[code]["paths"])):
                duration = sc[code]["durations"][i]
                node1 = sc[code]["paths"][i][0]
                node2 = sc[code]["paths"][i][1]
                row = [code, node1, node2, duration]
                writer.writerows([row])
                graph.add_edge(node1, node2, duration=duration,
                               date=sc[code]["dates"][i], vehicle=code,
                               agent=sc[code]["agents"][i],
                               method=sc[code]["methods"][i])

            for node in sc[code]["standing"]:
                if "durations" not in graph.node[node]:
                    graph.node[node]["durations"] = []
                graph.node[node]["durations"].append(sc[code]["standing"][node])

    file.close()
    return graph


def general_stats(sc):
    stats = {"total": [], "transport": [], "idle": []}

    for key in sc:
        transport = sum(sc[key]["durations"])
        idle = sum(sc[key]["standing"].values())
        stats["total"].append(transport + idle)
        stats["transport"].append(transport)
        stats["idle"].append(idle)

    histogram(stats["total"], "Hours",
              "Histogram␣of␣total␣hours␣taken␣for␣all␣vehicles")
    histogram(stats["transport"], "Hours",
              "Histogram␣of␣hours␣in␣transport␣for␣all␣vehicles")
    histogram(stats["idle"], "Hours",
              "Histogram␣of␣hours␣idle␣for␣all␣vehicles")
    plot_median_mad(list(stats.keys()), stats,
                    "Median␣and␣MAD␣of␣total,␣transport␣or␣idle␣process␣duration",
                    "Process")


def search(method=None, agent=None, vehicle=None, node=None):
    durations = []
    time = []
    for edge in G.edges.data():
        if (not vehicle or edge[2]["vehicle"] == vehicle) and \
                (not method or edge[2]["method"] == method) and \
                (not agent or edge[2]["agent"] == agent) and \
                (not node or edge[0] == node):
            durations.append({"from": edge[0], "to": edge[1], "meta": edge[2]})
            time.append(edge[2]['duration'])

    return time


def histogram(data, label, title):
    plt.title(title)
    plt.xlabel(label)
    plt.ylabel("#␣Occurences")
    plt.hist(data, rwidth=0.95, align='mid')
    plt.savefig("figures/"+title + ".pdf")
    plt.close()
```

```
def plot_bar(x, y, title, label):
    plt.title(title)
    plt.ylabel("Vehicles")
    plt.xlabel(label)
    plt.bar(x, y)
    plt.savefig("figures/"+title+".pdf")
    plt.close()


def table(header, data, title):
    print("——————␣" + title + "␣——————")
    t = tt()
    t.add_row(header)
    for row in data:
        t.add_row(row)
    print(t.draw())
    print("\n")


def plot_median_mad(options, data, title, label, ylabel="Hours␣taken"):
    x, y, e, tabledata = ([] for i in range(4))
    for option in options:
        x.append(option)
        median = np.median(data[option])
        y.append(median)
        e.append(np.median(np.abs(data[option] - median)))

    plt.title(title)
    plt.ylabel(ylabel)
    plt.xlabel(label)
    plt.errorbar(x, y, e, linestyle='None', marker='o', capsize=4)
    plt.savefig("figures/" + title + ".pdf")
    plt.close()

    for stat in data:
        time = data[stat]
        median = np.median(time)
        tabledata.append([stat, median, np.median(np.abs(time - median)),
                          np.min(time), np.max(time)])
    table([label, "Median", "MAD", "Minimum", "Maximum"], tabledata, title)


def transport_stats():
    transports = {}
    tabledata = []

    for edge in G.edges.data():
        date = edge[2]["date"].strftime('%d/%m/%Y␣%H:%M')
        method = edge[2]["method"]
        if method not in transports:
            transports[method] = {"amount": 0, "dates": []}
        if date not in transports[method]["dates"]:
            transports[method]["dates"].append(date)

        transports[method]["amount"] += 1

    x, y1, y2, y3 = ([] for i in range(4))

