Structure of Charity Networks

Annika Wipprecht

Somerville

University of Oxford

A thesis submitted for the degree of

MSc in Mathematical Modelling and Scientific Computing

2010/11

Acknowledgements

I would like to offer my sincerest thanks to the following people for their help and support during this thesis:

My supervisor Mason Porter who offered me the possibility to work in Network Science, offered guidance, useful comments, encouragement and providing me with the Facebook data set. Further, I would like to acknowledge Marc Stein for supplying the consumer data of FirstGiving. I also like to thank Kathryn Gillow for her help during the course.

Abstract

In this thesis, we study two data sets provided by the online fundraising company FirstGiving. People can use FirstGiving to create a personalized fundraising web page for a charity. From now on these people are called fundraisers. The fundraiser can distribute the web page among his or her friends, family members or acquaintances. These people can then donate to the charity via FirstGiving as well. The data set consists of the consumer data. We obtained two files. One file contains personal information and demographic data about the users of FirstGiving. The other file consists of the transaction data between donors and charities. In the first part of this thesis, we construct networks from the data set using charities and people as vertices. A tie between a person is formed whenever he or she donated to a charity. We study networks depicting the data from various time slots by applying network diagnostics to these networks. These network diagnostics include the degree and degree distribution of vertices, the path between vertices, the closeness centrality and the betweenness centrality. This is followed by deriving and testing a collective behaviour model imitating the donation behaviour of people using FirstGiving.

Contents

List of Figures

Chapter 1

Introduction

1.1 Context and Motivation

FirstGiving [10] is an online company that enables private fundraisers to collect money for a cause of their choice by providing tools to create a web page. A person wanting to fundraise money online creates an account with FirstGiving and then has the opportunity to use the tools provided by FirstGiving to build a personalized web page that raises awareness for a charity partnering with FirstGiving. The fundraiser then distributes the web address of the site to friends, family members, and other acquaintances. If someone is interested in donating money to the charity via the fundraiser, the money is transferred to a bank account through FirstGiving.

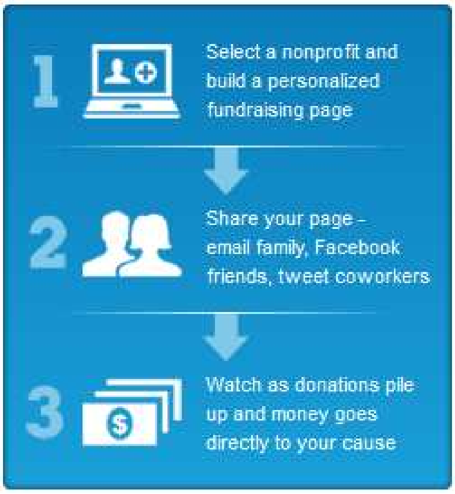
1
Select a nonprofit and build a personalized fundraising page
2
Share your page – email family, Facebook friends, tweet coworkers
3
Watch as donations pile up and money goes directly to your cause

Figure 1.1: Fundraising with FirstGiving. Picture was taken from the FirstGiving web site [10].

FirstGiving was created in 2003, and since then has helped to raise more than $1 billion for charities world-wide. We obtained the data of the transaction processes of the US based company starting from the beginning of the company (2003) up until 2010 from Marc Stein. The text files containing the information are of a size of more than 3GB. This data set includes the identification number of a donor, the precise time to the second, and the amount in US dollars when the donor donated to a charity via a certain fundraiser. Further, the data set also includes personal characteristics of selected donors. That is, we have the personal details of 1,048,574 people out of 3,244,874. The personal information about the customers include, for example, the age and gender, the state they are coming from together with the zip code and city, a tendency in numerical values of whether a person is is interested in donating to or fundraising money for a charity, information about the range of the household income, and, finally, a classification into different social groups made by an external company, Mosaic. Mosaic splits the US population into 13 social groups such as “Affluent Suburbia” or “Blue-collar Backbone”. For this dissertation, we limit our interest to a selected number of characteristics: we will only be concerned about information of the age, state residence, and household income.

The aim of this thesis is to study the data set by following the donation behaviour of the customers and users of FirstGiving over time. Then we want to apply network diagnostics to understand the structure of the networks extracted from the data set. This enables us to say more about the entire data set. That is, can we say anything about the people using FirstGiving? If one donates via FirstGiving do people donate more than once? Further, we aim to use this information to construct a model simulating donor behaviour.

The structure of this thesis is as follows. First, we extract some demographics from the consumer data. This is followed by a section where we state the donations and definitions of the network and network diagnostics used in this report. This is followed by a section where we discuss the results of applying the diagnostics to networks constructed from the second data file that contains information about the donation transactions. In the second part of this report, we introduce the concepts of epidemic models. Consumer behaviour is based on collective behaviour, we will discuss some literature on the issue before stating assumptions that lead to our consumer model. We then study and test the model. Finally, we test whether the model leads to a similar network structure as observed from networks constructed from the data of FirstGiving.

