GOING NATIVE

Doing data science to understand user behaviour

ABSTRACT

In recent years the emergence of data-driven research methods has opened up new possibilities for developing understanding of user behaviour of digital products and services. To understand their impact, the primary researcher of this paper took on the perspective and role of a data scientist. By executing a case study in collaboration with a third party organisation that provides a file exchange service, we were able to use network analysis, data mining and machine learning in order to gain a new understanding of usage of the service. We describe in detail the approach taken and how this approach provided new insights into user behaviour. We report on how the effects of sharing the research results with the participating organisation, and finally how data driven methods, findings and predictions can become a powerful contribution to the arsenal of research techniques that are available to anyone studying the usage of digital products and services.

GRADUATION THESIS
Media Technology
MSc Programme,
Leiden University

AUTHOR
J.A. Schelling (1605976)

PRIMARY SUPERVISOR
Dr. P. W. H. van der Putten
Leiden University

SECONDARY SUPERVISOR
Dr. I.J. Mulder
Delft University of Technology

1. INTRODUCTION

The convergence of computation and communication has created a society that thrives on information (Witten et al, 2016). We record our interactions with digital artefacts at a daily increasing rate (Turner et al, 2014 via van der Aalst, 2016). Often these digital artefacts exist within a complex ecosystem of networked information technology products and services. Some scholars have termed this transformation *'datafication'* (Lycett, 2013). Cukier & Mayer-Schönberger, (2013) describe datafication as *"the ability to render into data many aspects of the world that have not been quantified before."* Others talk about *digitalization* (van der Aalst, 2016) in this context.

What is novel about these terms is that they describe a process that is distinctly different from the process of digitisation. Where digitisation focussed on the transformation of analogue reality, media, and signals into digital information that a computer can read, *digitalisation* or *datafication* refers to the process *"...of how contemporary social life, science and business are impacted by the usage of digital communication infrastructures."* (van der Aalst, 2016). As Cuvier & Mayer-Schönberger posit: *"it is a far broader activity; taking all aspects of life and turning them into data."*

This transformation is not happening silently. The process of deriving valuable insights from these data has become known as *data science*. A convergence of scientific fields that includes *machine learning*, and *data mining*. It can be understood as the recording of data through computational systems, analysing these data for patterns, structures, or expectations, which gives us information. Understanding the relations between these patterns produces knowledge. This knowledge allows us to make decisions, and gives us the opportunity to act on these decisions.

This growth in the availability of data is termed *'Big Data'* (Manyika et al, 2011) in popular media, and different authors (Gillon et al, 2012; Mithas et al, 2013) suggest that the use of big data and analytics can create value for organisations, as their usage can help organisations better understand *'its business and markets'* (Chen et al, 2012) and organisations using these analytical insights can be used to *'guide both future strategies and day-to-day operations'* (LaValle et al, 2011 via Sharma et al, 2014). In typical descriptions of the organisational context of big data, such as given by Abbassi et al (2016), the people involved with knowledge derivation tend to be systems architects, developers, data scientists and analysts. Yet, within organisations that develop digital products and

“... Planet Earth has never been as tiny as it is now. It shrunk due to the quickening pulse of both physical and verbal communication...

We never talked about the fact that anyone on Earth, at my or anyone’s will, can now learn in just a few minutes what I think or do, and what I want or what I would like to do.”

— Frigyes Karinthy, 1929, From the short story ‘Láncszemek’

services, designers play an important role in guiding the development and implementation of these strategies into products and services, as well as the evaluation of the day-to-day customer satisfaction of such a product or service.

With the emergence of these data-driven research methodologies, designers are in need of experience and skills to help them operate in an environment in which these types of data-driven analyses take place and these data-driven methodologies are used to inform decision-making. In traditional user-centered design practice, designers have been trained to understand the user's context through qualitative research methods (Norman and Draper, 1986). Solely focussing on qualitative research methods to inform a designer's understanding of user behaviour and subsequent decision making, could become problematic in a setting where quantitative data-driven methodologies are used. And while there is development of data-driven methodologies in design (Liikkanen, 2017), most of these new methods are not enhancing the research topics of designers.

The aim of this project is to understand the potential that data-driven research methods have in the design and understanding of usage of digital products and services. We start from the premise that digital products and services, due to their fundamentally digital, computational and networked nature, generate behavioural data about their usage, which can be studied to better understand how a service is used. This view inspired by examples of research projects that have been carried out into the behaviour of people interacting through digital platforms (Kramer et al, 2014) as well as new research tools that have recently appeared (Bakshy et al, 2014), and developments happening in the academic fields of the Computational Social Sciences (Cioffi-Revilla, 2010) as well as Network Science (Barabasi, 2015).

To understand the impact of these developments, the primary researcher of this paper took on the perspective and role of a data scientist. He familiarised himself with the typical skills, tools and methodologies of a data scientist, and undertook a realistic real-world data science project by collaborating with a third party organisation that provides a digital file exchange service.

In this article we report on the process, experience and our findings (see figure 1). We first extend the descriptions of the developments we have sketched in this introduction. To explore the functioning of these data-driven research methodologies we collaborated with a third party file exchange service to study the usage of their service within two countries in which the service is active. We report on the setup and results of this case study, and the reception of the results within the participating organisation. Finally, we reflect on the impact that the adaptation of data driven research methods can have on design research activities.

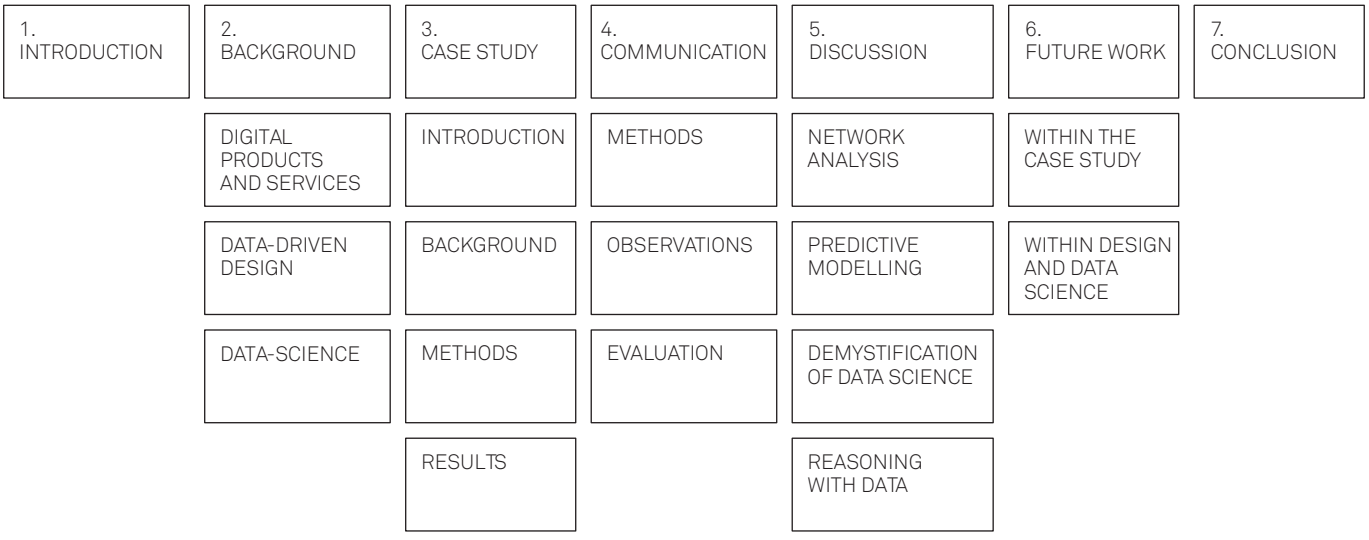


Figure 1. — Visual outline of this report

2. BACKGROUND

In this section we explore the background of data-driven research methodologies and their relation to the design of digital products and services. First we examine what constitutes a digital product or service and how these are designed. We then describe what data-driven methodologies are currently being applied in the design of products and services. We examine the development of the emerging field of data science, and conclude by examining typical data science and data mining processes that informed how we approached our case study.

2.1 DIGITAL PRODUCTS AND SERVICES

It has become commonplace in the Western world for people to interact with each other using the digital infrastructure we described in the previous section. The infrastructure that we built over the past 25 years has allowed us to digitally mediate our interactions with each other, and also with organisations. These interactions are often organised within the frames of digital products and services. Within the digital domain this distinction is somewhat hard to make; when a person performs a Google search, is he or she using it as a *product*, or is Google providing a *service*? Regardless of how we characterise the usage of these digital experiences, when designing these digital products and services designers often rely on design processes and research methodologies stemming from HCI research or traditional product design, for example Boeijen et al (2014).

During the 1980s designers started experimenting with bringing social scientists into the design process to get a better understanding of what to design instead of designing what was asked for. The adaptation of their research methodologies led to a better understanding of the social, cognitive and emotional needs of people (Sanders & Stappers, 2014). While many of these methodologies have been developed from a sound research practice rooted in product design and social sciences research, the complex network of information technologies in which digital products and services exist, makes them different from traditional products and services, and thus there is an opportunity to develop new design processes and research methodologies that are in closer keeping with the digital and computational nature of these digital artefacts.

2.2 DATA-DRIVEN RESEARCH METHODOLOGIES IN DESIGN

A significant difference when it comes to digital products and services is that they generate behavioural data from their users about their usage, that, when used via the internet is available to the creators of the product or service. This opportunity has been seized upon during the past few years and is being effectively exploited by large internet based organisations such as Microsoft, Facebook, Google and Yahoo (Likannen, 2017; Fisher et al, 2012). The different practices emphasise incremental improvements and selecting candidate designs based on exposing candidates to different groups of users, and selecting the most optimally performing candidates based on a previously determined variable. Within this approach, design becomes a form of hill-climbing, and different design candidates can be thought of as embodiments of different hypotheses (Schrage, 2014), with the behavioural data indicating the validity of the hypothesis, and thus the design.

This wide use of data can be characterised as a move toward a more evidence based practice within the design of digital products and services. Conducting large scale experiments, and performing A/B or multivariate tests is becoming common practice (King et al, 2017; Kohavi & Thomke, 2017). However, this approach has also revealed that designer-intuition is often at odds with data coming from tests or experiments, and according to Likannen (2017), *'... (this) forces designers to embrace failure and have a receptive mindset towards continuous iteration.'*. This approach aims to reach the perfect *exploitation* of the digital product or service.

However, using data only to validate hypotheses with regards to design changes, is just a single approach of using behavioural data from digital products and services. By building models of the usage of a service, we can also realise descriptions or predictions that have generalisation, predictive and sometimes explanatory or causal power. Some

studies (Kramer et al, 2014) (while controversial), for example study the extent to which social phenomena that arise in real-world interaction also arise within digital interactions, also enhancing our knowledge and understanding about how people us a digital product or service.

Within this study, we specifically looked at whether behavioural data can be used to enhance the *understanding* of the usage of a service. In order to do so, we looked at the emerging discipline of data science to provide us with a potential approach.

2.3 DATA SCIENCE

For the purpose of this study we keep to the definition of data science as it is given by Wil van der Aalst (2016), with an emphasis on translating data into value. This value may be provided in the form of predictions, automated decisions, models learned from data or data visualisation delivering insights. Each of these activities can be framed as a research activity with the aim of developing new knowledge.

With data science being an emerging and interdisciplinary field there are many (sometimes conflicting) views on the field (Cao, 2017). One can raise the question to what extent the field of data science aims to generate new generalisable knowledge, as much of it is executed in practice (O'Neill & Schutt, 2013). Some consider it a fancy term for statistics (as described in O'Neill and Schutt (2013) as well as Aalst (2016)), but it has its roots in many different fields. Aalst (2014) provides an extensive overview, in which we find obvious ancestor fields such as statistics, data mining / KDD, machine learning, predictive analytics and computer science, but also non-obvious fields, such as behavioural/social sciences, and business modelling and marketing. Since most data are generated by or from the interactions of people, it is important for the analysis of these data to understand human behaviour and the social context in which humans operate. Because data science is about deriving value from data, it can also be used to further business goals. We'll examine this specifically in our case study when we look at creating a predictive model to predict the likelihood a non-paying user will become a subscriber to a service.

Dhar (2013) defines data science as *"the study of generalisable extraction of knowledge from data"*, linking it to the practice of knowledge discovery and data mining, which in essence doesn't ask *"what data fits a prescribed pattern?"*, but rather *"which patterns fit the provided data?"*. The focus here is on deriving robust patterns from which predictions can be made. Extracting interesting patterns from a data set is non-trivial, as we will see in the case study, and constructing features that reveal interesting or promising patterns often requires a creative step. Dhar's observes that knowledge discovery can differ between scientific disciplines we visualise these differences in figure 2.

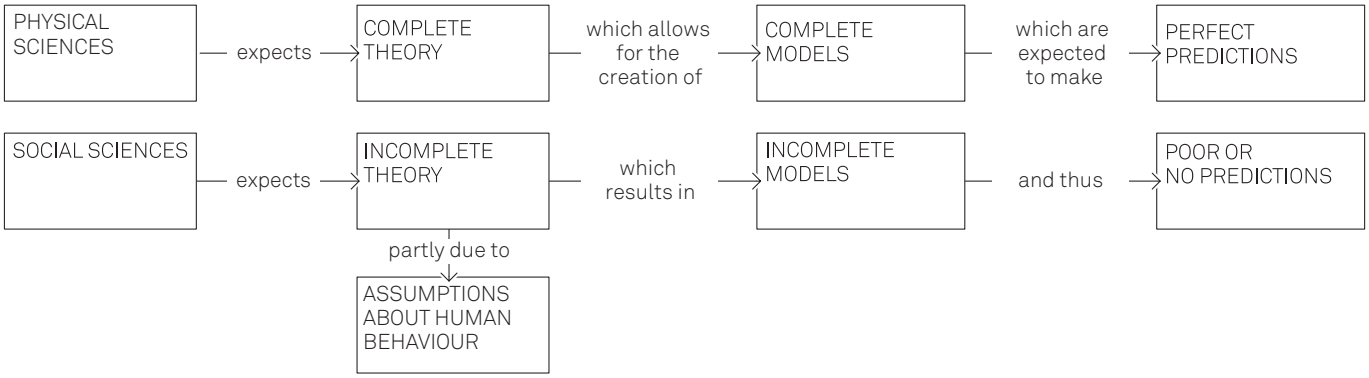


Figure 2. — Differences in knowledge discovery according to Dhar (2013)

Since data-science emphasises the quantitive study of phenomena through large data sets, models can often be made more complete, and as such could potentially realise better predictions. In order to epistemically assess whether this new knowledge gained through models is actionable, having strong predictive power is a major requirement. The emphasis on predictive power is especially strong in the machine learning and knowledge discovery in databases (KDD) communities (Dhar, 2013).

2.3.1 DATA SCIENCE PROCESS

Having an overview of a typical data science process provides the process that is necessary to follow through the steps of our case study. With that in mind we selected a typical practitioners' data science process model (O'Neill & Schutt, 2013) that show the steps that are expected in a data science project. This model has quite some similarities with the industry standard CRISP-DM (Chapman et al, 2000) process model shown in figure 3, which, while popular is no longer actively maintained.

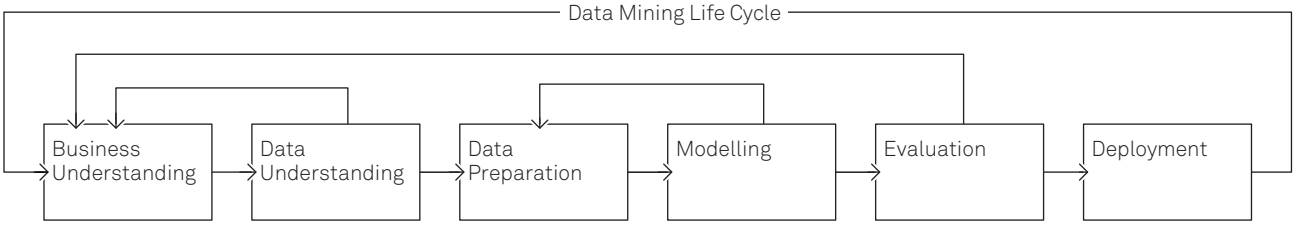


Figure 3. — Cross Industry Standard Process for Datamining (Chapman et al, 2000)

Like a CRISP-DM project, a typical data science project goes through multiple stages. Usually, research starts from a problem definition, but having a problem is not a firm requirement. Researchers could also find an opportunity to perform research from a personal curiosity, or a speculative question.

As can be seen in figure 4, first we have the 'real world' from which raw data is collected. The next step is then to cleanse this data and get it ready for analysis. Once we have cleansed data set there is the opportunity to do some exploratory data analysis (EDA). Through the EDA we might discover that our data is not as clean as we previously thought, or might discover outliers or gaps in the data that need to be filled by collecting more data.

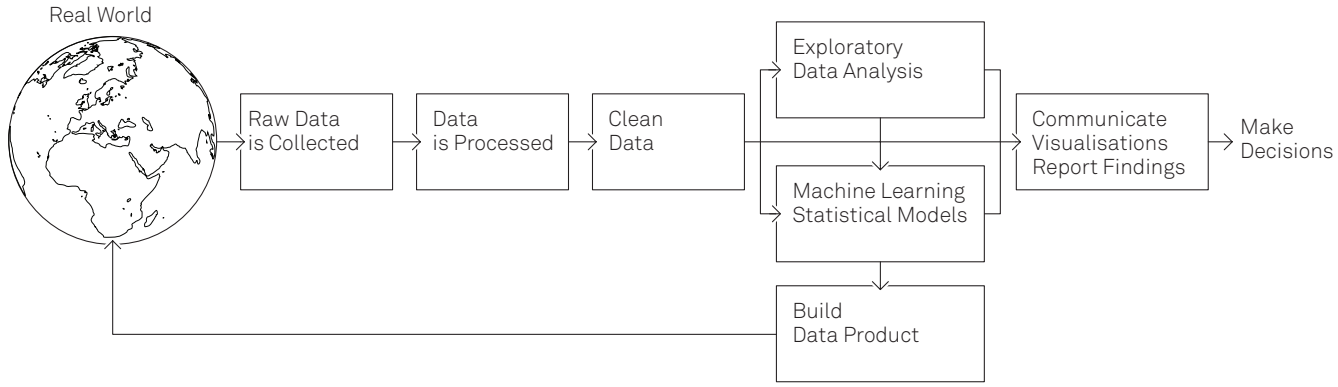


Figure 4. — A model of data science in practice (O'Neill & Schutt, 2013)

Next, we design a model that solves the particular problem or captures an answer to the question the researcher posed. Typical problems are prediction problems, classification problems or plain descriptive problems. After this step results can be interpreted, visualised, reported or communicated. Alternatively (and a step often found in practice) is the creation of a data product. A specific type of product or feature that relies on data in order to perform its task, such as a spam classifier, or a search ranking algorithm or recommendation system. If a data science project is taken into a production environment, researchers need to be aware that they've now created a feedback loop, where their data product interacts with the environment in which they've performed their original research.