    for method in transports:
        x.append(method)
        amount = transports[method]["amount"]
        batches = len(transports[method]["dates"])
        y1.append(amount)
        y2.append(batches)
        y3.append(amount / batches)
        tabledata.append([method, amount, batches, amount / batches])

    table(["Method", "Amount", "Batches", "Avg/batch"], tabledata,
          "Transport␣stats")

    plot_bar(x, y1, "Transportation␣-␣Vehicles␣per␣method", "Transport␣method")
    plot_bar(x, y2, "Transportation␣-␣Batches␣per␣method", "Transport␣method")
    plot_bar(x, y3, "Transportation␣-␣Average␣amount␣of␣vehicles␣per␣batch",
             "Transport␣method")


def find_node_agent_type():
    types = []
    agents = []
    nodes = []
    for edge in G.edges.data():
        # link nodes and method
        if edge[0] not in nodeType:
            nodeType[edge[0]] = edge[2]["method"]
            types.append([edge[0], edge[2]["method"]])
        # link agent and method
        if edge[2]["agent"] not in agentType:
            agentType[edge[2]["agent"]] = edge[2]["method"]
            agents.append([edge[2]["agent"], edge[2]["method"]])
        # link nodes and agent (multiple agents per node possible)
        if edge[0] not in nodeAgent:
            nodeAgent[edge[0]] = []
        if edge[2]["agent"] not in nodeAgent[edge[0]]:
            nodeAgent[edge[0]].append(edge[2]["agent"])
            nodes.append([edge[0], edge[2]["agent"]])

    table(["Node", "Method"], types, "Transportation␣method␣per␣location␣(node)")
    table(["Agent", "Method"], agents, "Transportation␣method␣per␣agent")
    table(["Node", "Agent"], nodes, "Agent␣per␣location␣(node)")


def node_stats():
    location = {}
    method = {}

    for node in G.nodes.data():
        if "durations" in node[1]:
            location[node[0]] = node[1]["durations"]
            if nodeType[node[0]] not in method:
                method[nodeType[node[0]]] = []
            method[nodeType[node[0]]] = method[nodeType[node[0]]] + node[1]["durations"]
            histogram(node[1]["durations"], "Hours",
                      "Histogram␣of␣idle␣times␣for␣" + node[0])

    plot_median_mad(list(location.keys()), location,
                    "Median␣and␣MAD␣of␣idle␣times␣per␣location", "Location")
    plot_median_mad(list(method.keys()), method,
                    "Median␣and␣MAD␣of␣idle␣times␣per␣method", "Method")
```

```
def process_edge_stats(options, find, title):
    stats = {}

    for option in options:
        if find == "Method":
            stats[option] = search(method=option)
        elif find == "Agent":
            stats[option] = search(agent=option)
        else:
            stats[option] = search(node=option)

        histogram(stats[option], "Hours", "Histogram "+option)

    plot_median_mad(options, stats, title, find)


def edge_stats():
    nodes = list(G.nodes)
    nodes.pop()
    time = search()
    histogram(time, "Hours", "Histogram duration all transportation processes")
    data = {"Average": time}
    plot_median_mad(["Average"], data, "Median and MAD of transportation duration",
                    "Duration")

    process_edge_stats(possibleMethods, "Method", "Median and MAD of transport methods")
    process_edge_stats(possibleAgents, "Agent", "Median and MAD of transport agents")
    process_edge_stats(nodes, "Location", "Median and MAD of locations")


def network_stats():
    n = G.number_of_nodes()
    m = G.number_of_edges()

    print("Number of nodes are: " + str(n))
    print("Number of edges are: " + str(m))


filename = "Sample2"
possibleMethods = []
possibleAgents = []
```

```
nodeType = {}
agentType = {}
nodeAgent = {}
arrivals = {"Early": [], "Precise": [], "Late": [], "Diff": {}}

file = load(filename)

G = process(file)
network_stats()

find_node_agent_type()

edge_stats()
node_stats()
general_stats(file)
transport_stats()
```