Chapter 2

The Data Set

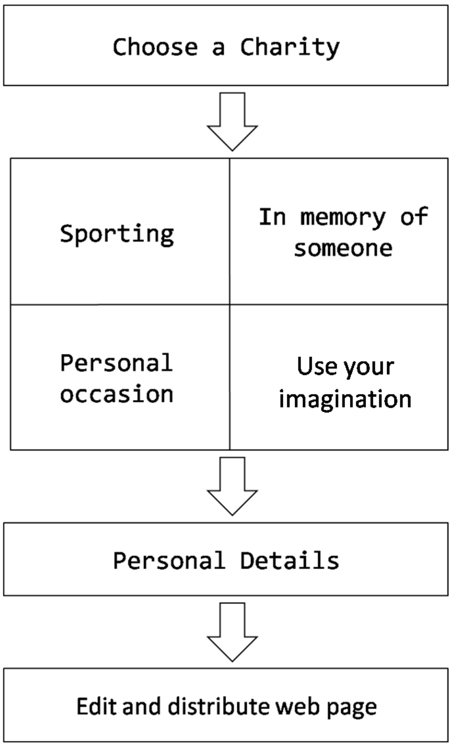

Figure 2.1: Steps to create a personalized fundraising webpage. Schematic indicating steps to create a web page using FirstGiving [10].

Before applying network diagnostics, we give a general overview of the data set. Whenever using FirstGiving to start ones own fundraising campaign (see Figure 3.1), one can choose one of their partnering non-profit charities, or an event, team, or person who already supports a charity. Suppose one wants to create a web page then one has to choose what the initial occasion for the fundraising is. These are categorized into four events: “Sporting” - for example dedicating a marathon run for a charity, “In memory of someone”, “Personal Occasion” - as in wedding or birthday celebration, and “Use your imagination” - that includes fundraising campaigns such as “grow a beard” or “quit smoking” [10]. FirstGiving targets people who want to fundraise because of a personal reason or a special occasion. After choosing a cause, one has to decide how long one wants to fundraise for: 6, 12 or 24 months long. Then, FirstGiving asks for personal information such as name, address, city and state of residence together with the zip code, and country. Afterwards, the fundraiser has the opportunity to edit a standardized web page, and distribute the address of the web page to friends and acquaintances. However, the newly created web page can also be accessed via the FirstGiving web page by either looking for the fundraiser or looking for the cause, and then choosing the fundraisers page to donate money. Beside the credit card details, the donor has to provide information about his address (again city, zip code and state) with no further information needed in order to place a donation.

The data set is split into two files. Both files include the entire customer data of the United States based company from 2003 until 2010. The first file contains information about 2,805,391 users. The file includes information about, among others, the age, household income, and state.

The second file contains the transaction data. That is the time whenever someone donated money via a fundraiser. In this file, we have the donor ID, the fundraiser ID via which the donor donated money, the amount donated in dollars, a number code for the charity, and the time and day up to seconds when the transaction took place.

Starting with the file containing the consumer data, we did some basic summary of the demographics. Thus, from the 1,048,574 rows contained in the file, there are 975,710 distinct consumers registered. Many consumers were registered more than once. Probably[1], they used FirstGiving multiple times, and some for their personal details changed. However, about 94% of the consumers are registered once, 5% twice, and the remaining 1% more than twice (with one person being registered 144 times).

Figure 2.2 shows the distribution of the donors of each state from the last 10 years.

[1] We tried to contact Marc Stein, but he did not respond. Thus, while analysing the data set, we made the assumption that consumer IDs were distributed once and only assigned to one person. Note that this might not be the case as another scenario might be that users get deleted after a certain amount of time and then the number is redistributed.

From the entire list, 1,111,025 entries come from the US. This number is less than it should be as some people made spelling errors, and whenever unnoticeable or the spelling error was made less than 10 times we disregarded these entries. Further, we also ignored any donations made from outside the US. Otherwise, the abbreviations for the 50 US states follow the standard list. Washington D.C. (DC) is included in the state list.

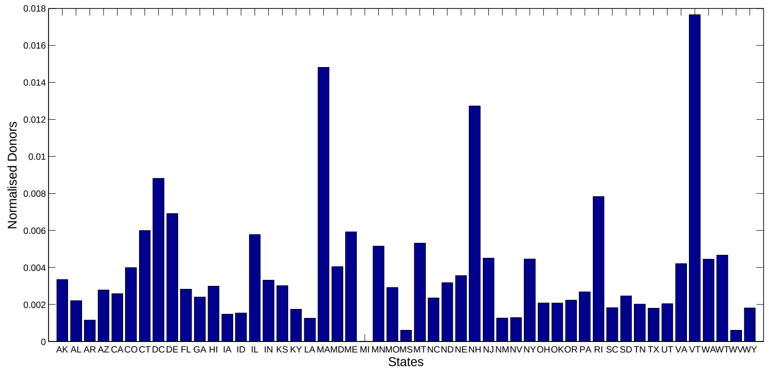

Figure 2.2: Number of donors normalised by population of each US state.

We normalised the data of the Figure 2.2 by the population of each state recorded by the census of 2010 [7]. The highest contributions of consumers (both, donors and fundraiser) are living in Vermont, Rhode Island, New Hampshire and Massachusetts with Californian and Massachusetts having the most contributors in a non normalised version. FirstGiving has its headquarters in Boston. Thus, it probably started its promotion of the company in this area. California has the highest population in the US. Thus, it was to be expected that the most consumers are coming from this state. However, when looking at the normalised graph we see that California has a rather low percentage of the population donating via FirstGiving. The lowest contribution of percentage of donors is from Michigan.