2.4 CONCLUSION

It has become common for people to interact with each other and organisations through a networked digital infrastructure. Product and services built on top of this structure are often designed by incorporating qualitative research methods adapted from HCI research or traditional product design literature. In recent years, the usage of data-driven research methods to improve the design of digital products and services have emerged.

These methods emphasise large scale experimentation and iterative improvements in order to enhance the design of a digital product or service. However, data-driven research methods can also be applied to enhance the understanding of the usage of a digital product or service. In order to understand the potential for data-driven research methodologies to enhance the understanding of the usage of a digital product or service, we collaborated with a third party organisation. We report on this research in our case study.

3. CASE STUDY

In this section we describe the case study that was executed as part of this project. To explore how data-driven research methods might affect the understanding of the usage and user behaviour of a digital product or service, we collaborated with an organisation that provides a digital file exchange service. For the purpose of this research, the company has requested to remain anonymous. We first describe how this collaboration was established and what questions were formulated to guide the research. We then report on the methodology and results of the research. Finally we examine how the results from the research efforts were received during a workshop with different stakeholders within the company.

3.1 INTRODUCTION

The core product of the service allows users to cary out a basic task, which is the exchange of large format digital files. While this seems like a trivial service, the technical complexity of this task is something which is hard to cary out for the average user, as there are many restrictions on how large files can be transmitted through email or other digital networks. Another aspect of the exchange of large digital files is the conformation that this large file was downloaded correctly, as well as giving conformation to the user that initiated the sending of the file that a file was downloaded by the receiving party.

3.1.1 COMING UP WITH QUESTIONS THAT DATA-DRIVEN RESEARCH CAN ANSWER

To establish a number of questions or hypotheses that could drive the research, we met with the organisation that facilitates the file exchange service and approached them with a proposal in which we described how studying the usage data of the service using data-driven research methods could help them in gaining a better understanding of how the service is being used. Figure 5 shows a conceptual model of the file exchange service functions.

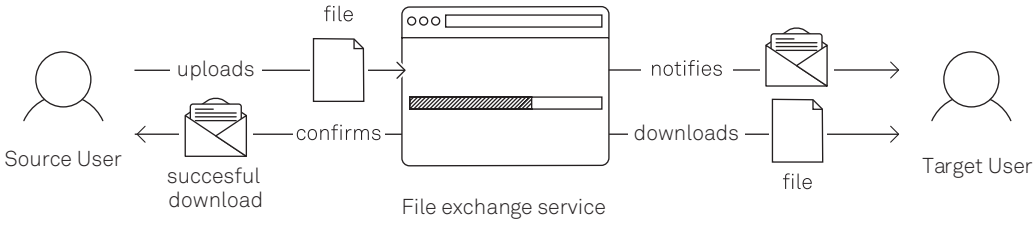


Figure 5. — Conceptual model of the file exchange service.

To study the behaviour of users in aggregate we proposed to model the usage of the service as a network, to determine wether we could observe the social organisation behind the usage of the service.

During our initial conversation we also found that one of the issues that the organisation was interested in was determining wether data-driven research could shed some light on another question they had around user behaviour. The service is offered in two tiers, a free tier, as well as a paid subscription service. To discover the motivations for becoming a paid subscriber, the company had done qualitative research, but the results were inconclusive, as users gave highly diverse motivations for subscribing. Could data-driven research methods provide another way of finding which shared attributes of a user could provide better insight into whether he or she would become a subscrib-

er? For this we proposed to create a predictive model that predicts the conversion from a free user to a paying subscriber. Examining the attributes that could lead to a successful prediction would provide an answer to this question.

So to summarise, our data-driven research consisted of two parts; network analysis to determine whether we could model the social organisation behind the service to gain new insights about product usage, as well as predictive modelling in order to predict the conversion from a free user to a paying subscriber. In the next section we describe some of the backgrounds of these specific data-driven research methods.

3.2 BACKGROUND

For the execution of our data-driven research we agreed with the participating company on two strands of research, network analysis as well as predictive modelling. In this section we give a short background on both fields of study and how they fit within the research executed for this case study.

3.2.1 NETWORK ANALYSIS

For many users the usage of the file exchange service tends to be transient, but together with the participating company we hypothesised that the interactions that people have through the service is influenced by the way they are socially organised and having a model of the social organisation could provide new insights into product usage.

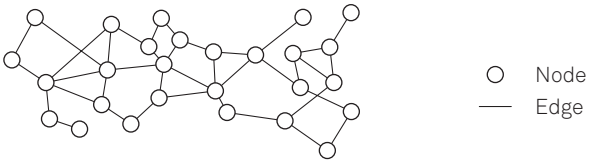
The service facilitates the interactions between millions of individuals, and can be thought of as a *complex system*. Network science deals with the study of the behaviour of complex systems, by modelling their interactions (Barabasi, 2015). When dealing with phenomena that arise out of complex systems, it is often the case that collective behaviour cannot be explained sufficiently by looking in isolation at the behaviour of the entities that make up the group. In short, while the service isn't a social network like Twitter or Facebook, it should be possible to observe a network of human interactions from the data of individual file exchanges. Understanding the properties of this network could potentially influence the understanding of the usage of the service.

Using network analysis, we are be able to study the usage at different scales:

- On a macro country level
- On a meso community level
- On a micro individual user and neighbourhood level

At its core, a network describes a pattern of interconnections among a set of things or entities. However there is a great diversity of contexts in which networks are evoked (Easly & Kleinberg, 2010). In this study we're interested in understanding the complex web of social interactions

BASICS OF GRAPH THEORY



Graphs are abstract structures that consist of sets of nodes and edges. They describe entities that are are in some sense related.

DIRECTED AND UNDIRECTED GRAPHS

6 nodes
8 edges

6 nodes
12 edges

Graphs come in many flavours. Shown here are a directed graph, and an undirected graph. In a direct graph, a connection from A to B is not the same as a connection from B to A. In an undirected graph edges are counted twice.

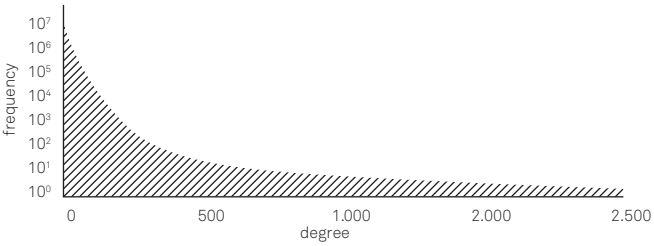
DEGREE

node 'B' has an *in degree* of 4 and an *out degree* of 3

node 'B' has a *degree* of 5

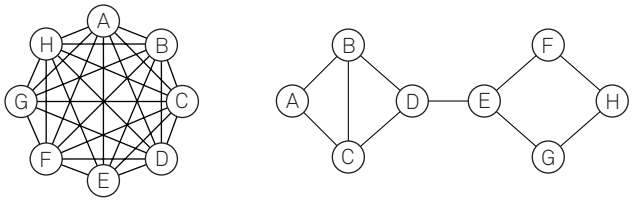
Degree indicates how many edges connect to a node. In directed graphs, we differentiate between in degree (number of incoming edges) and out degree (number of outgoing edges)

DEGREE DISTRIBUTION



The degree distribution shows the distribution of degrees for the whole network. Real world networks often have a skewed degree distribution; a large majority with a low degree, but a small number nodes with a high degree acting as hubs.

DENSITY



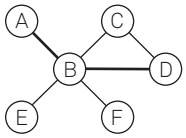
8 nodes
56 edges density: 1.0

8 nodes
20 edges density: 0.35

Graphs can be dense or sparse. This is determined by how close the number of edges is to the maximal number of edges. Graphs representing human phenomena are often sparse.

PATHS & DISTANCE

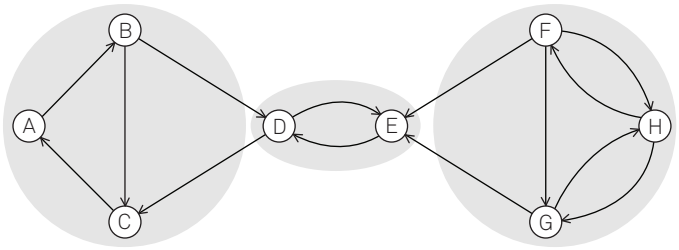
The shortest path from A to D is through B.



The distance between A and D is 2.

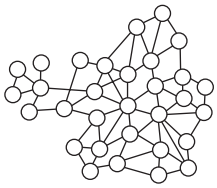
Paths describe routes from nodes to other nodes. In Network Science we're especially interested in *shortest paths*. The distance between two nodes is the number of edges in the shortest path.

COMPONENTS

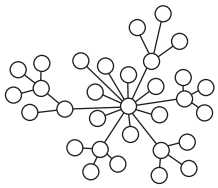


Components describe local areas within the graph where nodes are more strongly connected to each other.

ASSORTATIVITY



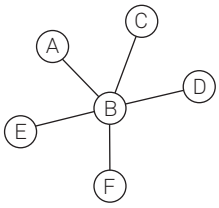
Degree assortativity



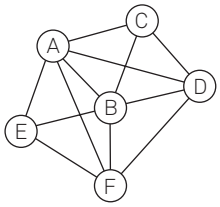
Degree disassortativity

Assortativity describes the property of how similar nodes with similar properties attract each other. Degree assortativity is a well known example, but it is also relevant for other attributes.

CLUSTERING COEFFICIENT



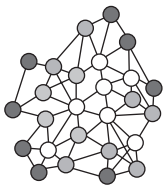
B has a clustering coefficient of 0.



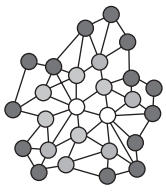
B has a clustering coefficient of 0.7

The clustering coefficient describes the extent to which nodes tend to cluster together. In real-world networks we tend to find a high average clustering coefficient compared to random graphs.

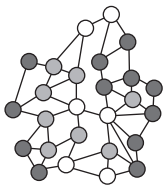
CENTRALITY



Degree Centrality



Closeness Centrality



Betweenness Centrality

Centrality measures indicate the most important node in a graph. Applications include identifying influential individuals (degree centrality), the core nodes within the network (closeness centrality), or key infrastructure nodes (betweenness centrality).

that take place between people using digital products and services through networked digital information systems. This understanding can concern itself with understanding the structure of a network, but also its aggregate behaviour and its dynamics. Like the field of data science we've described previously, the field of network science is highly interdisciplinary. Some authors even argue that network analysis can be seen as a sub-field within data science (O'Neal & Schutt, 2013).

The key concepts of network science are rooted in graph theory, a branch of mathematics. Yet, unlike the abstractions of mathematics, network science is described as highly empirical. Network theory seeks to apply its theories on data derived from the real world, and its value is judged by the measure in which it offers insights about the properties and behaviour of a complex system. It is this emphasis on empirical measurement and observation that suggested to us that there is potentially a good fit between the large scale of complex social interactions that are mediated by digital products and services, and the study of these interactions through the lens of network science. The quantitative nature of these networked information systems provides us with the empirical data that network science requires, whilst the context within which a digital product or service is placed provides us with the social phenomena we can study using network science.

3.2.2 DATA MINING, MACHINE LEARNING, AND PREDICTIVE MODELLING

As we stated in the previous section, one of the questions that was posed to us by the organisation participating in our research, was if our research could provide a better understanding to the conversion of a free user to a paying subscriber. To do so we applied several data mining and machine learning techniques.

In data mining, there are two extremes for expressing a pattern; a black box whose working is incomprehensible, or a transparent box, whose construction reveals the structure of a pattern (Witten et al, 2016). The difference is in whether the patterns that are mined can be represented in such a way that the structure can be examined, reasoned about and used to inform future decisions. These patterns are called structural patterns.

It's a bit harder to give a concise definition of machine learning, as learning can mean many things, but since we previously created an operational definition for data mining, we need to come up with an operational definition for learning. We could state that a thing learns when it can change its behaviour so it can perform better in the future. This ties learning to performance rather than knowledge (Witten et al, 2016). However, when we discuss the change of behaviour to improve performance, we usually talk about *training*, instead of *learning*. So, by *machine learning*, we mean to train a *model* that can perform a task, such as correctly (within an accuracy range) predicting the class of an observation that wasn't part of the original training dataset. But sometimes the goal is

not to predict, but to better understand data as it finds abstractions that describe the observed behaviour.

Since we're interested in determining what user attributes contribute to the conversion of a free to paid user, we want to train a model that can predict this conversion. By using algorithms that make their internal structure visible, we can reason about the attributes of a user that predict this conversion.

3.3 METHOD

We briefly describe the tooling that was used as part of this case study, after which we will describe the data processing steps that were taken to prepare the data we received. We then constructed our models and carried out the analysis necessary to achieve our previously stated goals. Using the models we derived new information about user behaviour and usage of the service.

3.3.1 TOOLING

To carry out the research as part of this case study we used numerous tools. We distinguish between two categories of tools, computational tools and ready-made tools. Often when performing data science research the size and scale of the data requires a computational approach to create a successful analysis or data product. This means that the researcher performing the modelling and analysis writes a computer program that is specifically tailored to the research activities that are carried out. Ready-made tools on the other hand, are used to carry out more general, but specific research tasks.

The chosen toolset used to do the computational analysis is based on a set of interlocking technologies based on the Python (van Rossum, 1995) programming language that is often used in scientific computing. For basic numerical manipulation we used the NumPy library (van der Walt et al, 2011). Built on top of this is the Pandas data analysis library (McKinney, 2010) that allows the manipulation and analysis of structured data. To create the network model we initially started with the NetworkX library (Hagberg, Schult and Swart, 2008), but it became unsuitable due to performance issues at the scale of the network model we were creating. This led us to switch to using the Graph-Tool (Peixoto, 2014) library for our network modelling and analysis. Finally, results from our analysis were visualised using the MatplotLib (Hunter, 2007) and Seaborn libraries (Waskom et al, 2017), which provide data visualisation capabilities. Tying all these disparate pieces of technology together is the Jupyter interactive computing environment (Kluyver et al, 2016), which allows researchers to combine code, interactivity, text and graphics into a notebook environment that documents their research in reproducible form.

To gain further insight into subcomponents of the network models we used Gephi (Bastian et al, 2009), a popular software tool for network visualisation and analysis. To visualise networks larger than 10.000 nodes and edges we used GPUGraphlayout (Brinkman et al, 2017). Other visualisations were made using RAWGraphs (Mauri et al, 2017), a frontend for the D3.js (Bostock, Vadim, and Heer, 2011) visualisation library. Processing (Reas & Fry, 2007) was used to create geographical visualisations. To create the predictive model we relied on WEKA (Waikato Environment for Knowledge Analysis) (Witten et al, 2016), a machine learning workbench that provides a researcher with capabilities to both explore and experiment with machine learning algorithms. It supports several standard datamining tasks, such as data preprocessing, clustering, classification, visualisation, regression and feature selection.

Keeping with the practical data science model we examined in section 2, we needed to go through a sequence of data processing steps to create a dataset that was suitable for exploration and modelling.

3.3.2 DATA PROCESSING

To create the appropriate models that we can use to analyse our data, we need to transform the raw data into structured data that we can easily manipulate, query and visual-

ise. The initial form in which we received the data takes the form of a transactions-log. A long list of events, describing the exchanges of files between different users at a specific moment in time. Before we can create our models, we first need to go through a number of intermediary data processing steps in order to successfully create models that can give us new insights and predictions.

These steps are as follows:

1. Acquiring the data in its raw form
2. Parsing raw data into structured (tabular) data
3. Data cleansing - removing unnecessary, incomplete or damaged data
4. Feature construction - deriving new features from existing features or using other data sets
5. Descriptive statistics - describing and summarising the data

DATA SET ACQUISITION

The initial data set was received as a CSV (comma separated values) text file dumped from the service's database. The received data set covers the file exchange data during a period of 34 days, starting from the first of January 2017, lasting until the third of February. The data set consists of transactions originating from two countries with a comparably sized user base. Table 1 shows the basic properties of the data set.

DATASET PROPERTIES BEFORE CLEANSING	
Start of the observation period	02/01/2017 00:00
End of the observation period	04/02/2017 08:39
Duration of the observation period	33 days 08:39:30
Number of countries in the dataset	2
Number of file exchanges	4.203.885
Number of file uploaders	1.159.653
Number of file downloaders	2.326.862
Number of unique email adressess	3.023.805

Table 1. — Dataset properties

DATA PARSING

In order to create the appropriate models that we can use to analyse our data, we need to structure the data using different data types. The data types of all the features in the data set are listed in table 2. For brevity identical features of both the source and target users are listed on the same row.

DATA FEATURE	DATA TYPE	DESCRIPTION
`exchange_id`	integer	Unique number identifying this file exchange
`filesize_bytes`	integer	total size of the exchanged files
`source_email, target_email`	string	e-mail address of both the sender and receiver
`source_ip, target_ip`	string	ip address of both sender and receiver
`source_country_code, target_country_code`	string	the ISO-3166 country code of the location of both sender and receiver
`source_upload_timestamp, target_first_download_timestamp`	timestamp	the exact moment in time the file was uploaded and stored on the servers of the service, as well as the exact moment when it was first downloaded by the target user
`source_user_type, target_user_type`	string	categorical value, tells wether a user is a subscriber to the paid variant of the service

Table 2. — Dataset features and datatypes.

DATA CLEANSING

After parsing the raw data into the proper structured format, it is ready for data cleansing. The process of cleansing data aims to verify that a data set is complete (no missing values), accurate (each variable conforms to a specific data type), consistent, and relevant. A second step that is popular in data science practice is ensuring that the data is

'tidy' (Wickham, 2014). This ensures that each feature in the data set forms a column, each observation forms a row, and that each observational unit forms a table, as can be seen in figure 6. Tidy data facilitates easy manipulation, visualisation and modelling.

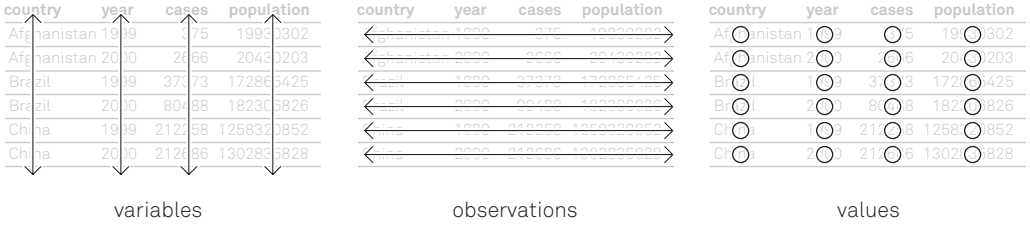


Figure 6. 'Tidy' Data (Wickham, 2014).

During the cleansing step we found that there were numerous rows in which variables had missing values. After close investigation we determined that there were 123 entries in which there was no `source_email` set. These 123 entries were removed from the data set. Furthermore, a significant amount of file exchanges did not have both the `first_download_timestamp` and `target_ip` set. We interpreted these missing values as a file exchange not being successful. After removing these unsuccessful file exchanges we were left with 3.204.518 successful file exchanges as can be observed in table 3.