The next demographic we looked at was the age of the donors. We plotted the number of donors against the information about the age a person entered when donating money. This plot can be seen in Figure 2.3. Here, it becomes clear that the data should be taken with caution as the data indicates it is possible that 80 plus year olds used an internet company to donate but less likely. Nevertheless, Figure 2.3 shows that the majority of users of FirstGiving are in their late thirties to early fifties. The data plot has a mean value of $\bar{\mu} = 43$ and a variance of $\bar{\delta} = 24$. However, this and the

Abbreviation	Range of household income
A	\$ 1,000 - \$14,999
B	\$ 15,000 - 24,999
C	\$ 25,000 - \$34,999
D	\$ 35,000 - \$ 49,999
E	\$ 50,000 - \$74,999
F	\$ 75,000 - \$99,999
G	\$ 100,000 - \$124,999
H	\$ 125,000 - \$149,999
I	\$ 150,000 - \$174,999
J	\$ 175,000 - \$199,999
K	\$ 200,000 - \$249,999
L	\$250,000 +

Table 2.1: Range of the household income of donors

following conclusions drawn from the plots should be taken with caution because as mentioned before the user might enter wrong information making the data set biased.

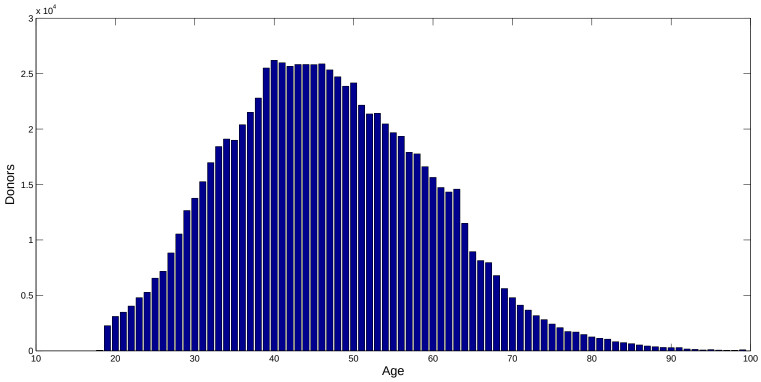

Figure 2.3: Age distribution of donors.

Finally, we also plotted the income distribution of the donors in Figure 2.4. Table 2.1 shows the abbreviations used in Figure 2.4. The majority of people donating have an household income between $\$50,000-\$74,999$. The median household income in 2010 in the US was \$50,221 [7]. Figure 2.4 indicates that most of the contributors using FirstGiving earn above the mean income of US households. However, again this information might be bias for the same reason as for Figure 2.3.

The next several figures give a summary of the observations made in the second data file. This file contains information about the time evolution of the donations

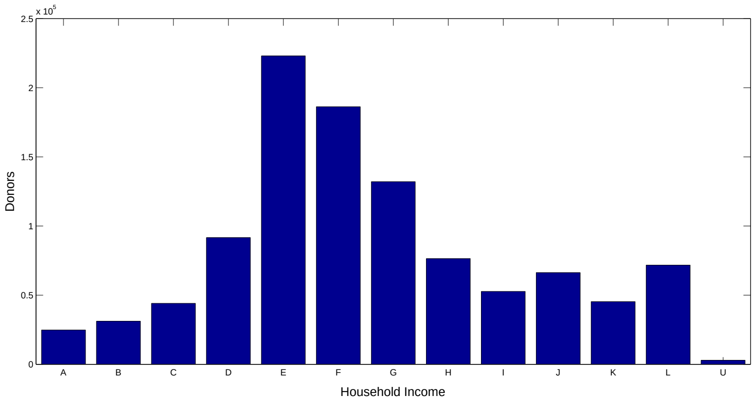

Figure 2.4: Distribution of household income of donors.

made. Thus, the figures show the time evolution using different units. First, we started with a year plot for donors' contributions to charities (Figure 2.5).

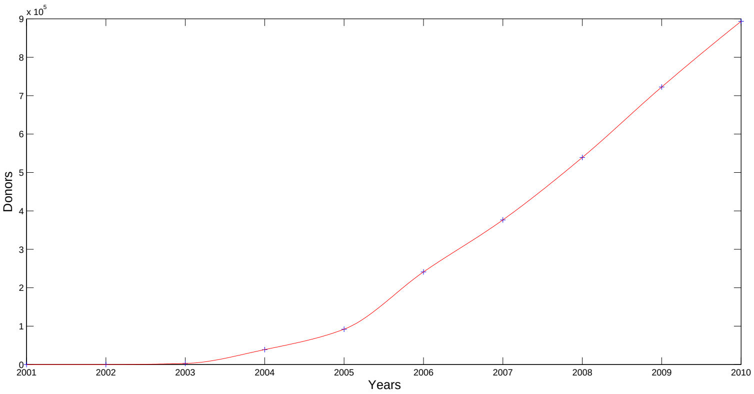

Figure 2.5: Number of donors vs. year.

In Figure 2.5, the crosses indicate the data measured from the file. The red line connects the data points. We can see that in the years 2001 and 2002 no donors contributed donations via FirstGiving, as the company was set up in 2003. It seems to be that there is a bend in the curve at 2005 having a sharper slope afterwards. The next plot shows the contributions of donors for different months. Again, we normalised the data this time by the amount of donors in a year.

The data points in Figure 2.6 indicate the fraction of donors in a certain month. We can see that most donations take place in April and September. April is the month when many people receive tax returns. The peak in September is likely explained by donations dedicated to the terror attacks at September 11th 2001.