DATASET PROPERTIES AFTER CLEANSING	
Start of the observation period	02/01/2017 00:00
End of the observation period	04/02/2017 08:39
Duration of the observation period	33 days 08:39:30
Number of countries in the dataset	2
Number of file exchanges	3.204.518
Number of file uploaders	1.008.413
Number of file downloaders	1.707.447
Number of unique email adressess	2.385.469

Table 3. — Dataset properties after the data cleansing step.

FEATURE CONSTRUCTION

Our next step is to construct new features (attributes) based on the features that are already present in the data set, or that can be derived from combining with, or filtering through other data sets. These newly constructed features help us in our analysis by allowing us to look at aspects and patterns in the data set that might otherwise not be visible.

When examining the frequency counts for the domains occurring within the data set, we found that the top 20 most frequent domains, are from companies that provide free email services. Table 4 gives an overview of the top 10 domains.

RANK	DOMAIN	COUNT
1.	gmail.com	517.750
2.	hotmail.com	87.476
3.	yahoo.com	65.707
4.	*omitted for confidentiality*	38.070
5.	*omitted for confidentiality*	24.718
6.	me.com	18.535
7.	*omitted for confidentiality*	17.794
8.	*omitted for confidentiality*	14.756
9.	outlook.com	12.963
10.	icloud.com	12.831

Table 4. — Top 10 of most frequently occurring email domains in the dataset.

In order to make better sense of the domains that are present in the data set, we decided to filter the domains using another data set. Since we found such a large presence of free e-mail providers in the data set we looked for a way to identify wether a domain belonged to a provider of free e-mail services. We found a public project called 'Freemail' (White, 2017) that keeps an extensive list of domains that provide free or disposable e-mail addresses (see table 5).

Based on their list we created a table of domains with a categorical variables, to label a domain as a provider of free,disposable, or blacklisted email addresses. Disposable e-mail addresses are often services that provide an email address for one-time usage. Using this list we filtered the domains in our data set. Domains that were not present on the list obtained from 'Freemail' were labelled as private.

After labelling all the domains present in the data set we made a non-obvious observation: while the top 10 email domains in the data set consisted of addresses from free e-mail providers, we actually found that 62,2% of the labelled domains were labelled as coming from private domains (see table 6).

DOMAINS IN FREEMAIL DATASET	COUNT
free e-mail domains	4.316
disposable e-mail domains	349
blacklisted e-mail domains	4

Table 5. —Summary of the properties of the Freemail dataset.

DOMAIN TYPE	COUNT	%
private	1.485.912	62,2%
free	899.481	37,7%
disposable	76	0,01%

Table 6. —Results of the domain classification step.

GEOLOCATED IP ADRESSES	COUNT	%
unique ip's	2.215.298	100%
succesfully identified	2.002.523	90,4%
unsuccesfully identified	212.775	9,6%

Table 7. — Results of the geolocation step.

Having labelled the types of domains we find in the data set, we proceeded with, geolocating the IP addresses in the set. Geolocation is the estimation of the geographical location of an object, such as a computer connected to the internet. By geolocating the IP addresses in the data set it becomes possible to say something about the geographical distribution of the data set, as well as the geographical location of the user base present in the data set. The primary motivation for geolocating the users in the data set is that having the geographical location allows us to visualise the network of users geographically. Which helps us to understand something about the spread of users across the world. IP-Geolocation is inherently imprecise, but it is precise enough in order to pinpoint an IP address to a particular city. It should not be used to identify particular households. Table 7 shows the results of the geolocation step.

We used a freely available database from a commercial vendor (Maxmind, 2017). This free database can be used to successfully locate 90% of all IP's in the IPV4 range. While it doesn't cover the entire set of IP's in our data set, after geolocation 2.002.523 ip's were successfully located. For our geographical network visualisation we removed all results from the data set that were unsuccessfully identified.

After taking these steps we now have five extra features for every observation in the data set. In order to distinguish between data from our primary source in our data set and data from third party sources, we post-fixed the columns that were derived using these third party sources with a reference to the original data source.

SPLITTING THE DATA SET

Having completed the main feature construction step, we split our data set in 3 individual sets:

- One set containing the full data set of file exchanges
- One set containing the file exchanges originating from country A
- One set containing the file exchanges originating from country B

After splitting our data set into these three sets we needed to take care of an anonimisation step, in order to be able to perform analysis and / or storage in a remote computing environment, which we ended up doing when performing the network analysis.

ANONIMISATION

While care is taken to handle sensitive private information in accordance to regulations and agreements made with the data providing company, during our research project there were occasions when it was be necessary to perform an analysis in a remote computing environment, because it was better equipped to handle specific types of analysis. In order to create a version of our data set for these types of occasions we went through a basic anonimisation step. First, each feature of the data set was split into its individual column, after which the unique values were counted. These unique values were then merged with the original column to create an ID value for every occurrence of a value in a column in the data set. Simultaneously a lookup table was generated, containing the combination of ID's with the original datapoints. An anonimised copy of the data set could then be restored to its original state by going through the lookup tables and reassociating every ID with its original value. A procedure like this does not re-arrange any patterns that are present in the data, it just removes directly identifiable features such as e-mail addresses and makes it somewhat difficult to reconstruct the original data. If any of the patterns were to have identifying value, a third party that obtained a copy of the data set would have to go through numerous difficult steps in order to recreate the original data set without access to the lookup tables.

With our anonimisation features present in the data set we could move to the next step and collect basic statistics about our data set.

BASIC STATISTICS

Having taken care of anonomisation, we can now make a statistical summary of our data sets and compare the two different countries present in the set, as can be seen in table 8. As the current shape of the data set is in the form of a transactions-log we might observe some patterns that we want to explore further at a later stage, when we have created the user and network models.

FILE EXCHANGE SUMMARY	FULL DATASET	COUNTRY A	%	COUNTRY B	%
total number of file exchanges	3.204.518	1.708.984	53,3%	1.495.534	46,7%
unique email adressess	2.385.469	1.228.509	51,5%	1.183.169	49,5%
file exchanges from free to free domains	709.959	261.803	36,9%	448.156	63,1%
file exchanges from free to private domains	494.686	239.049	48,3%	255.637	51,7%
file exchanges from private to free domains	368.645	188.786	51,2%	179.859	48,8%
file exchanges from private to private domain	1.672.824	1.038.651	62,1%	634.173	37,9%
file exchanges from free user to free user	2.969.300	1.598.311	53,8%	1.370.989	46,2%
file exchanges from free users to subscribers	95.552	43.891	45,9%	51.661	54,1%
file exchanges from subscriber to free user	137.696	66.372	48,2%	71.324	51,8%
file exchanges from subscriber to subscriber	12.848	5.336	41,5%	7.512	58,5%

Table 8 . — Basic statistics of the file exchange dataset.

One the patterns we can already spot, is that while total usage is lower in country B it has more subscribers than country A. Also, in country B we observe that there are many fewer exchanges between users with private email domains. As e-mail domains are often part of organisations, this suggests that in country A the professional usage of the service is higher than in country B. As we can see from the number of destination countries we can infer that there's a global audience using the service. All of these observations merit further exploration, when analysing the network models and predictive models.

With these steps we've transformed our raw data into structured data, cleansed the data set and constructed new features from existing features, as well as through combination with other data sets. We took care of anonimizing the data for external processing, and split the data into three unique sets. In this shape the data set is now ready to make different models.

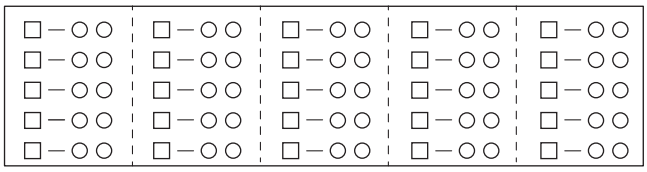
3.3.3 NETWORK MODELS

To construct our network model we started with the cleansed transaction data. Examining our data set we find that for every `exchange_id` we have an ip-address, country code, email address, and domain and username of both source and target users. Since the `exchange_id` is shared by both source and target users, we can *connect* both users through the file exchanges they perform using the service. So with this information we can construct a network with two different node types. *Users*, connected to each other through *file exchanges*. However, two-mode networks are rarely used for analysis, since most network measures are only defined for one-mode networks. In order to create a one-mode network out of our two-mode network we need to *project* the two-mode network to a one-mode network. Opsahl (2013) describes a procedure for this: *"Projection is done by selecting one of the sets of nodes..."*_ (in our case country codes, domains or e-mail addresses), *"...and linking two nodes from that set if they were connected to the same node of the other type."* Information from the two-mode network can be incorporated into the one-mode network through edge weights and properties.

With this procedure and our data set we were able to construct three different types of network models:

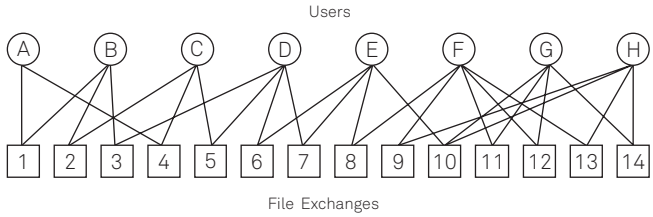
- A country model, representing the flow of file exchanges between countries
- A domain model, representing the flow of file exchanges between domains
- An e-mail model, representing file exchanges between different e-mail addresses.

CREATING A GRAPH FROM TRANSACTION DATA



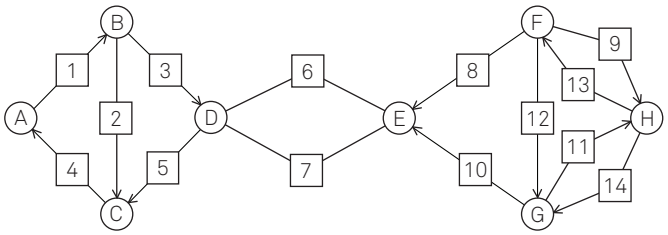
Our transactions dataset consists of file exchanges and source and target user values associated with the file exchange.

CONSTRUCTING A BIPARTITE GRAPH



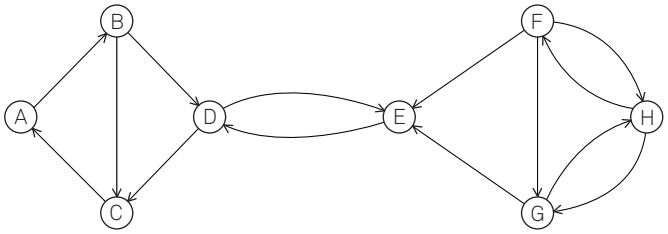
From this transaction data we can create a bipartite graph. In a bipartite graph there are two types of nodes in the network (users and file exchanges). All edges have their endpoints in different node sets.

TWO MODE NETWORK



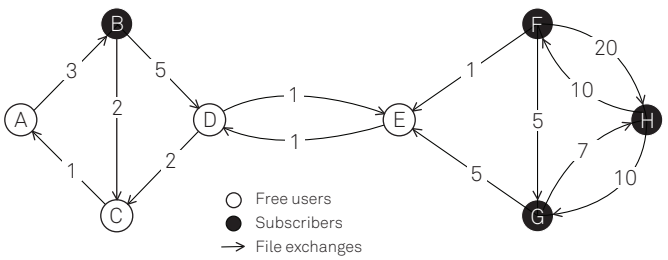
This two mode network has the same structure as the bipartite graph, only the node positions were rearranged to show how every user is connected to another user through a file-exchange.

PROJECTING A TWO-MODE NETWORK TO A ONE-MODE NETWORK



Since every user is connected to another user through a file exchange we can project the network to a one-mode network. In this network users remain nodes, but file exchanges become the directed edges of the network.

ADDING NODE AND EDGE ATTRIBUTES



Finally, we can add node and edge attributes to represent the usage of the service. Edge weights for example can be added to show the frequency of communication, as well as node properties showing if a user is a subscriber.

Each of the three models can answer different questions about the usage of the service. And every model represents the usage of the service at a finer level of detail, as countries in the data sets number in the hundreds, domains in the data sets in the thousands, and e-mail adresses in the millions. Since we split our data sets into three sets, we created these three different models for every subset of our data set. One for country A, one for country B, and for the combination of both countries, as it is likely that usage of the service is not delimited by borders.

We modelled our networks as directed graphs. So a file exchange from source A to target B is not the same as a file exchange from source B to target A. This allows us to say something about the reciprocity the service's users. It's unlikely that all users use the service to the same extent. It's more likely that we'll observe users who are strong senders, just as well as there might be users that are strong receivers, and we want to learn about both different types of users.

In order to create the graphs, we for every model we went through the following basic steps:

1. Create an edge-list using the transaction data.
2. Calculate the edge weights, for the different models.
3. Create a graph using the edge list.
4. Collect basic graph properties.
5. Add node and edge properties to the graphs.
6. Collect network measures.
7. Extract the Largest Connected Component (LCC) as well as other components.
8. Describe properties of the components.
9. (Optionally) visualise the different graphs.

For our analysis we constructed the network for the entire duration of the observation period. During this analysis we did not examine any dynamic properties of the network.

COUNTRY NETWORK

The country network is made up out of the two countries from which exchanges in our data set originate as well as the target countries that the file exchanges were downloaded from. In this network model the nodes represent the different countries and the edges represent the number of file exchanges taking place between the countries in the data set. Since we only had two originating countries in our data set, only back and forth traffic between country A and country B could be observed. For all the other countries we could only observe file exchanges originating in country A and B being downloaded. With this model being a limited network, aside from basic properties we did not collect any network measures.

We created an edge-list using the country codes that were present for every file exchange in the data set. With this edge-list we created the country network as a directed graph. For every edge we calculated the weight based on the frequency of file exchanges going into the edge direction. For every edge we then calculated the relative amount of exchanges going into that direction compared to the whole network. Figure 7 shows a simplified example of the structure of the country network.

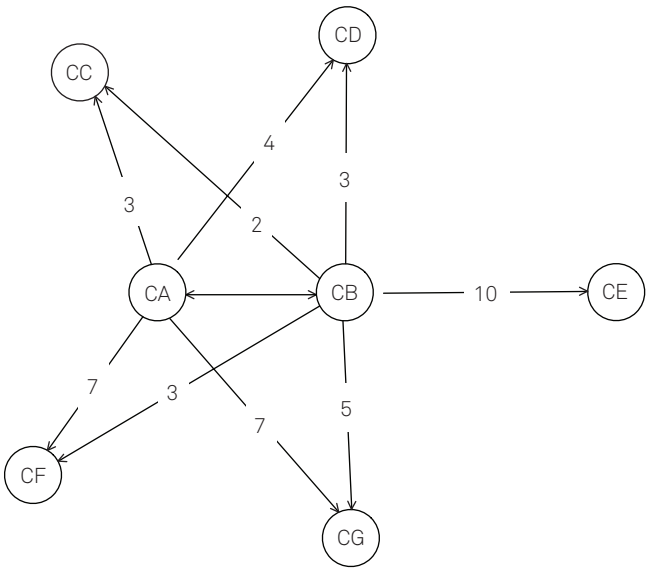


Figure 7. Example of the country network model. For privacy reasons the country codes are replaced with generic codes.

DOMAIN NETWORK

The domain network is made up out of the domains that are present in the data set. In this network model the nodes represent the domains in the data set and the edges represent the file exchanges taking place between the domains. The edge weights show the frequency of a file exchange taking place in a specific direction. As domains are often representative of the organisation that a user is associated with, the domain network model can tell us more about usage of the service inside and across different organisations. For the domain networks we collected network measures and extracted the largest connected components.

We created an edge-list using the domains that were present for every file exchange taking place between different domains. The domains are taken from the `source_domain` and `target_domain` variables. With this edge list we created the domain network as a directed graph. For every edge we calculated the weight based on the frequency of file exchanges going into an edge direction. Figure 8 shows a simplified example of the structure of the domain network.

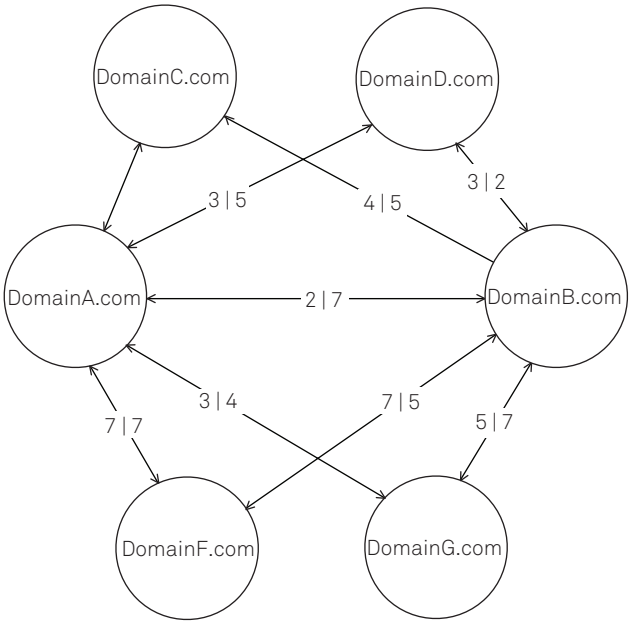


Figure 8. Example of the domain network model. Domain names are replaced to protect users' privacy.

E-MAIL NETWORK

The e-mail network is created based on the e-mail addresses present in the data set. In this network model the nodes represent individual e-mail addresses and the the edges represent the file exchanges between e-mail addresses. As users can specify different e-mail addresses as a source there is potentially a many to one relationship between e-mail addresses and individual users. For the creation of this model we assumed that each unique e-mail address mapped to a unique user.

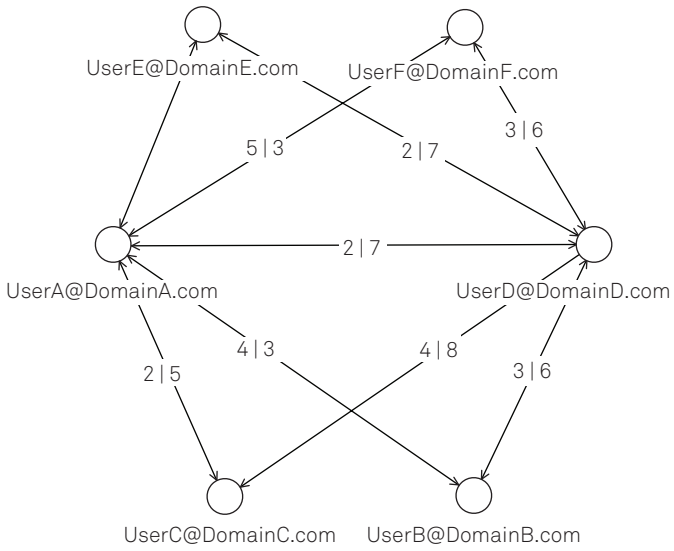


Figure 9. Example of the e-mail network model. Actual e-mail adresses not shown to protect anonimity.

We created an edge-list using the `source-email` and `source-target` variables, for every exchange taking place between different e-mail addresses. With this edge list we created the e-mail network as a directed graph. For every edge we calculated the weight based on the frequency of file exchanges going into an edge direction. Figure 9 illustrates the structure of the user network. We then added a number of node and edge properties that we could use when further studying the network (Table 9).