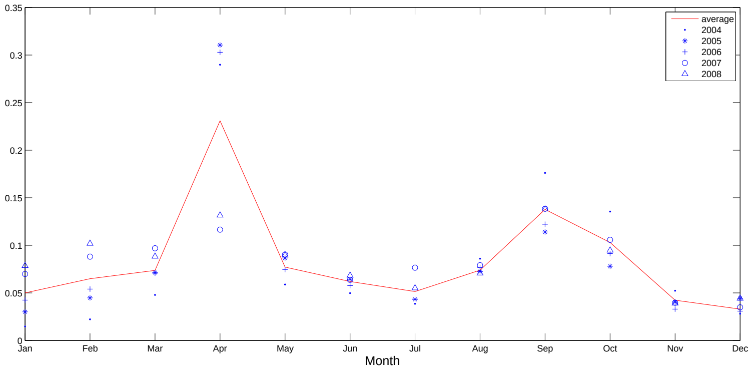

Figure 2.6: Number of donors vs. month.

Again, we can see a change in behaviour from 2005 to 2006, as the donors are more equally distributed over all months after 2005. The years before, the peaks in April and September are more pronounced.

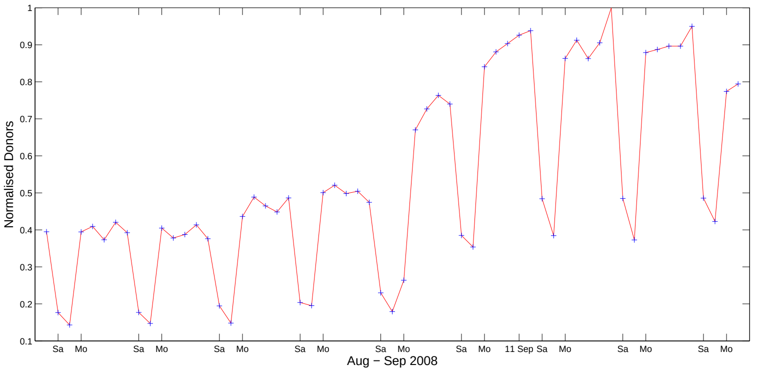

Figure 2.7: Number of donors vs. days.

In Figure 2.7, we also plotted the normalised number of donors (whereby we normalised such that the maximal number of donors in both months is equal to one) against 61 days. The plot shows the data for August 1st until September 30th with blue crosses. Again, to give the eye guidance, we connected the data points with a solid red line. Interestingly, most of the donors give money during weekdays rather than on weekends. However, there does not seem to be a preferable day during the working days when more people donate money. We also can see again that there are more donors on September than in August. Also, there is no obvious increase of the number of donors at September 11th, but the number of donors does seem higher around this day.

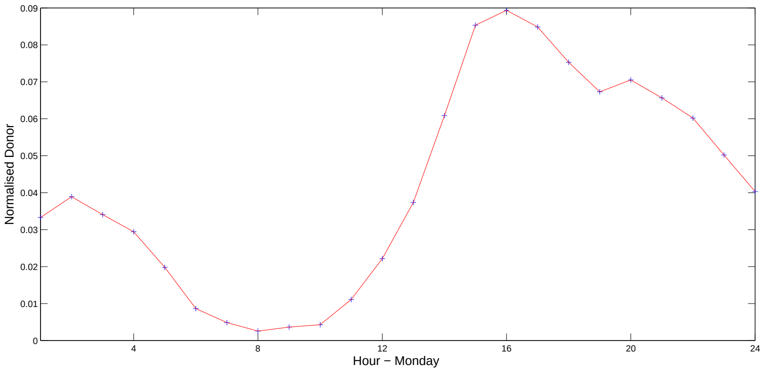

Figure 2.8: Number of donors vs. Mondays in August until September 2008.

Finally, we plotted the number of donors against the hours of the day. Here, we distinguish between Mondays (Figure 2.8) and Saturdays (Figure 2.9). We used again the same time slot as for Figure 2.7, i.e. every Monday and Saturday in August and September 2008. The time saved in the file is most likely East Coast Time, but we do not know this with certainty. However, there are 5 different time zones in the US. Thus, we needed to assign the right time for each donor using the state information of each consumer from the previously discussed file. Note that for some donors, there was no data saved in the consumer files. We therefore did not include the donors without information. Also, states that have more than one time zone are considered to belong to only one time zone. Further, we normalized every day by the total amount of donors of each day. Then we took the mean of all days (9 Mondays and 8 Saturdays).

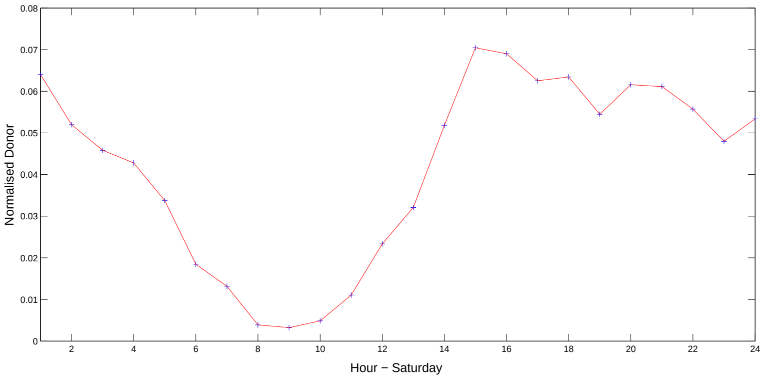

Figure 2.9: Number of donors vs. Saturdays in August until September 2008.

The plots can be seen in Figure 2.8 and Figure 2.9. Though one might expect the daily routines of people to be different on weekdays and weekends. Interestingly, most donors give money during the afternoon. On Mondays, we can see a peak of donations around 16.00 and then during the evening the amount of donors decreases. Note, however, that on Saturdays donations reach a maximum at 14.00 but donors are also active in the evening and at night time. Certainly, we have to have in mind that some donors might enter wrong information but it seems that quite a lot of people also donated during the night - again more on Saturday than on Monday. However, this information should be taken with caution, as we are not sure whether the donation transactions are stored in Eastern time. From the plots, we see that the donation behaviour in the sense varies that Saturdays people donate more throughout the day whereas Mondays people tend to donate at the end of usual business hours or the early evening.

In this chapter, we observed from the data that users of FirstGiving on average are middle-aged people with good salaries. Further, we can conclude that people are more inclined to donate money via FirstGiving in a month whenever they seem to get a tax return or in memory of September 11[2]. Nonetheless, there is no particular peak visible during September 11. Also, people donate more during the week then over the weekend. However, the hours during which people donate seem to be similar Mondays and Saturdays. We also ploted similar graphs with donations and charities against years, months, days and hours, but each of these graphs behaved in a similar way to the respected graphs with donors. Thus, we decided to omit the graphs for brevity.

[2] August 29th was the day when Hurrican Katrina reached the coast of Louisiana. We plotted the number of donors against days in September and October 2005, and the plot looked similar to Figure 2.7.

Chapter 3

Networks

3.1 Network Representation

This section gives an overview of standard notation in network theory but also details some network diagnostics used in this work. The definitions were taken from Newman [15], and Barrat *et al.*[3]. Note that we already discussed some centrality measures such as degree, and eigenvector centrality, closeness and betweenness centrality, in a Special Topic [24].

A network is commonly understood as a system that can be represented as a graph, $G = (V, E)$, with a set of nodes or vertices, V, representing the elements of the system and a set of links, ties or edges, E, connecting the elements and representing some sort of interaction between the elements. We will denote the cardinality of V as N throughout this work. Further, elements of V are usually denoted with lower case letters, i.e. we say, the vertex i is an element of V. If vertex i is connected to vertex j, then $\{i, j\}$ denotes an edge in E. A network is said to be undirected if for $\{i, j\} \in V$ then $\{j, i\} \in V$ for all i, j in E. A network is directed if this is not the case. We say the cardinality of E is m.

A common representation of a network is an adjacency matrix. An adjacency matrix, A, for a simple graph, that is a graph where vertices are connected with at most one edge, has elements $a_{ij\,[i,j\in\{1,\ldots,N\}]}$ such that

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \quad \forall\ i, j. \tag{3.1}$$

Obviously, for a directed graph, A is symmetric. Actually, equation (3.1), represents the entries for a simple unweighted graph. If the network has multiple edges distributed between the vertices, then the number of ties connecting the vertices replaces the 1. A weighted graph has instead of the number of ties, a weight as a non-zero entry. For example, if the network represents a phone network with the set of vertices being people and the set of ties are calls made between the people than a weight might be the duration of the call.

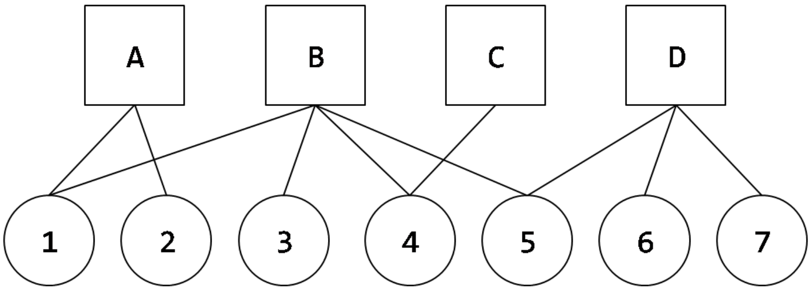

Figure 3.1: Schematic of a bipartite network. The vertices are forming the following subsets: $X = \{1, 2, 3, 4, 5, 6, 7\}$ and $Y = \{A, B, C, D\}$.

For some networks, the set of vertices has two distinct subsets, X, and Y. If no ties connect different elements of the subsets but the ties run between elements of distinct subsets, then a network is said to be bipartite. For example when one tries to group people into common groups of classes, then one vertex subset contains the people and the other the classes. A link between two vertices exists whenever a person belongs to a certain class.

The adjacency matrix for a bipartite network is still of size $N \times N$. However, we do not need to store the entire adjacency matrix but only the necessary information can be stored in a smaller matrix, the incidence matrix B with elements $b_{ij\,[i \in \{1,\ldots,n\}, j \in \{1,\ldots,g\}]}$ where n and g are the numbers of elements of the respective subsets.