After populating the graph with node and edge properties we collected a number of network measures and extracted the Largest Connected Components. These are the components in the network that hold the largest amount of nodes and every node is connected to every other node. We extracted both strongly connected components as well as weakly connected components. We then extracted the next largest 50 components for further study. We calculated network measures for all the extracted components, as well as component summaries based on the node and edge properties. These components in turn were visualised for further analysis and presentation. Depending on the scale of the components we used either Gephi (for networks with less than 10.000 nodes and edges), or GPUGraphLayout (for networks with over 10.000 nodes and edges).

	PROPERTY	DESCRIPTION
NODE PROPERTIES	`email`	the e-mail address of this user
	`username`	the username of the e-mail address of this user
	`domain`	the domain of the e-mail adress this file exchange originates from
	`country_code`	the country_code that's associated with the e-mail adress
	`component_id`	the id of the component in the network this node is part of
	`smallest_file_exchange_size`	the file size in bytes of the smallest file the user sent
	`average_file_exchange_size`	the average file size in bytes of the files the user sent
	`largest_file_exchange_size`	the filesize of the largest file the user exchanged
	`user_type`	the inferred user type of the user
EDGE PROPERTIES	`weight`	the number of file exchanges happening during the observation period in that direction

Table 9. Node and edge properties for the e-mail network.

3.3.4 PREDICTIVE MODELS

We created two predictive models based on variations of the same data set. Our goal was to train a model to see if it could successfully predict the conversion of a free user to a subscriber. We created two versions of the data set as we were curious to see wether shared neighbourhood properties had any influence on the conversion of a user. In order to create a successful predictive model we had to go through a number of steps:

1. Prepare the user data using the transactions data set and data gathered from the e-mail network model.
2. Derive neighbourhood features using the e-mail network model.
3. Combine the neighbourhood features with the user data.
4. Run experiments in WEKA with the prepared data sets.
5. Evaluate results of the experiments.

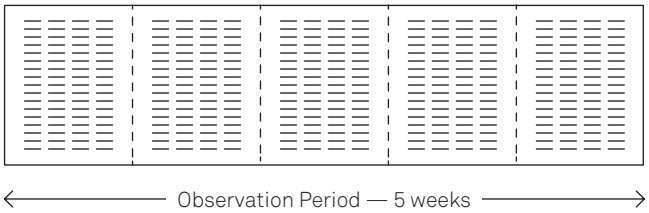
DATA SET CONSTRUCTION

To construct the data set for the predictive model we started with the cleansed transactions data set. We split the data set into two time periods. A predictor period, and an outcome period. Since the entire data set had data for 5 weeks of usage, we split between the first 3 weeks for the predictor period and the final 2 weeks for the outcome period. With the predictive model we'd like to predict the behaviour of the users from the predictor *in the outcome period*. In our case, the want to predict whether the users in the predictor period will become a subscriber in the outcome period. The primary benefit for becoming a subscriber is that it allows a user to send larger files, as free users can only send files under the maximum of a predefined file size.

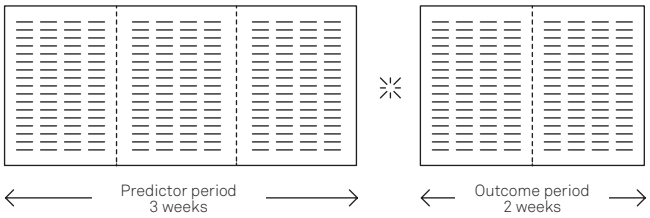
However, here we ran into a problem with the transaction data. The user type was recorded at the transaction level, meaning that it could vary during the observation period, for example on wether a user was logged in as a subscriber or not. Due to technical issues during the observation period, we found that subscribers were not necessarily automatically logged in to the service if they were sending files that were under the file size limit set for the free users. To resolve this ambiguity we wrote an algorithm

CREATING THE DATASET FOR THE PREDICTIVE MODEL

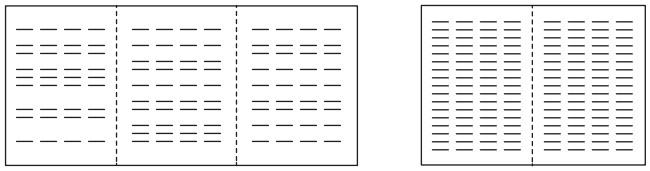
1. We take the dataset for the entire observation period,



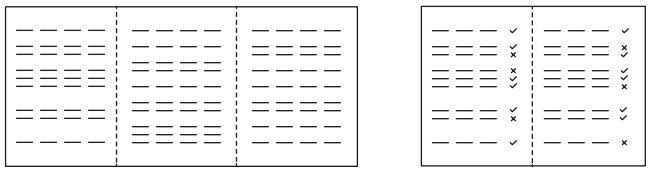
2. And split it into two periods, a predictor period and an outcome period.



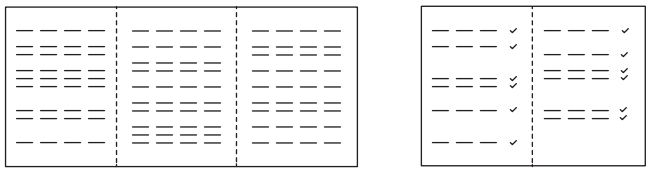
3. We select all free users in the predictor period,



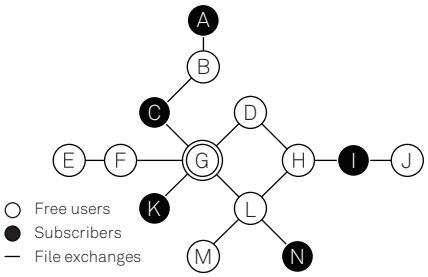
4. And create a flag to see if they've become subscribers in the outcome period.



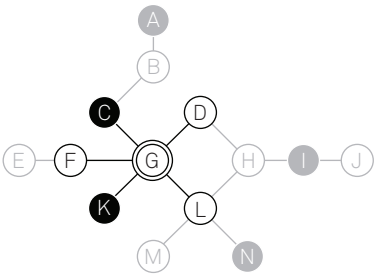
5. We select all users that have become a subscriber.



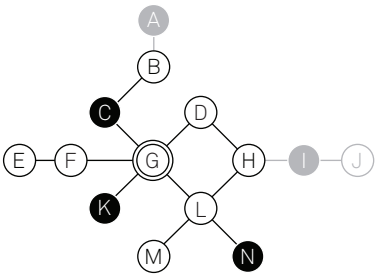
DERIVING NEIGHBOURHOOD FEATURES FOR A NODE IN A GRAPH



In this example graph we have 13 nodes. Out of these 13 there are 5 subscribers and 8 free users. We're interested in the neighbourhood properties of node G.



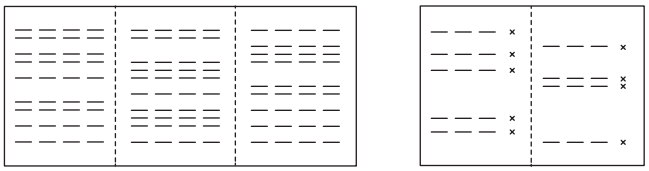
The first degree neighbours of node G are nodes C, D, F, K and L. So, node G has 2 neighbours that are a subscriber.



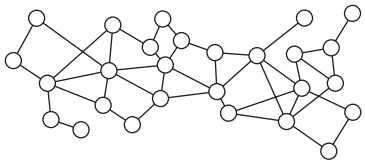
The second degree neighbours of node G are the neighbours of node C, D, F, K and L, being nodes B, E, H, M and N. So there's one second degree neighbour that's a subscriber (node N)

CREATING THE DATASET FOR THE PREDICTIVE MODEL

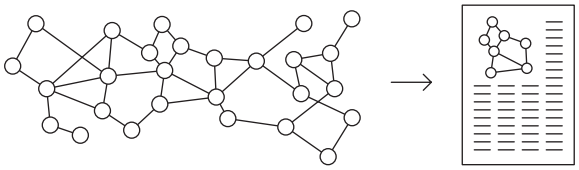
6. Select an equaly sized random sample of users that have not subscribed during the outcome period.



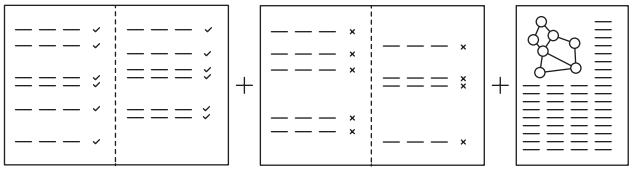
7.Create the network of e-mail adresses as it is at the end of the predictor period.



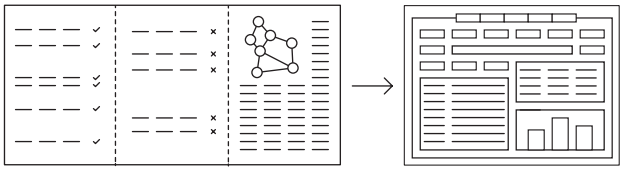
8. Derive the network attributes for the selected individual users.



9. Combine the data from the users that became subscribers, with the non-subscribers, and their network attributes.



10. This combined dataset forms the input for the WEKA datamining workbench.



to remove all the users for whom the `user_type` property was ambiguous (so not either a free user *or* a subscriber).

We proceeded to select all free users in the predictor period, and created a flag that determined if they became a subscriber in the outcome period. We then selected all users for whom this flag was set to 'yes'. This resulted in a selection of 2.252 users. Comparing this number to the total number of users, we elected to selected an equally sized sample of non-subscribing users, as the number of subscribers we could select was relatively low compared to the total number of users. So finally we selected 4.504 users.

We then recreated the e-mail address network but only for the duration of the predictor period. That way we could derive network and neighbourhood properties based on the state of the network as it existed at the end of the predictor period.

DERIVING NEIGHBOURHOOD FEATURES

Aside from the regular network features, we also wanted to derive a number of neighbourhood features, so properties of users that our users interacted with through the service.

We wanted to evaluate wether having subscribers in the local neighbourhood of a user had any influence on their conversion to a paid user. To evaluate this we had to examine the properties of neighbouring nodes of or chosen users. We specifically examined if a user had neighbouring users who were also a subscriber. We hypothesised that having a larger number of subscribers in a users neighbourhood in the predictor period could influence the user in also becoming a subscriber in the outcome period. For every user we selected we retried their first degree and second degree neighbours (neighbours of neighbours).

We created two versions of our user data set. One *with* neighbourhoods and one without. Table 10 shows the variables contained in the final data sets.

With these data sets we were now ready to create our predictive models using WEKA.

VARIABLE	DESCRIPTION
`in_degree`	the number of exchanges received by the user
`out_degree`	the number of exchanges sent by the user
`smallest_file_exchange_size`	the file size in bytes of the smallest file the user sent
`average_file_exchange_size`	the average file size in bytes of the files the user sent
`largest_file_exchange_size`	the filesize of the largest file the user exchanged
`domain_type`	the domain type of the user (free / private)
`n_free`	number of neighbours that are free users
`n_subscriber`	number of neighbours that are subscribers
`nn_free`	number of neighbours of neighbours that are free users
`nn_subscriber`	number of neighbours of neighbours that are subscribers
`became_subscriber`	classification that we want to predict; wether the user became a subscriber.

Table 10. Variables for the predictive model.

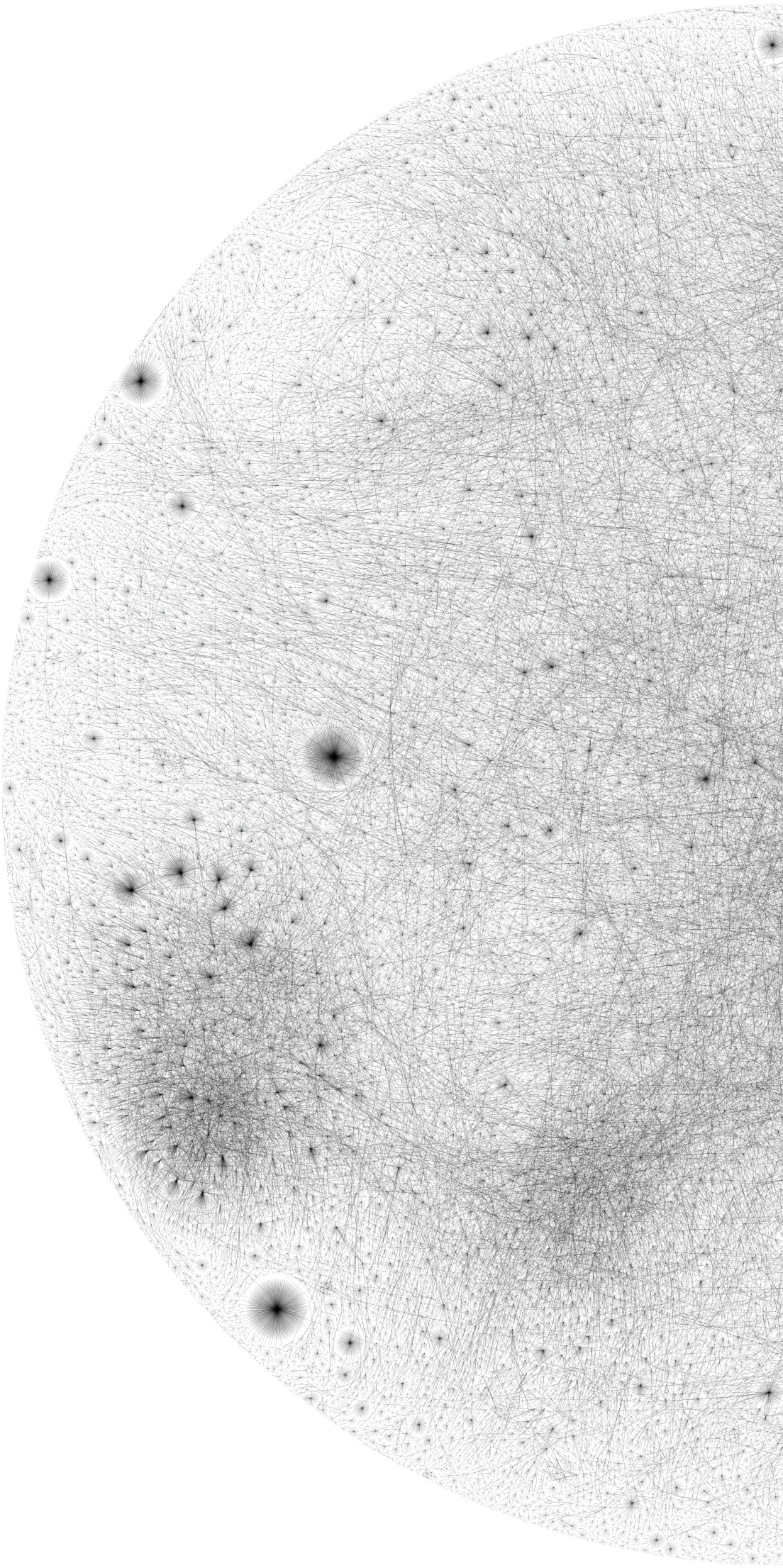
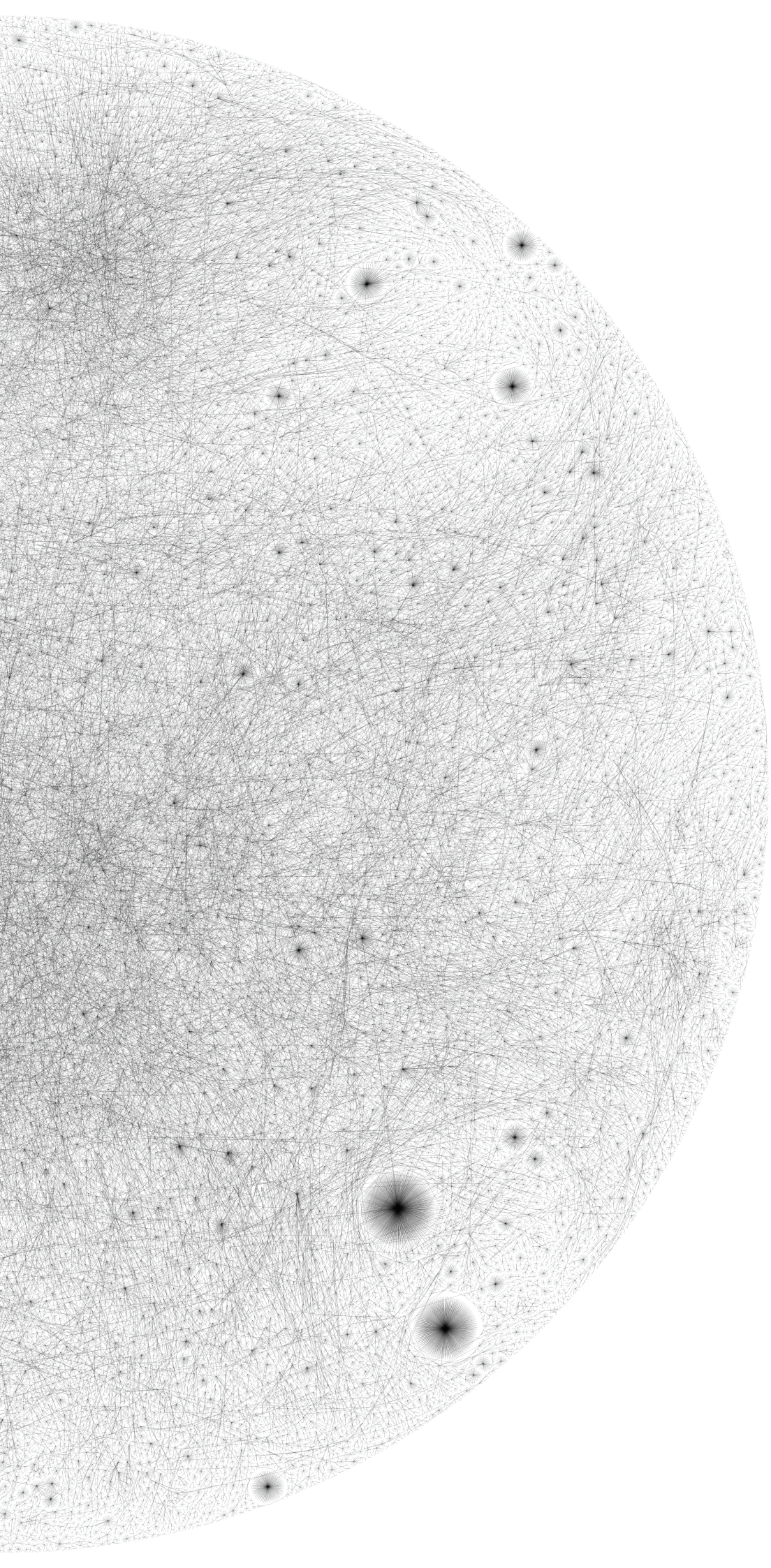


Figure 10.
Country B
Weakly Connected Component
01/01/2017 — 03/02/2017

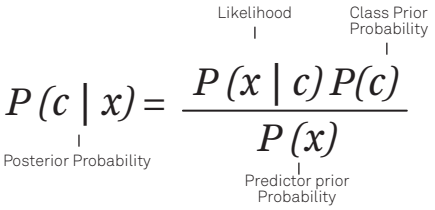
270.869 nodes,
312.542 edges



EXPERIMENTAL SETUP

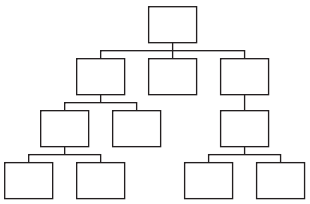
To train our predictive models we used the WEKA toolkit. This is an open source machine learning environment that makes it possible to use a graphical user interface for data analysis and predictive modelling. It also provides basic functionality for preprocessing a data set before training. It can be used for exploration as well as experimentation. We set up an experiment to train two predictive models using our data sets, one with neighbourhood features, and one without. We then selected the following algorithms to train our models:

NAIVE BAYES



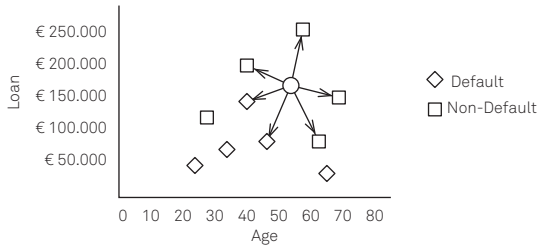
The Naive Bayesian classifier is based on Bayes' theorem, with independence assumptions between variables. Bayes' theorem provides a way of calculating the posterior probability. A Naive Bayes classifier assumes that the effect of a predictor (x) on a given class (c) is independent of the values of other predictors (Sayad, 2017).

DECISION TREES (J48)



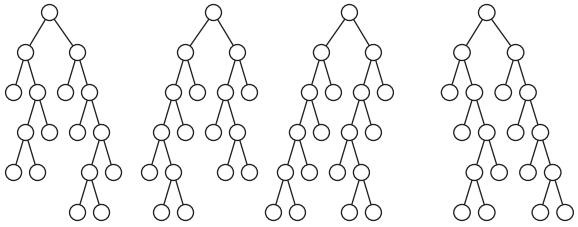
The J48 algorithm is WEKA's implementation of the popular C4.5 decision tree algorithm (Quinlan, 1993). It splits a dataset down into smaller and smaller subsets that best seperates the classes, while simultaneously the associated decision tree is produced. This decision tree can then be navigated in order to determine the classification of an observation.

NEAREST NEIGHBOUR (LAZY IBK)



The LazyIBK algorithm is a K-nearest neighbour classifier (Aha, 1991), and classifies new observations based on a similarity measure (often a distance function). It's a technique that has been in use for statistical estimation and pattern recognition since the 1960s.

RANDOMFOREST



RandomForest is an example of an ensemble learner. It is made up out of multiple decision trees. The basic principle is that a group of "weak learners" can come together to form a "strong learner" (Breiman & Cutler, 2001). It introduces randomness into the learning algorithms' input to create varied trees, that when taken together produce better predictions than a single tree.

In order to compare these models we used a ZeroR classifier to create a baseline. This algorithm ignores all predictor variables and just returns the most frequent class that is in the training set. Hence, in our 50/50 data set it should return a result of 50%. The diagram above gives a brief overview of the different machine learning algorithms and their tradeoffs. Each of these algorithms uses a different approach in order to build a predictive model. We configured WEKA to run 10 experiments for every algorithm, using 10-fold cross validation in order to yield an overall estimation error. Cross-fold validation is a technique where the data set that is used to perform the machine learning on is divided into n-sized parts, after which each single part is held out in turn, and the model is trained on the remaining n-parts. The final remaining part is then used to check the error-rate of the trained model. In order to get a valid estimation of the error rate, the created model should be tested on data that the model was not previously exposed to. Cross-fold validation ensures that we get an reliable error rate for the trained models. We then performed a significance test of the different algorithms compared to the baseline. For this we used a corrected Paired T-Tester.

3.4 RESULTS

In this section we report on the results from our modelling step, focussing on the information gathered from the three different network models, as well as the experimental results from the training of the two predictive models.

3.4.1 NETWORK MODELS

We created three different network models, each at different observational magnitudes. The country network examines the service's usage at an international level, with a network that numbers in the hundreds of nodes and edges. The domain network looks at the usage of the service across internet domains, numbering in the hundreds of thousands and finally the e-mail network examines the usage of the service down to the individual email address, which is in the millions.

COUNTRY NETWORK

The country network models the flow of file exchanges within and between countries. Since our dataset contained only file exchanges originating in two countries, we could mostly only observe one directional traffic, except of course for traffic between country A and country B. We created three versions of this model, using data from both countries, using data from only country A and using data from only country B.

When we started examining the network properties we found that both countries communicated with 220 other countries, but there is some variation in which countries, as the network from the combined countries has 227 nodes, see table 11. Considering that the ISO 3166 standard for country codes defines 248 different countries, both countries contain users that collectively exchange digital files with most of the rest of the world. Since there were only two originating countries in the dataset, we found it to be more useful to talk about file exchanges flowing out of these countries, than the whole network. We examined the amount of file exchanges flowing inside and between countries and found in both cases that slightly over 80% of all exchanges happen within country borders, while slightly less than 20% of the file exchanges move across country borders (table 12). We suppose that we might be seeing the Pareto-principle (80/20 rule) in action, but as we have data originating from only two countries we can not draw a hard conclusion from this observation.

For presentation purposes we visualised the flow of file exchanges from the two countries and their top 20 destinations. For this we elected to use a Sankey diagram, as they are well suited for visualising flows (see figures 11 and 12) .

NETWORK	NODE COUNT	EDGE COUNT
all	227	440
Country A	220	220
Country B	220	220

Table 11. Network properties of the country network.

	NATIONAL	INTERNATIONAL
Country A	80,77%	19,23%
Country B	81,91%	18,09%

Table 12. Percentage of file exchanges happening nationally and internationally in country A and B.

DOMAIN NETWORK

The domain network models the file exchanges between different domains associated with e-mail addresses. As domains are often tied to organisations that individual users are affiliated with, the domain networks can shed some light on usage of the service within an organisational context. The basic network properties of this network can be observed in table 13. Of course, this is difficult for accounts associated with free e-mail providers. Table 14 shows an overview of the basic network measures (Easly & Kleinberg, 2010). What's interesting to observe is the higher density of the network in country A, being 12% more dense that country B. We speculate that this might have something to do with the amount of time the service has been active in both countries. This could be corroborated by the difference in assortativity between both countries.

Finally when we extracted the largest connected component we found that in country A it contains 90,7% of all domains, whereas the largest connected component in country B contains only 84,6% of all nodes. The combined network holds 88,6% of the nodes. This is likely due to the fact that Country B is country A's second highest destination country, and the opposite holds true for country B. Table 15 shows the node counts for the different largest components.

NETWORK	NODE COUNT	EDGE COUNT
all	589.240	1.103.748
Country A	315.180	650.577
Country B	298.957	463.187

Table 13. Network properties of the domain network.

NETWORK	ALL	A	B
DENSITY	0.318^{-5}	0.655^{-5}	0.518^{-5}
DEGREE ASSORATIVITY	0.061	0.058	0.066
IN DEGREE ASSORTATIVITY	0.057	0.055	0.056
OUT DEGREE ASSORTATIVITY	0.058	0.057	0.065

Table 14. Network measures of the domain network.

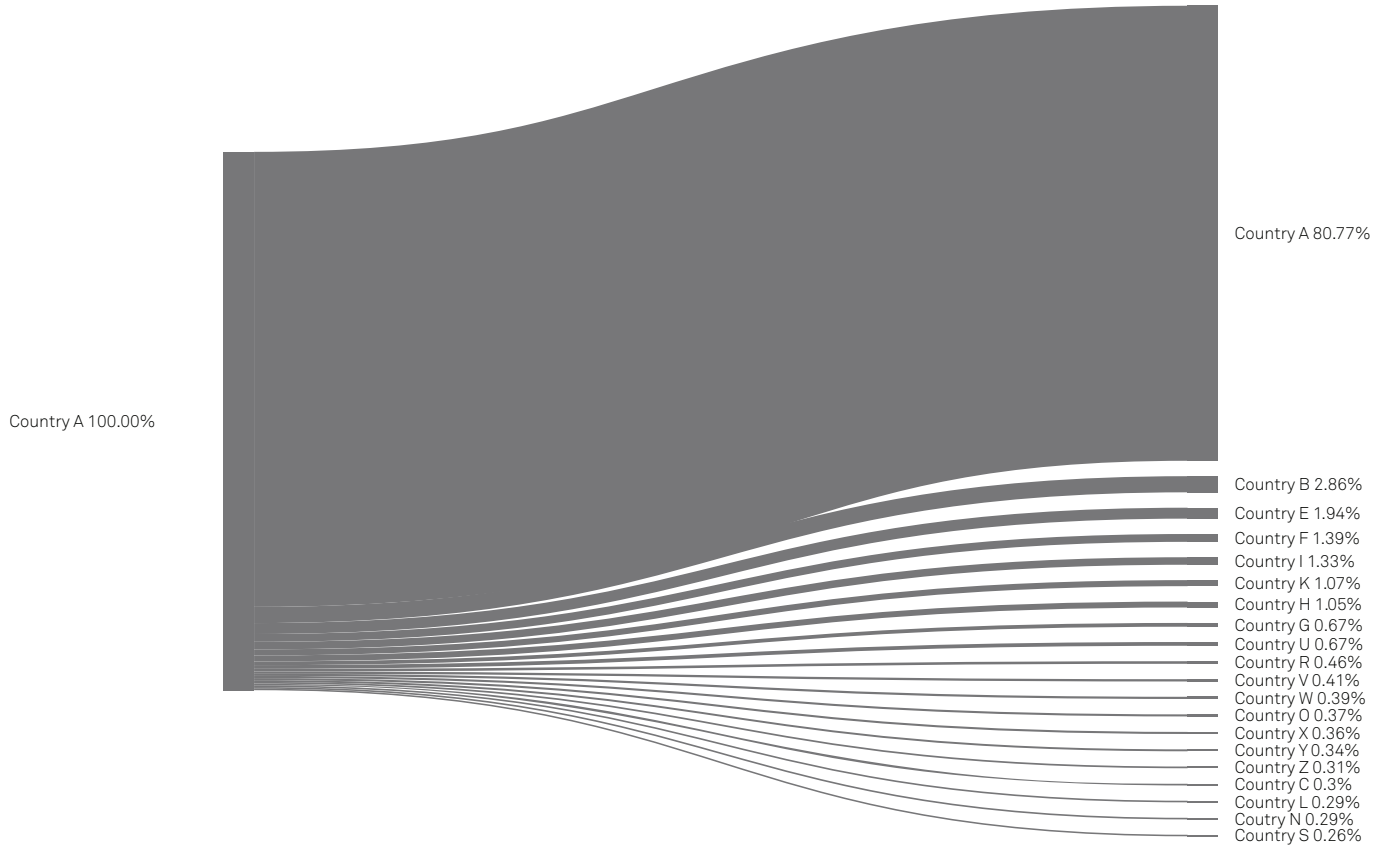


Figure 11. Sankey diagram for file exchange flows in Country A. Country names anonimized for confidentiality

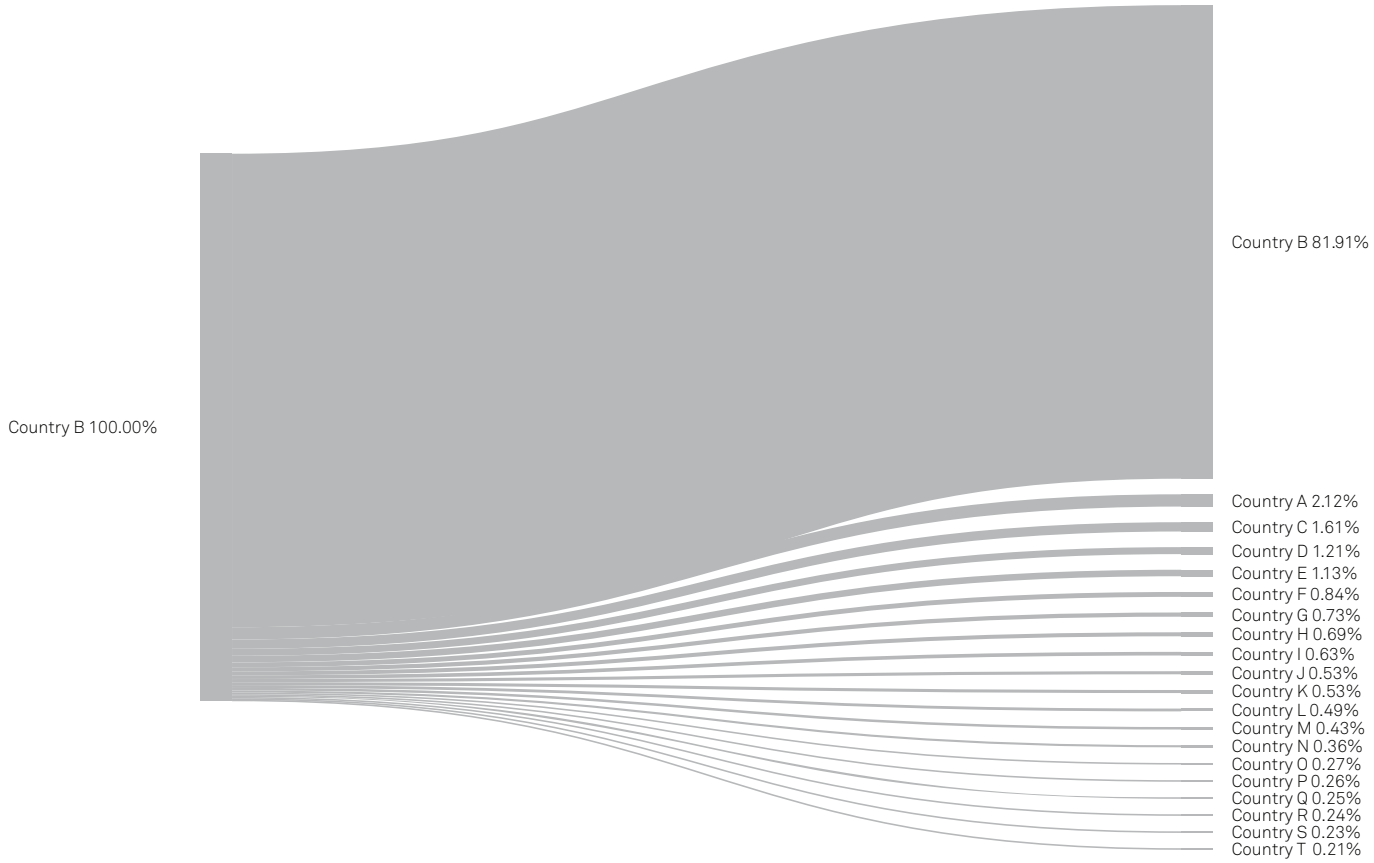


Figure 12. Sankey diagram for file exchange flows in Country B. Country names anonimized for confidentiality

NETWORK	NODES IN LCC	%
all	522.221	88,6 %
Country A	285.973	90,7 %
Country B	252.918	84,6 %

Table 15. Node counts for the largest connected components in the domain networks.

During the presentation of the results to the participating organisation we explored the makeup of the different organisations in the networks in more detail, and we found that while the organisation has a very specific idea around the target user groups that make up the usage of the service, there are enough examples of organisations using the service falling outside of this definition, suggesting that a broader definition of the target users of the service is in order. We explored the different forms of usage we found in detail with the E-mail network.

E-MAIL NETWORK

The e-mail network maps the interactions between the individual e-mail addresses of users. While the service allows users to specify multiple e-mail addresses as a source, for this model we assumed that each unique e-mail address mapped to a unique user. This network model can provide insights at the most granular level. Examining the basic properties (see table 16) of the network we find that the number of users in both countries are quite equally sized, with country A making up 51,5% of the users in the network, while country B holds around 49,5%. As with the domain network the network in country A is a little bit denser than country B, but not to the same amount as we see in the domain network (see table 17).

NETWORK	NODE COUNT	EDGE COUNT
all	2.385.469	2.180.019
Country A	1.228.509	1.179.856
Country B	1.183.169	1.001.860

Table 16. Network properties of the e-mail network.

When we examined the degree distribution, we learned that in both countries close to 60% of the users actually only interact with the service as a receiver (as those nodes have an out degree of 0) and that 30% of the users use the service only to send a file exchange, but never have received anything throughout the observation period. Tables 18-23 show the distribution for both the in and out degree between 0 and 4, as these degree values are true for around with around 99% of the nodes. The full degree distribution diagrams are available in Appendix B. Aside from the anomalously high value for the out degree of 0, the degree distributions found in the different network models hew closely to distributions found in most real-world networks.

NETWORK	ALL	A	B
DENSITY	0.38^{-6}	0.78^{-6}	0.72^{-6}
DEGREE ASSORATIVITY	0.048	0.050	0.043
IN DEGREE ASSORTATIVITY	0.023	0.023	0.019
OUT DEGREE ASSORTATIVITY	0.031	0.029	0.033

Table 17. Network measures of the e-mail networks.

ALL			COUNTRY A			COUNTRY B		
IN DEGREE	FREQUENCY	%	IN DEGREE	FREQUENCY	%	IN DEGREE	FREQUENCY	%
0	678.022	28,4%	0	341.056	27,7 %	0	345.119	29,1 %
1	1.441.657	60,4 %	1	726.887	59,1 %	1	739.796	62,5 %
2	178.264	7,4 %	2	103.977	8,4 %	2	70.924	5,9 %
3	48.575	2 %	3	30.649	2,4 %	3	16.162	1,3 %
4	18.443	0,7 %	4	12.259	0,9 %	4	5.453	0,4 %
OUT DEGREE	FREQUENCY	%	OUT DEGREE	FREQUENCY	%	OUT DEGREE	FREQUENCY	%
0	1.377.056	57,7 %	0	697.554	56,7%	0	702.762	59,3 %
1	592.050	24,8 %	1	304.785	24,8%	1	290.048	24,5 %
2	187.377	7,8 %	2	100.199	8,1%	2	87.485	7,3 %
3	88.192	3,6 %	3	47.557	3,8%	3	40.672	3,4 %
4	48.586	2,0 %	4	26.580	2,1%	4	22.020	1,8 %

Table 18-23. Top 5 of the in and out degree distributions of the respective networks.