The incidence matrix represents a two-mode bipartite. To return to a one-mode network, we can use a one-mode projection. The projection matrix, P, is formed by multiplying the incidence matrix with its transpose or vice versa. For example, if P has elements p_{ij} then P is constructed in the following way of a $g \times n$, simple, unweighted, incidence matrix B with elements b_{ij}:

$$p_{ij} = \sum_{k=1}^{g} b_{ki} b_{kj} = \sum_{k=1}^{g} b_{ik} b_{kj}, \tag{3.2}$$

or in matrix form

$$P = B^T B. \tag{3.3}$$

For this case, P is an $n \times n$ matrix where the diagonal elements indicate how many elements of the subset X of size n are connected with vertices from the subset Y of size g (which is the degree of these vertices which we will define later). The off-diagonal entries are weights, namely, the common elements between the single vertices. If one multiplies the matrix B with its transpose the following way:

$$P = BB^T, \tag{3.4}$$

then this results in a $g \times g$ matrix where the diagonal elements indicate again how many ties lead to elements in Y from elements in X, and the the off-diagonal elements represent the number of vertices in Y that are connected to the same elements in X.

3.2 Network Diagnostics

As mentioned before, the diagonal elements of a one-mode projection matrix indicate how many vertices of one subset are connected to the same element in X in the bipartite network. However, one can also just sum the entries of one row or column of an unweighted adjacency matrix for one vertex. This sum is called the degree of a vertex, and is written in terms of elements of an $N \times N$ undirected matrix A in the following way:

$$k_i = \sum_{j=1}^{N} a_{ij}. \tag{3.5}$$

For a directed network, the number of edges leading to a vertex might differ from the number leaving the vertex. Thus, for a vertex i, we define the in- and out- degree, respectively, as:

$$k_{i_{\text{in}}} = \textstyle\sum_{j=1}^{N} a_{ij} \quad \text{and} \quad k_{i_{\text{out}}} = \textstyle\sum_{j=1}^{N} a_{ji}, \tag{3.6}$$

and k_i is then the sum of $k_{i_{\text{in}}}$ and $k_{i_{\text{out}}}$. From these equations, we get the total number of edges for an undirected network:

$$2m = \sum_{i=1}^{N} k_i. \tag{3.7}$$

The mean degree $\langle k \rangle$ of a network is then defined to be

$$\langle k \rangle \ = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} k_i \ = \tfrac{2m}{N}. \tag{3.8}$$

The degree distribution[1], $p(k)$, is defined to be the frequency of the vertex degrees [2, 15]. A degree distribution that was discovered in a couple of networks is the power-law degree distribution [4]. Power-law distributions were noticed, for example for the frequency of unique words occuring in "Moby Dick" [9, 16], for the degrees of proteins in the known protein interaction network of yeast [9, 13], and the population of cities [16]. The degree distribution, $p(k)$, of a power law follows

$$p(k) = Ck^{\alpha}, \tag{3.9}$$

where C is a constant. Further, $p(k)$ has a singularity for $k \to 0$. Therefore, there must be a $k_{\min} > 0$ at which the power law behaviour stops for any k such that $k_{\min} > k \geq 0$. Thus, $k_{\min}$ is a lower bound, and C is called the normalization constant as $p(k)$ has to satisfy the following for all $k \geq 0$:

$$\sum_{k=0}^{\infty} p(k) = 1. \tag{3.10}$$

Substituting equation (3.9) into equation (3.10), we get

$$C \sum_{k=0}^{\infty} k^{-\alpha} = 1. \tag{3.11}$$

Hence,

$$\begin{aligned} C = & \quad \tfrac{1}{\sum_{k=0}^{\infty} k^{-\alpha}} \\ \simeq & \quad \tfrac{1}{\int_{k_{\min}}^{\infty} k^{-\alpha} \mathrm{d}k} = \quad (\alpha - 1) k_{\min}^{\alpha-1}, \end{aligned} \tag{3.12}$$

which implies that

$$p(k) \simeq \frac{\alpha - 1}{k_{\min}} \left(\frac{k}{k_{\min}} \right)^{-\alpha}. \tag{3.13}$$

We are interested in whether the degree distribution determined from our data fits a power-law distribution. Therefore, we need a method for parameter fitting. We will use the algorithm provided by Clauset *et al.* and the next paragraph contains the derivation of the formulas used the in the algorithm Clauset *et al.* provided [9].

Clauset used the cumulative distribution function (CDF). That is

$$P(k) = \sum_{k'=k}^{\infty} p(k'), \tag{3.14}$$

[1] The degree of a vertex is an integer value. Therefore, we are only interested in discrete distributions.

i.e. the fraction of all degrees greater than k' or the probability that a chosen vertex i has degree k' or greater $[P(k)) = \Pr(K \geq k)]$. We can approximate $P(k)$ by

$$\begin{aligned} P(k) = &\ C\textstyle\sum_{k=0}^{\infty} k'^{-\alpha} \simeq \ C\int_{k}^{\infty} k'^{-\alpha} dk' \\ = &\ \tfrac{C}{\alpha-1}k^{-(\alpha-1)} = \ C\big(\tfrac{k}{k_{\min}}\big)^{-(\alpha-1)}, \end{aligned} \tag{3.15}$$

where we assume $\alpha \geq 1$ to ensure that the series is convergent. Thus, we are left with two parameters that we need to determined from the data. First, we find an approximation for α. For this, let us assume $k_{\min}$ is known. Then, we can apply the method of maximum likelihood on $P(k)$. That is, the probability that the data has the same CDF as the theoretical distribution:

$$P(k|\alpha) = \prod_{i-1}^{M} \frac{\alpha - 1}{k_{\min}} \left(\frac{k_i}{k_{\min}}\right)^{-\alpha}, \tag{3.16}$$

where M is the size of the data sample. This probability is called the likelihood. Next, we apply the logarithm on equation (3.16):

$$\begin{aligned} L = &\ \ln(P(k|\alpha)) = \ln\big[\textstyle\prod_{i-1}^{M} \tfrac{\alpha-1}{k_{\min}}\big(\tfrac{k_i}{k_{\min}}\big)^{-\alpha}\big] \\ = &\ \textstyle\sum_{i=1}^{M} \big[\ln(\alpha-1) - \ln(k_{\min}) - \alpha\ln\big(\tfrac{k_i}{k_{\min}}\big)\big] \\ = &\ M\ln(\alpha-1) - M\ln(k_{\min}) - \alpha\textstyle\sum_{i=1}^{M}\ln\big(\tfrac{k_i}{k_{\min}}\big). \end{aligned} \tag{3.17}$$

For the maximum likelihood, we require $\mathrm{d}L/\mathrm{d}\alpha_{\max} = 0$, and solving equation (3.17) for $\alpha_{\max}$ gives

$$\bar{\alpha} = \alpha_{\max} = 1 + n\left[\sum_{i=1}^{M} \ln\frac{k_i}{k_{\min}}\right]^{-1}, \tag{3.18}$$

where the bar above α denotes the estimate of the exact solution from the data set. In order to estimate the error, we consider the exponent of the last line of equation (3.17), i.e.

$$P(k_i|\alpha) = ae^{-b\alpha}(\alpha-1)^M, \tag{3.19}$$

where $a = e^{-M\ln(k_{\min})}$ and $b = \sum_{i=1}^{M}\ln(k_i/k_{\min})$. The mean of α can then be approximated to

$$\begin{aligned} \langle\alpha\rangle \approx &\ \tfrac{\int_0^{\infty} e^{-b\alpha}(\alpha-1)^M \alpha \mathrm{d}x}{\int_0^{\infty} e^{-b\alpha}(\alpha-1)^M \mathrm{d}x} \\ = &\ \tfrac{M+1+b}{b}, \end{aligned} \tag{3.20}$$

where we used $\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}$. The second moment of α is then

$$\begin{aligned} \langle\alpha^2\rangle = &\ \tfrac{\int_0^{\infty} e^{-b\alpha}(\alpha-1)^M \alpha^2 \mathrm{d}k}{\int_0^{\infty} e^{-b\alpha}(\alpha-1)^M \mathrm{d}k} \\ = &\ \tfrac{M^2+3M+b^2+2b+2Mb+2}{b^2}. \end{aligned} \tag{3.21}$$

Thus, the variance is given by

$$\begin{aligned} \sigma^2 = &\ <\alpha^2> - <\alpha>^2 \\ = &\ \tfrac{M+1}{b^2}. \end{aligned} \tag{3.22}$$

The standard deviation is

$$\sigma = \sqrt{M+1}\Big[\sum_{i=1}^{M} \frac{k_i}{k_{\min}}\Big]^{-1}. \tag{3.23}$$

For $M >> 1$, we can approximate this to

$$\sigma \approx \frac{\alpha - 1}{\sqrt{M}}. \tag{3.24}$$

In order to find $k_{\min}$, Clauset *et al.* minimizes the distance between the CDF of the data and the best-fit model for values higher then $\bar{k}_{\min}$, i.e.

$$C = \max_{k \geq k_{\min}} |S(k) - P(k)|, \tag{3.25}$$

where $S(k)$ is the CDF using the data set for $k \geq k_{\min}$, $P(k)$ is the corresponding CDF for the power-law model that best fits the data in the region $k \geq k_{\min}$. We chose $\bar{k}_{\min}$ such that the value that D is minimized which is known as the Kolmogorov-Smirnov statistic.

Finally, in their algorithm, Clauset *et al.* also uses the p-value test that indicates whether the power-law distribution is a good fit for the data set. For this, they estimate the parameters as described before and then use these parameters to generate new data sets. Finally, p is the fraction of time whenever $D' = |S(k') - P(k')|$ is less then the original D. Further, Clauset says that for $p \approx 0.0$, one can rule out that the distribution is actually a power-law distribution. For $p \leq 0.1$ there is a moderate indication that the power-law distribution models the actual behaviour, and for $p \geq 0.1$ the power-law distribution is a good model. However, it is still not for sure that the observed degree distribution actually is a power-law distribution. Thus, the p-test is more of an indication whether an observed distribution is not a power-law distribution.

Another important concept to determine the network structure is that of a path between vertices. A path is defined to be the set of edges between two vertices, i.e. $E_{p_{i_1 i_n}} = \{\{i_1, i_2\}, \{i_2, i_3\}, ..., \{i_{n-1}, i_n\}\} \subseteq E$ and $\{i_1, i_2, ..., i_n\} \subseteq V$. The length of a path $p_{i_1 i_n}$ between vertex i_1 and i_n can be defined as $p_{i_1 i_n =}|E_{p_{i_1 i_n}}|$, i.e. the cardinality of $E_{p_{i_1 i_n}}$ [3]. A network is said to be connected if there exists a path between all pairs of vertices in the network. If a path does not exist between two vertices i and j, then we say, following the definition of Barrat *et al.* [3] that $p_{ij} = \infty$. If $E_{p_{ij}}$ is the smallest possible subset of E containing a path between vertex i and j, then the path is called a *geodesi*c path or the *shortest* path between i and j. The diameter, d, of a network is usually defined as the longest geodesic path that differs from infinity. Further, the distance matrix P is the matrix with elements p_{ij}.