Next we examined the component distribution. In our case we specifically looked at the weakly connected components, as it is not necessary to both send and receive files using the service. The component distribution shows the makeup of the network, and the extent to which the entire network is connected. We were surprised to find that in all three network models the amount of nodes found in the largest connected component (The Giant components) did not meet our expectations. In networks associated with human phenomena often, the largest connected component holds the majority of the nodes, however in our network models we found that less than 50% of the nodes were in the giant components of the respective networks. We speculate that there could be several factors contributing to these (relatively) small giant compo-

nent sizes. As we only had a sample containing usage data from two countries, it could very well be that if we were to model the network for the observation period using all countries, the giant components could hold a large percentage of the nodes. The fact that the combined network model (ALL), has a giant component that in terms of number of nodes is larger than the combination of the giant components from both Country A and Country B (see table 24), points to international usage that could provide the 'connective tissue'. It does bring into question to what extent the usage of the service is delimited by borders. Another reason could be that our observation period was too short in order to observe all connections within the network.

Next, we more closely examined the components. For presentation purposes We visualised the Giant Component of Country B (see figure 10 on page 20/21), using the GPUGraphLayout software on Leiden University's Data Science cluster. For these purposes we used the anonimised version of the dataset we created.

Within the visualisation we found numerous denser structures, which we further explored using Gephi. As Gephi is poorly suited to visualising large networks, we focussed on smaller components, and extracts from the giant component of country A. Within these extracts we found that users tended to cluster based on industry. Using the domain names for each node we found that usage tends to be delimited by industry. Within these sections of industry we also found that depending on the industry, other node properties may vary. For example, figure 13 shows an extract of the network in country A. Nodes in this extract belong to domains that are strongly associated with the publishing industry in that country. Here we see a strong cluster of users with private domains (black nodes) and few domains associated with free e-mail providers (white nodes). Comparing this to figure 14, that shows an extract from the same network with nodes with associations with the music industry in that country, where we see many users with domains from free e-mail providers (white nodes) interacting with a few nodes from private organisations (black nodes).

For each network model we extracted the next 50 largest components, and calculated some network measures. Based on these measures we selected several components for visualising different usage scenario's of the service. In figure 15 for example, we see usage of the service across different companies within the fashion industry. Figures 16 and 17 show usage of the service in education related scenario's. In figure 16 we see one node interacting with many other free-email addresses, and the private domains in the graph all belong to domains associated with higher educational institutes. Figure 17 also shows usage, but in this case it's many free-email addresses with a single private domain. On closer inspection we found that the private e-mail address was used for study grant application purposes by an energy company. Figure 18 shows a component extracted from the network in country A that captures the usage within a single company and its associates. The white nodes in the network all belong to the

NETWORK	NODES IN LCC	%
all	876.590	36,7 %
Country A	570.425	46,4 %
Country B	270.869	22,8 %

Table 24. Node counts for the largest connected components in the e-mail networks.

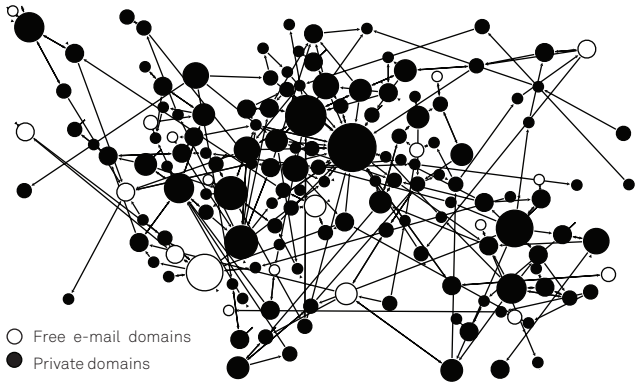


Figure 13. Extract from the network in country A. This shows the usage of users associated with the publishing industry.

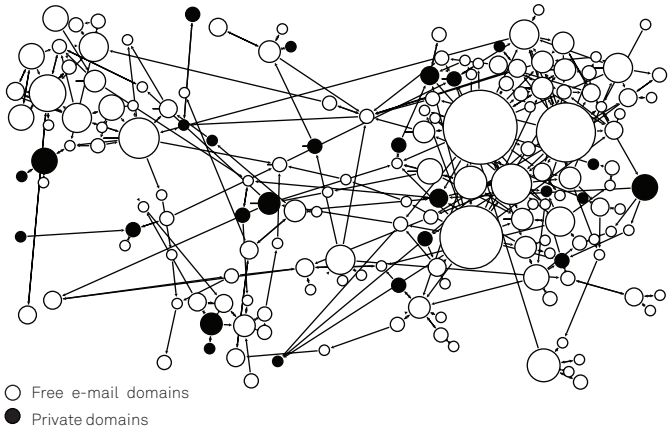


Figure 14. Extract from the network in country A. This shows the usage of the service within country A's music industry.

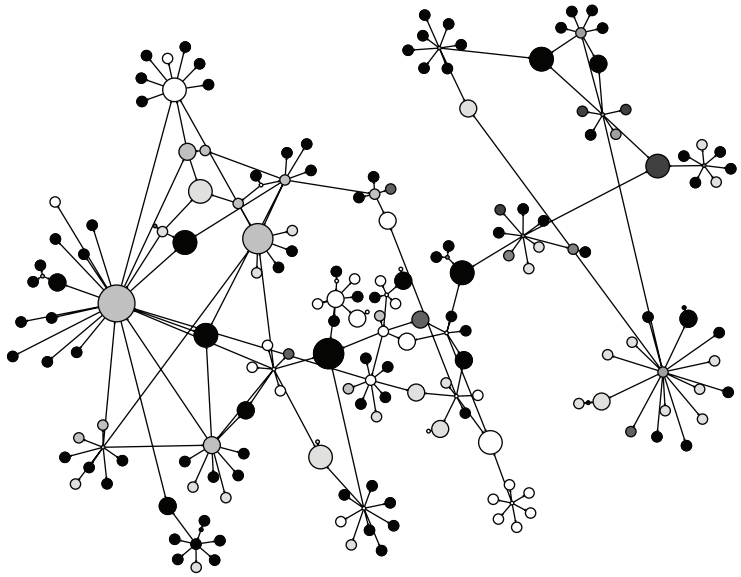


Figure 15. Component from the network in country A. This network shows the usage of the service within a set of companies in the fashion industry.

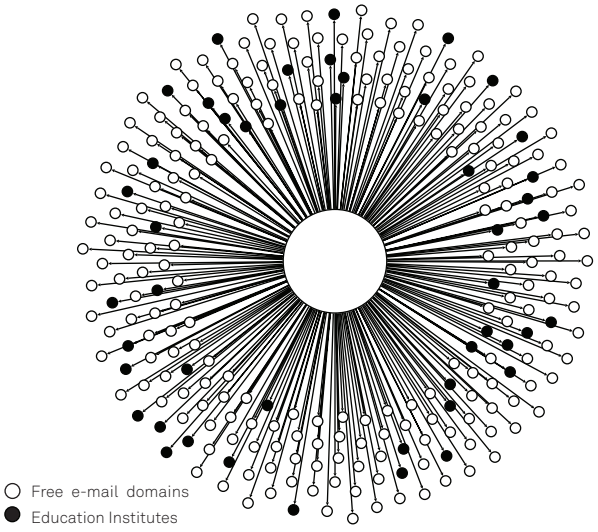


Figure 16. Component from the network in country A. The central node is an free e-mail adress, interacting with users in educational institutions and other free e-mail adresses.

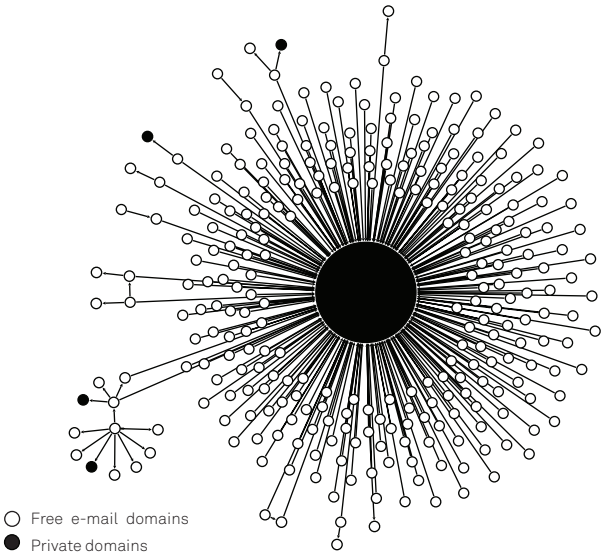


Figure 17. Component from the network in country A. The central node is an e-mail adress used for grant-applications at an international oil company.

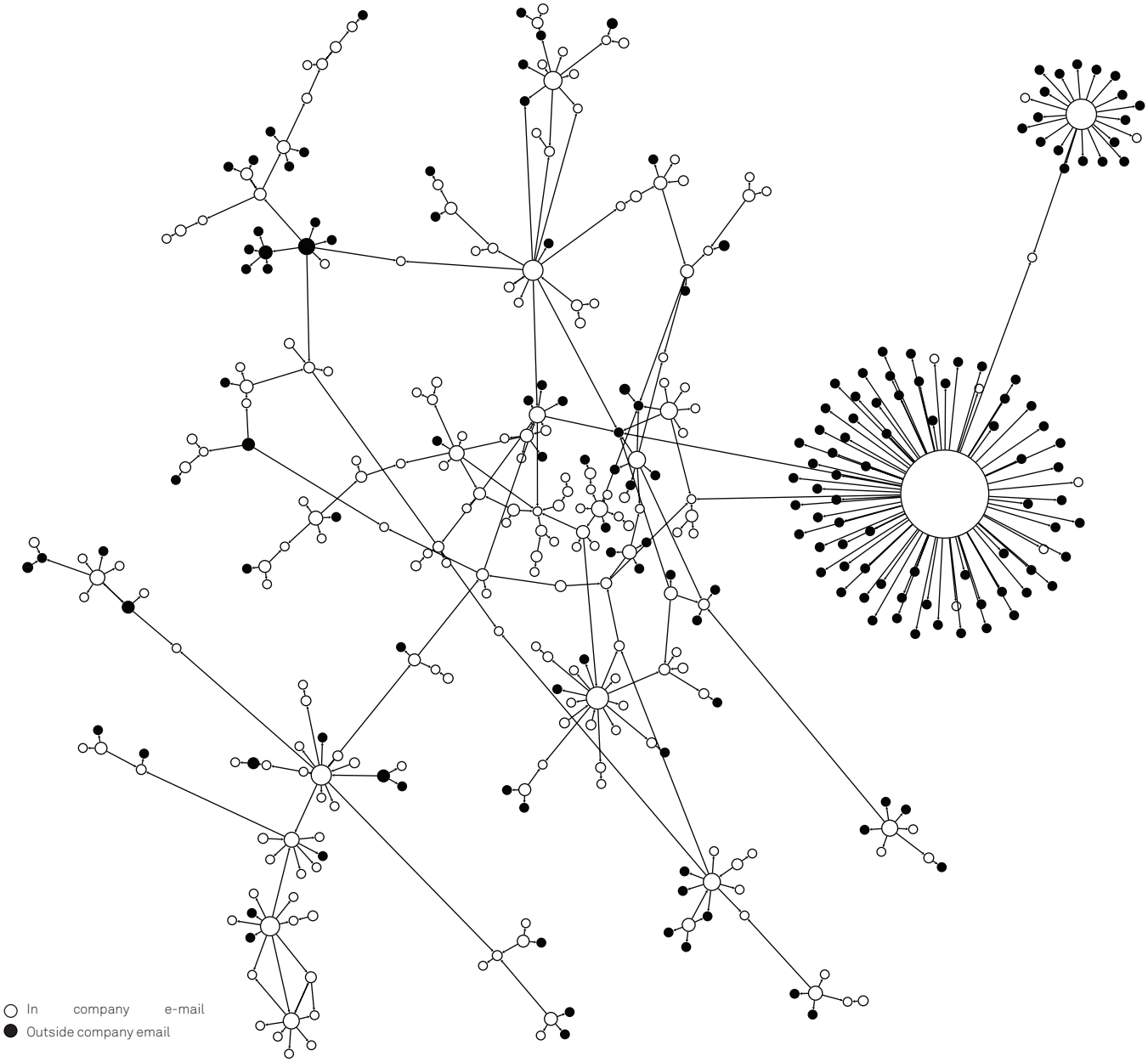


Figure 18. Component from the network in country A. It shows the usage of the service within an investment company for sustainable energy. White nodes represent in-company use, and black nodes destinations outisde the company.

same organisation, black nodes represent users outside of that specific company.

In general we found that the different network models provided a glimpse into the usage of the service as it had not previously conceptualised.

3.4.2 PREDICTIVE MODELS

We trained two sets of classification models. The goal for these classification models was to predict the conversion from a free user to a subscriber. One set of models was trained on a dataset that included attributes that were derived from the network of the users in the dataset (the ALL dataset) and the other set was trained on a dataset from which these network variables were withheld (the NO NETWORK set). To asses the quality of the classifiers we trained we need to look at the percentage of correctly classified instances (see table 25) for both datasets. This was tested with a (corrected) Paired T-Tester, and a significance level of 0.05. As expected, the ZeroR classifier just returns the classification and since we built our dataset with an equal amount of examples of users that convert and users that do not convert, it returns an error rate of 49.95%. This provides us with a baseline to compare the different algorithms. What is notable here is the difference in performance between the tree-based models (J48 and RandomForrest) compared to the other algorithms in the set (NaiveBayes and Instanced based k), which both seem to perform better on the set without network variables, whereas the tree-based algorithms perform slightly better on the dataset with network variables. This could be due to the fact that tree-based algorithms are less sensitive to the usage of more (correlated) predictors, as they perform implicit data selection.

Next, we need to evaluate the quality of the ROC (Receiver Operating Characteristic) Curves (see figure 19). These curves allow one to assess the quality of a binary classifier (which our models are, either you become a subscriber, or you don't), by creating a graph of True Positives (TP) rate, versus False Positives (FP) rate for every classification threshold. The True Positive rate is the number of predicted positives that are indeed positive expressed as a percentage of the total number of positives. The False Positive rate is the number of predicted positives that are actually negatives expressed as a percentage of the total number of negatives. A ROC curve that 'hugs' the upper left corner of the graph is considered a 'good' classifier, whereas a classifier that is below the x = y line, performs worse than random guessing (which is what the x=y line represents). Based on the ROC curve, the *Area under the Curve* or AOC value can be calculated, which quantifies the performance of the classifier. Examining the results in table 26, we find that the RandomForest classifier performs best when looking at this measure, with an AOC value of 0.83 for the dataset with network attributes, and 0.81 for the dataset without network attributes.

Finally we examined the *information gain* of each attribute (Witten et al, 2016). Information gain is often a good measure for indicating the relevance of an attribute, and is indicated with a value between 0 and 1. We found that

DATASET	ALL		NO NETWORK	
ZeroR	49,95%	(0,05)	49,95%	(0,05)
J48 Decision Tree	*78,08%*	*(1,74)*	*78,04%*	*(1,76)*
Naive Bayes	63,05%	(2,46)	68,22%	(3,31)
Lazy IBk	69,97%	(2,13)	70,11%	(2,17)
Random Forest	77,03%	(1,84)	76,22%	(1,87)

Table 25. Percentage correctly predicted results for the different algorithms and datasets.

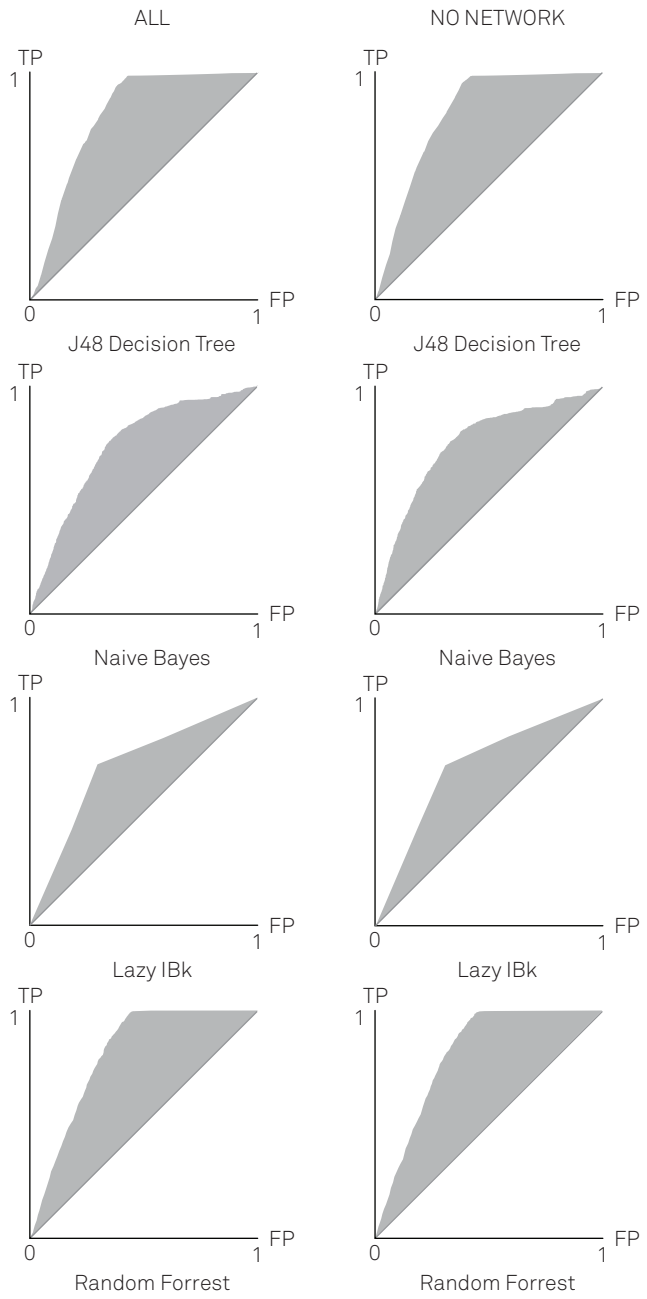


Figure 19. ROC curves for both datasets, showing that the RandomTrees classifier performs most accurately.

DATASET	ALL		NO NETWORK	
ZeroR	0.50	(0,00)	0.50	(0,00)
J48 Decision Tree	0.78	(0,02)	0.79	(0,02)
Naive Bayes	0.75	(0,02)	0.76	(0,02)
Lazy IBk	0.70	(0,02)	0.70	(0,02)
Random Forrest	*0.83*	*(0,02)*	*0.81*	*(0,02)*

Table 26. Area under Curve (AUC) results.

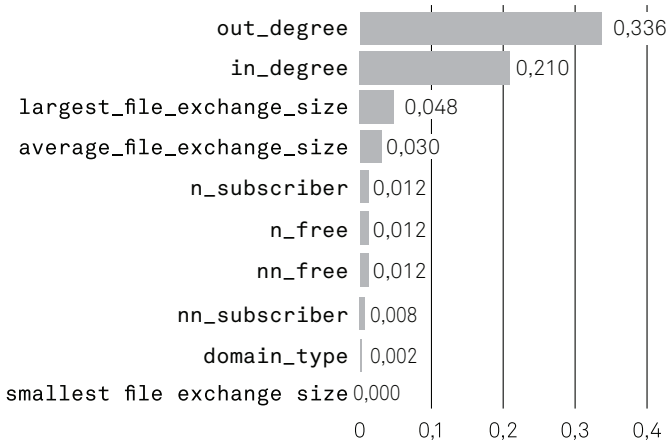


Figure 20. Information gain for each attribute in the network dataset.

the `in_degree` and `out_degree` attributes showed the highest information gain (0.210 and 0.336 respectively, see figure 20). In our model these user attributes represented the number of times file exchanges were sent and received during the predictor period. From this we may infer that when it comes the conversion of a free user to a subscriber, the frequency of use is most strongly predictive for the conversion of a user. What is interesting here is that the information gain chart provides the basis for a discussion around the business model of the service. Based on the results from this model the proposition for subscribing to the service could be adjusted, emphasising frequency of use, as opposed to for example file size.

We also found that, while the dataset with network attributes performs statistically significantly better, their contribution to the final classification is relatively small, with the results from the second degree neighbours equally sized or even 33% smaller (0,012 and 0,008 respectively).

4. COMMUNICATING RESULTS OF THE CASE STUDY

This section reports on the study and workshop on the results of the research described in the previous section. To evaluate the results and outcome of the analyses, a workshop was held with the participating company. The goal of this workshop was to evaluate the findings, but also to study the reactions of the participants to the research methodology and its potential applications. We first describe the setup and procedure and finally reflect on the reactions to the research results and workshop. This evaluation reflects on the context, purpose and effect of the research results and how it affected the various team members and knowledge and interaction with data-driven research and how it enhanced their thinking around the usage of the service.

4.1 METHOD

The workshop was held at the offices of the participating company and included 10 employees of three different sections of the company. Business, technology, and design. Video and audio was recorded during the workshop. The recordings of the discussions during the workshop were partly transcribed and analysed. The workshop took place at the offices of the company and included ten participants besides the primary researcher. At the start of the workshop the participants were given an open questionnaire on the topics of the case study. After the presentation a follow-up questionnaire was filled out to evaluate how their attitudes to data-driven research had changed.

The goal of the first questionnaire was to establish a baseline with regards to the topics that would be discussed during the presentation of the results. Each participant was asked to respond to four open questions with regards to the presentation: Their general expectations, Whether it was possible all of their users were connected in some way; What they would find most interesting if the could predict user behaviour and finally if they had any questions up front regarding network science, data mining, and machine learning.

We started by explaining the data processing steps taken create our dataset. Some of results from these steps already sparked a discussion with the participants. Having explained how we arrived at our final dataset we moved towards discussing the network models. We first presented the hypothesis that while common usage of the service is transient, it was likely that the social interactions that people have influence the interactions that are facilitated by the service. We then showed how one could think of the service as a complex system, and that, while the service does not present itself to the users as a network, it should be possible to observe a network, based on the data of the interactions between users. We elaborated on the scientific definition of complexity and how studying complex systems by studying their constituent parts, can not necessarily predict the behaviour of the whole system. To explain the differences between a complex system and a complicated system, we used the examples of a flock of birds and

a Rube Goldberg machine. The collective behaviour of the bird flock is difficult to explain and predict from studying a single bird, whereas the behaviour of a Rube Goldberg machine can be explained and predicted by breaking it down into its constituent parts and describing the causality between the parts. We explained how the field of Network Science has arisen during the past twenty years to facilitate the study of complex systems, and the kind of answers it could provide to questions regarding the usage of the service. We described how a network graph could be constructed from the transaction data, and could be used to create three different network models. We then described the properties and visualisations of the various networks and their implications for the usage of the service.

Before discussing the results of the predictive models, we first explained the basic concepts of data mining and machine learning. Our goal was to demystify the subject, and take away any hesitations that the participants may have had, and ensure that all members of the workshop were at a comparable level of understanding in order to interpret the results from the case study. We acknowledged that data mining and machine learning have a reputation for being dense and complex subjects, but that we would familiarise everyone to a degree of general comprehension. In order to do so, we used an idiosyncratic framing of the explanation of the basic concepts: *'Datamining and machine learning for Cocktail Parties'*. This frame was borrowed from the Dutch scientist Maarten Lamers, who uses this framing in order to open up Artificial Intelligence research to a popular audience. The general thought behind this was that if the introduction into the subjects of datamining and machine learning is successful, workshop participants would be able to converse with datamining and machine learning researcher about their work at a social gathering, such as a cocktail party. By reframing these explanations in such a way, we created an opportunity for all members of the organisation to approach the theory with an equal level of comprehension, regardless of background. In our explanation we emphasised the transparent box approach to machine learning, in which an algorithm produces a structure that can be studied, and could provide better insight into which variables were predictive for the conversion of a free user to a subscriber. We described how we created the datasets for the predictive models and discussed the different results from the various machine learning algorithms that we used and how strongly predictive the variables within the datasets were. The results from these models led into a broader discussion about what was learned during the workshop and what the research results implicated for what was known by the employees about the usage of their service.

We concluded the workshop with a final questionnaire in which we assessed how the participants' understanding of data-driven research was changed. We inquired what they were surprised to learn, how establishing that there is a network 'behind' the service changed their view of the service, how they could apply data mining and predictive modelling to their daily work now that they've seen it was successful in predicting the conversion from free users to subscribers, and finally any general observations that they had after completing the workshop.

4.2 OBSERVATIONS

In the baseline questionnaire we found that the employees of the company had different expectations from the workshop. The employees within the design department mentioned their unfamiliarity with data-driven research yet expressed optimism that the workshop might provide them with new insights with regards to the usage of the service, as well as familiarising them with subjects such as data-mining, machine learning and network analysis. The familiarisation observation was also shared with business analysts, who had heard of these types of research, but were unfamiliar with them, and as such unsure of how to apply them to the specific data that was generated by the service.

When it came to our hypothesis that the social organisation of users influenced the usage of the service in some way, we found that several members of the workshop made references to the popular theory of *'Six Degrees of Separation'*, or variations thereof. Other members referred to this principle through their experience with social networks, going so far as to actually sketch small network diagrams. Figure 21 shows a selection of responses from the workshop participants to the question whether it could be possible that there's a network organisation behind the usage of the service.

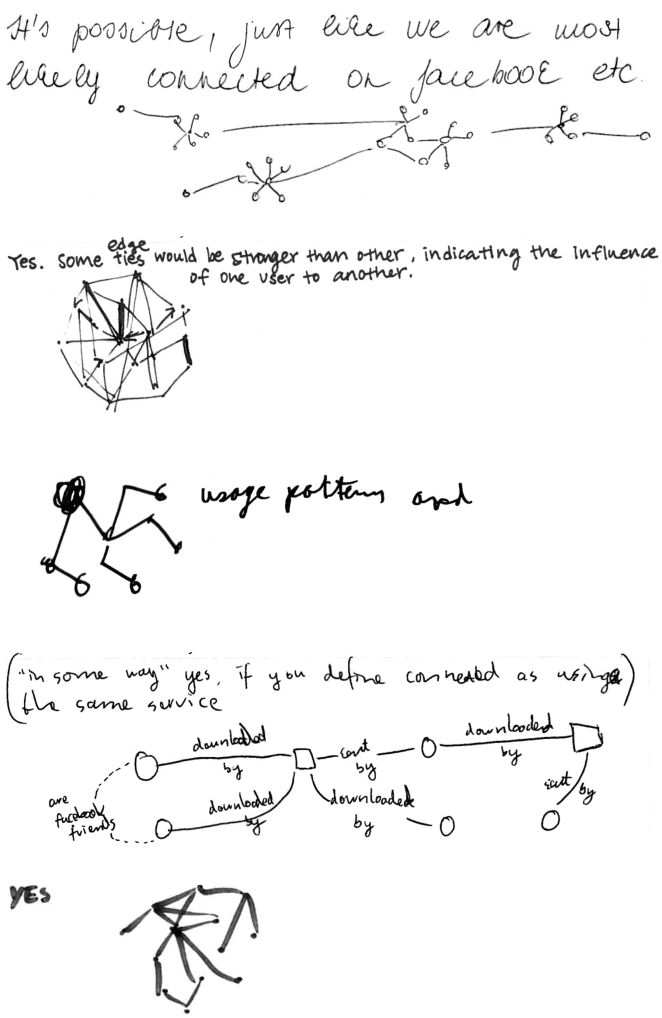


Figure 21. Responses from the workshop participants to the question whether it would be possible that there's a network organisation behind the usage of the service.

Regarding the possibilities of predicting user behaviour we found that users who were more familiar with predictive modelling and statistical analysis were able to more specifically formulate what kind of user behaviour they wanted to predict. Participants who indicated being less familiar with these techniques in most cases also expressed statements that were hard to predict based on the data that the company was collecting from their users.

One of the questions that was also raised by one of the analysts was regarding the limitations and self-inflicted bias set by the methodology or the 'inner nature' of the service.

During the workshop we found that many questions about the nature of connectivity within the service could easily be answered from studying the network models we created, but we observed a challenge within members of the organisation to translate from the concrete occurrences and thinking about functionality and usage of the service, to the abstract network model. The same also held in reverse. Figuring out the implications of findings in the abstract network models for the actual usage of the service proved a challenge for some members and seemed to be dependent on how well they understood the theory behind the abstractions.

With our post-hoc evaluation we found that the presentation of the results of our data-driven research efforts had noticeable effect on the participants. For those involved in the business development side of the organisation, seeing a successful application of data-driven research to their service convinced them of the feasibility and applicability of this type of research. Analysts were more focussed on the outcome of the predictive models, and what attributes had proven to be strong predictors of conversion behaviour. Interestingly, these results were somewhat corroborated by the user researcher, stating that in their qualitative inquiries they found extremely diverse motivations for becoming a subscriber, and little relation between the incentives the organisation provided by becoming a subscriber, and the actual motivation for people to do so. The finding that there was a network visible 'behind' the usage of the service proved to be valuable from a product design viewpoint, as the different industry clusters that were revealed could provide starting points for new product features, as well as perhaps having different problem situations that could provide opportunities for improvement of the design of the service. When asked how the participants of the workshop might apply data mining and machine learning techniques in their daily work, we found that for members coming from design and business appreciated now having a general understanding of how these techniques worked, and seeing them applied to their business made them realise that there could also be other areas of the business they could potentially be applied to.

4.3 EVALUATION

When we initially formulated this project, we specifically aimed understanding the potential that data-driven research methods have in the design and understanding of

usage of digital products and services. During the interactions with the organisation, we found that interest in insights into the usage of the product extends beyond merely those involved with the *design* of the product. This should not come as a surprise, but it calls into the question the initial framing with which we carried out the research. Instead of cordoning of understanding of user behaviour to those responsible for user experience, it was extended to other parts of the organisation, and widened the scope towards how *organisations* understand and learn about the usage of the product.

One of the things that we found most productive to observe was how the interpretation of the research results fostered a discussion between the different sections of the company, and what their individual perspectives and backgrounds brought towards the interpretation of the results, as well as potential new directions. Predictably each section focussed on an aspect that was relevant to their specific role in the organisation, but taken as a whole, having these models available to learn and reason about the usage of the service as a 'single source of truth' helped in developing the thinking about future directions of the service, as well as defining directions for future research.

Further more, in a company-wide presentation given after the results workshop, we found that attendees from almost every business function were willing and able to contribute to the discussion of the research results.

5. DISCUSSION

The previous sections described the application of data-driven research methods within a case study and the communication of the research results to the participating organisation. In this section an analysis of the data-driven research methodology, and the communication of its results is presented. Its aim is to assess the impact of data-driven research and how designers can move forward in an environment in which this of research is increasingly taking place. The aim of this discussion is to extract guidelines, links, and generalisations, from the results reported in sections 3 and 4. To properly review the results of our work we need to discuss our results on two levels; the case and its chosen approach, as well as the impact and potential of data-driven research methods in the understanding of user behaviour.

5.1 CASE STUDY

Within the case study, we need to assess what knowledge and new insights were gained by applying the data-driven research methodology. In our case we specifically focussed on using data-driven methods to enhance the understanding of user behaviour. To do so we used network modelling to learn about the connectedness between users, and predictive modelling in order to learn about which user properties had strong predictive power in predicting the conversion from a free user to a paying subscriber. In general, both techniques elaborated the understanding of *what was going on* in the usage of the service. We suspect that this is due to the focus of both analysis methods on making structures visible. The nature of these structures can then be reasoned about and speculated upon, providing new directions for future research and development of the service.

5.1.1 NETWORK ANALYSIS

We found that applying network analysis to the context and transaction data of a file exchange service was very powerful and valuable. Its powerful analytical capabilities stem from the many different perspectives network analysis provides. The micro (individual), meso (components and communities) and macro level (global) in the network models provide the opportunity to answer many different questions, as long as the researchers (and members of the company) are able to pose the question in such a way that it can be answered using network models. This requires familiarity with the way in which networks are modelled (graphs), as well as the various network measures that exists. Special attention should then be paid to the manner in which the translation happens from the implementation of the service and the data it generates to the abstract network models that are used to analyse the usage data. Also, attention should be paid to how representative the patterns *really* are. It's easy to spot appealing patterns in network visualisations, but for people unfamiliar with network analysis it is hard to estimate to

which extent these patterns generalise, and how these are translated back into information about how the service is being used.

5.1.2 PREDICTIVE MODELLING

The creation of the predictive model turned out to be relatively straight-forward, provided we had constructed the necessary properties in order to train the classifier. While we are pleased with the results of the models we trained, we do feel that it is currently more of a proof of concept, than a well-trained model on a well-constructed dataset. Due to the omission of subscription dates of individual users, we had to infer these dates based on the individual transactions. We would feel more confident with having these dates directly available to train the classifiers on. That being said, this initial model is already quite predictive and informational. We found it interesting that during the discussion of the research results with the participating organisation earlier qualitative user research conducted by the company pointed into similar directions with regards to the attributes that predict the conversion of a free user to a paying subscriber.

5.2 DATA-DRIVEN RESEARCH

As described in the introduction of this thesis, the emergence of data-driven research methodologies has led to their adaptation in the design and development of digital services. In the initial problem definition we stated that only designers are in need of experience and skills in order to to thrive in an environment in which data-driven research is the norm, but having seen the diversity of interest from different areas within an organisation that creates a digital service, it might be better to frame these research activities under the banner of organisational learning, rather than as an research activity that is done as part of design research. We would argue that a broader set of people within such an organisation are in need of these skills and experiences. So what might we do to enhance these skills and experiences? Based on the results from the case study and the communication of the results we point out a number of steps that can be taken to enhance the understanding of data-driven research.

5.2.1 THE DEMYSTIFICATION OF DATA SCIENCE

One of the first things we need to consider, is how data science is embedded in organisations, and how people in these organisations view it. The emphasis on *'translating data into value'* as several authors (van der Aalst 2016; Gillon et al, 2012; Mithas et al, 2013) define it, suggests that data science is often done inside a business or mixed environment rather than a purely scientific one. As such, we should assume that not all members of an organisation have the same level of experience, or are even familiar with the fields and its norms and values from which data science emerged. Boyd and Crawford (2012) define Big Data (which we described as the popular umbrella term for the growth in the availability of data, and its computational analysis) as a cultural, technological, and scholarly phenomenon which rests on the interplay of technology, analysis and mythology. In an environment in which decisions are taken that affect the future design of a service, we feel this mythical component of the Big Data phenomenon can become problematic. The *'aura of truth, objectivity and accuracy'* as boyd and Crawford describe it, could potentially shut down critical thinking around the implications that the results of data-driven research point to. This would make it seem logical to look at ways to *'level the playing field'* when it comes to the experience with data science.

An obvious strategy to pursue would be to educate members with diverse backgrounds in some of the underlying fields that make up data science. We did this in our workshop and presentation in which we explained the basic concepts behind network analysis and data mining. Based on the observations from the workshop we described in section 4, we feel that the relevance of these efforts is very much dependent on the extent to which participants were able to translate and integrate those concepts into their own working practice. This suggests that it's beneficial to present research results from data-driven research using different modalities, so that every member of the organisation can bring his or her own perspective and expertise to bear on the models and the directions and actions they point to.

5.2.2 REASONING WITH DATA

As established in the previous section, there are various forces shaping the emergence of data-driven research methodologies, one of which is the *market*. Within machine learning there is a tradeoff known as the "Exploration versus Exploitation" tradeoff. It applies to systems that want to acquire knowledge, and at the same time maximise the reward for their performance. Based on our experiences in conducting this research project, we feel that it is worthwhile to consider this tradeoff on a larger organisational level as well. While there might be direct and possibly short-term business benefit in *exploiting* certain patterns that are found using data-driven research, this could be to the detriment of the possibility of gaining a deeper understanding of first causes in the underlying patterns. It is tempting in an environment such as a digital service to treat the availability of data as a given, but an awareness of the root causes and phenomena that produce the data ensures that the study of these data can be relayed back into systematic study of these phenomena. In our case study we started from the hypothesis that the social organisation of people was somehow influencing the usage of the service, and that gaining insight into this social organisation could help the organisation learn more about the usage of the service.

Another thing to take into account is that *interpretation* is at the centre of data analysis. Any dataset, regardless of size is subject to limitations and biases. As such it is important to outline these biases, or misinterpretation of data is the result. It is important to consider how some of these biases are established by the design and implementation of the service.

Next, we need to consider how, when these types of research activities are undertaken by an organisation, the results of this research is communicated within the organisation. The different models and their accompanying visualisations we created in our case study establish a mediated relation between the service, its users, and the people attempting to understand the usage of the digital service. The representations operate as a model of the current usage of the service, and aid thinking about how the service might be designed in the future. The models and their representations generate what Ihde (1990) refers to as a *"hermeneutic relation"*, as technology (the various models and their representations) embodies the world, and through this technology the world is read, interpreted and explained by the people involved in the creation of the digital service. Leurs (2014) describes these hermeneutic relations as *people making sense of reality through technology.*

6. FUTURE WORK

When considering future work, we feel that there are many avenues left to explore. If anything, this study has shown that there is much to be gained from adopting a data-driven research perspective when designing digital services. Considering the fact that this study operates on two distinct levels (the case study, as well as the meta perspective of doing data science to foster the understanding of user behaviour) we have the following recommendations for future work.

6.1 WITHIN THE CASE STUDY

As previously mentioned the current predictive models could benefit from further iteration and development using better data about the subscriptions of users. The network models could also be taken into new directions, for example to analyse them from a dynamic network perspective. Furthermore, as we only explored the usage of the service within two countries, scaling the research effort to the entire user base would be an obvious first step. Also, results from the case study could be implemented in the platform, which would result in the creation of distinct data products. Another area would be tool development. Software tools such as Gephi are currently not equipped to handle the scale of networks that are found in modern digital services. This restricts the analysis of these networks to computational tools that are only accessible to analysts with programming skills or computer scientists.

Another direction to consider is how findings from data-driven research are communicated and shared within organisations. Our workshop and presentations showed that when care is taken in clearly communicating research results, and providing the audience with an opportunity to adopt the results into their own thinking around the design and development of a service, many potential new directions for applications and research directions can emerge.

6.2 WITHIN DESIGN & DATA SCIENCE

To evaluate how the emergence of data science can affect designers and illustrate steps that can be taken to move forward, we look at two distinct levels of design activity, to assess future steps that can be taken with regards to the profession, and the practice.

6.2.1 DATA SCIENCE AND THE DESIGN PROFESSION

The current study as conducted, is very much an exploration of the potential of data-driven research methods. Its emphasis is on using data science to develop organisational understanding of usage of a digital product or service. One major obstacle within the adaptation of these methods by researchers working within a design context would be their formal training. The emphasis on qualitative research, while still relevant and necessary, should be augmented with sound development of data-driven quantitative research skills. We speculate that this qualitative / quantitative gap is likely not closed by individual researchers, but by small groups of researchers, working in teams and making use of each others strengths. How this can be done is an important direction for future research.

6.2.2 DATA SCIENCE AND THE DESIGN PRACTICE

As we mentioned in our background on data-driven research methodologies in design, the current emphasis of data-driven research is focussed on moving towards an evidence based practice in the design of digital products and services. Through public testing of multi-variate designs the design with the best fit between user behaviour and organisational objectives can be converged upon. As this approach has revealed that 'design-intuition' can be at odds with data coming from tests or experiments, the practice of the designer changes; he or she needs to become able to embrace failure, and should be receptive towards continuous iteration. Yet, as we have shown, data-driven research can be much more than experimentation with multivariate designs. As we have shown in this article, network analysis and predictive analytics are also important forms of data-driven research that can provide structured insight into the behaviour of users and provide new insight and opportunities for designers to learn and build an understanding about the behaviour of their users.

Another important thing to consider here is how the 'fit' between user behaviour and organisation objectives is construed. Designers traditionally argue *for* the user. Their understanding of context, people and how things are used and understood, guides the design decisions that are made. If these are at odds with the results coming from data-driven enquiry, which solutions are then implemented? How can designers become able to articulate their findings from qualitative research in such a way that it can be reconciled with quantitative results coming from data-driven research?

Decision making emphasising maximal exploitation of a design might produce desired organisational outcomes on the short term, but what happens in the long term if these outcomes at odds with the contexts in which the products and services are used? Resolving these conundrums we feel is an important direction for future work.

7. CONCLUSION

This section aims to answer the key questions that were present in this research project. The goal for this project was to understand the potential that data-driven research methods have in contributing to understanding of the usage of digital product and services. We explored how the emergence of data science could provide us with an approach to study the usage of a digital service at scale, and allow us to generate new insights into its usage. Through this paper and the accompanying case study, we have explored the potential for the adaptation of data-driven research methods in helping an organisation understand the usage of its service in a novel manner. By executing a case study we were able to test the potential for these data-driven methodologies. We used network analysis in order to explore the social organisation of users using the service, and found it to be a powerful analysis method to analyse the usage and distinct levels of magnitude, revealing patterns of usage that were not previously visible. These patterns provided opportunities to improve the offering of the service. We used machine learning techniques in order to create predictive models that could predict the conversion of a free user to a paying subscriber. We were able to successfully train these models and gain new insight into user attributes that contributed to this conversion. When sharing these results with the participating organisation we found that our initial framing of using data-driven research methods as part of design research efforts, was misaligned, and could potentially be better thought of through the perspective of organisational learning. We showed that with clear communication and instruction, complex subjects such as machine learning an data mining can be opened up to many different members of an organisation that facilitates a digital service. We expect that opening up these techniques could help in demystifying data science, and ensure that its usage is critically evaluated in organisations, as the models created through its usage establish a mediated relation between the usage of the service, and those trying to interpret that usage and give direction to future development of these service.

All in all, we found that by adapting the perspective of a data scientist, and conducting data-driven research, many new avenues for exploration were opened up in the possibilities for learning about product usage. Its methods, findings and predictions are a powerful contribution to the arsenal of research techniques that are available to anyone studying the usage of digital products and services. Especially when taking into account how reciprocity can be established with qualitative techniques that are currently in use for learning about the social, cognitive and emotional needs of people. We foresee a bright future for the incorporation of data-driven research methodologies in the design of digital products and services.

8. ACKNOWLEDGEMENTS

We'd like to thank Frank Takes of LIACS / Leiden University, and Péter Kun of Delft University of Technology for their input during this project and a review of the first draft of this text.

9. REFERENCES

van der Aalst, W. M. (2016). Process mining: data science in action. Berlin, Heidelberg: Springer. http://doi.org/10.1007/978-3-662-49851-4

van der Aalst, W. M. (2014). Data scientist: The engineer of the future. In Enterprise Interoperability VI (pp. 13-26). Springer, Cham.

Abbasi, A. Sarker, S. and Chiang, R.H.L. (2016) "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," Journal of the Association for Information Systems: Vol. 17 : Iss. 2 , Article 3. Available at: http://aisel.aisnet.org/jais/vol17/iss2/3

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments. the 23rd international conference (pp. 283–292). New York, New York, USA: ACM. http://doi.org/10.1145/2566486.2567967

Barabási, A.-L., & Pósfai, M. (2016). Network science.

Boeijen, A. ., Daalhuizen, J., Zijlstra, J., Schoor, R. ., & Technische Universiteit Delft. (2014). Delft design guide: Design methods.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. IEEE transactions on visualization and computer graphics, 17(12), 2301-2309.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

Breiman, L., & Cutler, A. (2007). Random forests-classification description. Department of Statistics, Berkeley, 2.

Brinkmann, G.G., Rietveld K.F.D. and Takes, F.W. (2017) Exploiting GPUs for fast force-directed visualization of large-scale networks, in Proceedings of the 46th International Conference on Parallel Processing (ICPP), pp. 382-391.

Cao, L. (2017). Data Science. ACM Computing Surveys, 50(3), 1–42. http://doi.org/10.1145/3076253

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chen H, Chiang RH and Storey VC (2012) Business intelligence and analytics: from big data to big impact. MIS Quarterly 36 (4), 1165–1188.

Cioffi-Revilla, C. (2010). Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3), 259–271. http://doi.org/10.1002/wics.95

Cukier, K., & Mayer-Schönberger, V. (2013, March 29). The Rise of Big Data. Foreign Affairs, 1–14.

Dhar, V. (2013). Data science and prediction. Communications of the ACM. http://doi.org/ 10.1145/2500499

Draper, S. W., & Norman, D. A. (2009). User centered system design: New perspectives on human-computer interaction. Boca Raton, FL: CRC Press

Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets. Cambridge University Press.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. interactions, 19(3), 50-59.

Gillon, K., Brynjolfsson, E., Mithas, S., Griffin, J., & Gupta, M. (2012). Business analytics: Radical shift or incremental change?.

J. D. Hunter, "Matplotlib: A 2D Graphics Environment," in Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, May-June 2007. doi: 10.1109/MCSE.2007.55

Ihde, D. (1990). Technology and the lifeworld: From garden to earth. Indiana University Press.

Aha, D.W., Kibler, D., & Albert, M.K. (1991). Instance-based learning algorithms. Machine learning, 6(1), 37-66.

King, R., Churchill, E. F., & Tan, C. (2017). Designing with Data. "O'Reilly Media, Inc.."

Kluyver, T. Ragan-Kelley, B. Pérez, F. Granger, B, Bussonnier, M. Frederic, J. Kelley, K. Hamrick, J. Grout, J. Corlay, S. Ivanov, P Avila, D. Abdalla, S. Willing, C, Jupyter Development Team "Jupyter Notebooks – a publishing format for reproducible computational workflows" (2016) in Positioning and Power in Academic Publishing: Players, Agents and Agendas doi: 10.3233/978-1-61499-649-1-87.

Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. Harvard Business Review, 95(5), 74-+.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788–8790. http://doi.org/10.1073/pnas.1320040111

LaValle S, Lesser E, Shockley R, Hopkins MS and Kruschwitz N (2011) Big data, analytics and the path from insights to value. MIT Sloan Management Review 52 (2), 21–32.

Liikkanen. L. (2017). The data-driven design era in professional web design. interactions 24, 5 (August 2017), 52-57. DOI: https://doi.org/10.1145/3121355

Leurs, B.L.F. (2014) Tools for Proximity: Helping designers to make sense in the boardroom.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Mauri, M., Elli, T., Caviglia, G., Uboldi, G., & Azzi, M. (2017). RAWGraphs: A Visualisation Platform to Create Open Outputs. In Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter (p. 28:1–28:5). New York, NY, USA: ACM. https://doi.org/10.1145/3125571.3125585

Maxmind, (2017) GeoLite2 Free Downloadable Databases. Retrieved from: https://dev.maxmind.com/geoip/geoip2/geolite2/

Mithas, S., Lee, M. R., Earley, S., Murugesan, S., & Djavanshir, R. (2013). Leveraging Big Data and Business Analytics [Guest editors' introduction]. IT professional, 15(6), 18-20.

McKinney, W. (2010). Data Structures for Statistical Computing in Python . Proceedings of the 9th Python in Science Conference, 51-56.

O'Neil, C., & Schutt, R. (2013). Doing Data Science. "O'Reilly Media, Inc.."

Opsahl, T., 2013. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social Networks 35, doi:10.1016/j.socnet.2011.07.001

Quinlan, J. R. (1993). C4. 5: Programming for machine learning. Morgan Kauffmann, 38.

Reas, C., & Fry, B. (2007). Processing: a programming handbook for visual designers and artists (No. 6812). Mit Press.

Sanders, L., & Stappers, P. J. (2014). From designing to co-designing to collective dreaming: three slices in time. Interactions, 21(6). http://doi.org/10.1145/2685354.2670616

Sayad, S. (2017) Naiev Bayesian Classifier. Retrieved from: http://www.saedsayad.com/naive_bayesian.htm on 29-12-2017.

Schrage, M. (2014). The Innovator's Hypothesis: How Cheap Experiments are Worth More Than Good Ideas. MIT Press.

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. European Journal of Information Systems, 23(4), 433–441. http://doi.org/10.1057/ejis.2014.17

Tiago P. Peixoto, "The graph-tool python library", figshare. (2014) DOI: 10.6084/m9.figshare.1164194 [sci-hub, @tor]

Turner, V. Gantz, J.F. Reinsel, D. and Minton. S. (2014) The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. International Data Corporation, Framingham, MA, USA, 2014. http://www.emc.com/leadership/digital-universe/.

van der Walt, S. Colbert, S. C. and Varoquaux G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, vol. 13, no. 2, pp. 22-30 doi: 10.1109/MCSE.2011.37

Waskom, M. Botvinnik, O. O'Kane, D. Hobson, P. Lukauskas, S. Gemperline, D. Qalieh, A. (2017, September 3). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. http://doi.org/10.5281/zenodo.883859

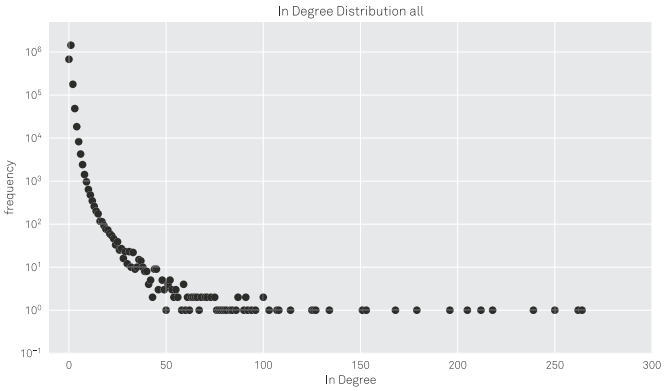
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining. Morgan Kaufmann.

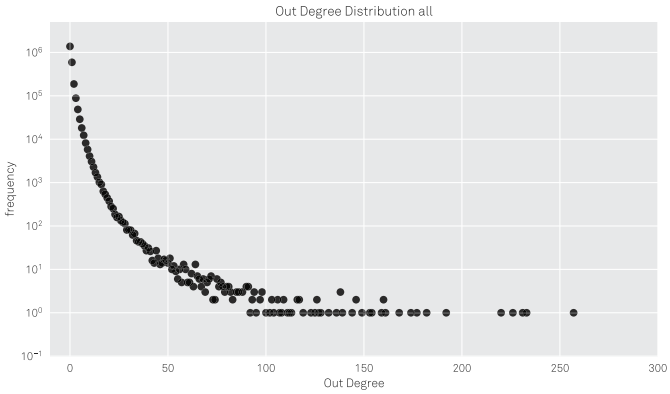
White, W. (2017) Freemail - A database of free and disposable email domains. https://github.com/willwhite/freemail

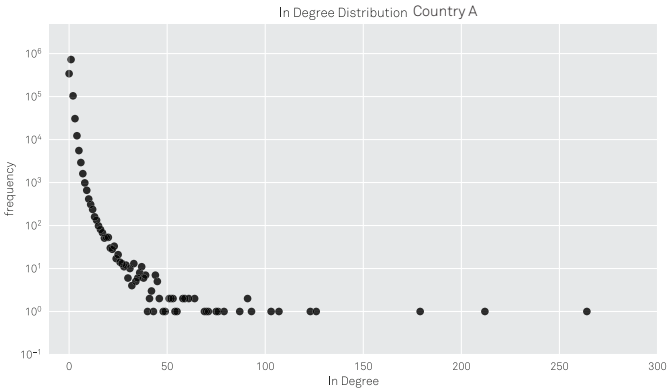
10. APPENDIX

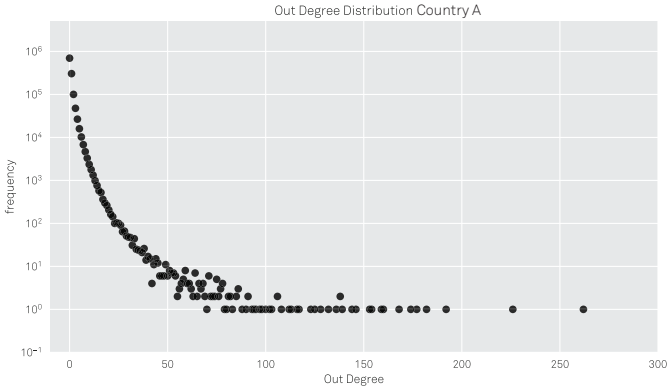
A. NETWORK MODELS

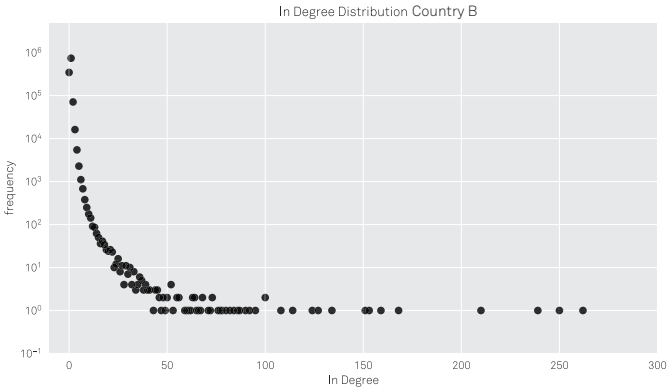
DEGREE DISTRIBUTION FOR THE COUNTRY, DOMAIN AND E-MAIL NETWORK MODELS

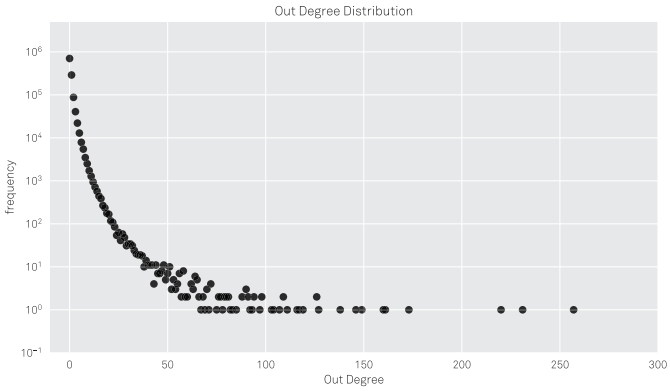




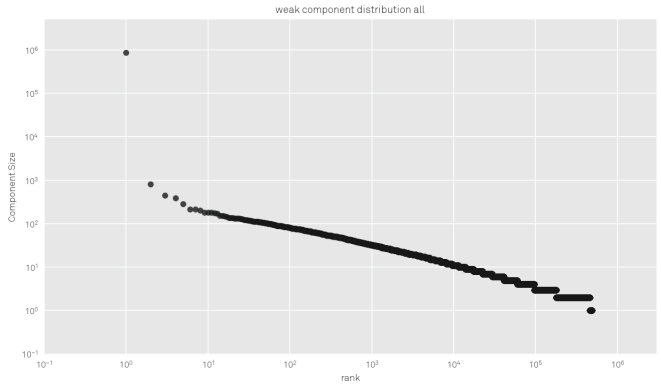


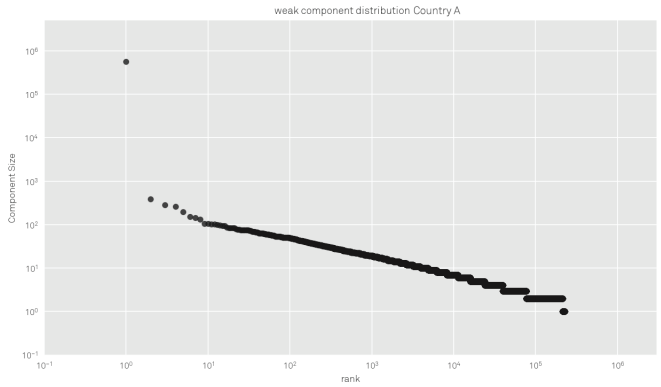


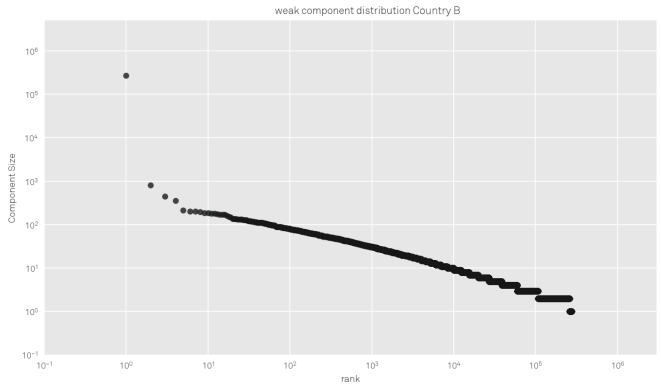




COMPONENT SIZES FOR THE COUNTRY, DOMAIN AND E-MAIL NETWORK MODELS







B. COMMUNICATION

QUESTIONS FOR THE EVALUATION OF THE RESULTS WITH THE PARTICIPATING COMPANY:

Baseline

- Today I hope to learn, see and discover...
- Do you think it's possible that all [...] users are connected to each other in some way? What would this connectivity look like?
- If I could predict the behaviour of our users, I would be most interested in predicting:
- Are there any questions you have upfront about networks, datamining or machine learning?

Post Presentation

- I was surprised to learn today that:
- Now that we've established that there is a network behind the service's usage, is there a way you can apply that knowledge in your daily work?
- Aside from predicting the conversion from free to paid users, with what you've learned today about datamining and machine learning, can you apply within your daily work?
- Are there any questions that you have that you would like to see answered?