A way to find the shortest path for a vertex i to other vertices is the breath-first algorithm [15]. The idea for the breath-first algorithm is the following: First, find all of the neighbours of i. These have distance d equal to one. Next, we find the neighbours of the neighbouring vertices and set $d = 2$, and so forth whereby by each iteration step, we increase d by one. We used the breath-first algorithm from Sporns [20]. For a bipartite network, the algorithm has an operation time of $O(m + N)$, where m is the number of ties, and N the cardinality of the set of vertices.

A network diadnostic that is derived from the geodesic path is the closeness centrality. For a vertex i, the closeness centrality can be defined[1] as [15]

$$c_i = \frac{1}{N-1} \sum_{\substack{j = 1 \\ j \neq i}}^{N} \frac{1}{p_{ij}}. \tag{3.26}$$

Using this definition, we can define $p_{ij} = \infty$ whenever a path between vertices i and j does not exist. Thus, a vertex gets a high closeness measure if the length of the geodesic paths from the vertex i to others is small, and has a low value whenever the geodesic path to other vertices is long. Also, this measure values vertices that have shorter paths higher then vertices that are further away from the vertex i. The computation time for the closeness centrality is of order $O(N)$ whenever the distance matrix already has been computed. Further, from this equation, the so-called harmonic mean distance $\langle p \rangle$ can be defined to be

$$\langle p \rangle = \frac{N}{\sum_{j=1}^{N} c_i}. \tag{3.27}$$

[1] This definition is not the original one that was developed by Sabidussi but a variation.

A criticism of the closeness centrality is that values tend to be spread rather closely. As an example, Newman [15] refers to the network of actors where the vertices are actors and an edge exists whenever two actors played in the same movie. The closeness centrality values hardly differ between the highest ranked actor that Newman determined to be Christopher Lee with a value of 0.4143 and the lowest ranked actress Leia Zanganeh who has a value of 0.1154. In between these two actors are, according to Newman, "about half a million other actors". This leads to big changes for networks that vary over time, i.e. where the edges slightly change, the centrality measure for the vertices might differ to a bigger extense then for example for the degree as it has a wider dynamic range [15].

Another centrality measure is the betweenness centrality. Here, the focus is more on whether a vertex lies on a geodesic path of other vertices. Thus, the betweenness centrality for a vertex i is given by [6, 15]

$$b_i = \frac{\sum_{st} n_{st}(i)}{n_{st}}, \tag{3.28}$$

where n_{st}^i is 1 if vertex i lies on a geodesic path and zero otherwise and n_{st} is the number of all shortest paths from s to t between s and t. To calculate betweenness centrality, we used algorithm from Rubinov [19] that is based on the algorithm proposed by Brandes [6]. This algorithm has an operation time of $O(N(m+N))$ for unweighted graphs. To understand the algorithm, we summarize the algorithm proposed by Brandes. First, he defines the pair dependency for a pair of vertices s and t to be

$$\delta_{st}(i) = \frac{n_{st}(i)}{n_{st}}, \tag{3.29}$$

i.e. the probability that a path from s to t goes through i. Then he defines the set of predecessors of a vertex i on shortest paths from a vertex s to be:

$$P_s(i) = \{j \in V | \{j,i\} \in E, d_g(s,i) = d_g(s,j) + \omega(j,i)\}, \tag{3.30}$$

where $\omega(j,i) = 1$ if $\{j,i\} \in E$ and equals 0 otherwise. Now, note that $\delta_{st}^i > 0$ only for $t \in V/\{s\}$ whenever v lies in the shortest path between s and t. Further, there exists exactly one edge $\{i,j\}$ with $i \in P_s(j)$. He then extense the pair-dependency also to edges. That is

$$\delta_{st}(i,e) = \frac{n_{st}(i,e)}{n_{st}}, \tag{3.31}$$

where $n_{st}(i,e)$ is the number of shortest paths from s to t that contain not only i but also $e \in E$. Thus, $\delta_{st}(i,e)$ is the probability that the path from s to t goes through i

and over e. Thus, we can say that the probability of all shortest paths starting from s going through v is

$$\begin{aligned} \delta_{s.} = & \sum_{t\in V} \delta_{st}(i) = & \sum_{t\in V} \sum_{j|i\in P_s(j)} \delta_{st}(i,\{i,j\}) \\ & = & \sum_{j|i\in P_s(j)} \sum_{t\in V} \delta_{st}(i,\{i,j\}). \end{aligned} \tag{3.32}$$

Then, because there is an edge from i to j, there are n_{si} paths that first go through i and then through j. Thus, the total number of paths going from s to $t \neq j$ containing i and $\{i,j\}$ is

$$\frac{n_{si}}{n_{sj}} n_{st}(j). \tag{3.33}$$

Hence,

$$\delta_{st}(v,\{i,j\}) = \begin{cases} \frac{n_{si}}{n_{sj}} & \text{if } t = j \\ \frac{n_{si}}{n_{sj}} \frac{n_{st}(j)}{n_{st}} & \text{if } t \neq j \end{cases}, \tag{3.34}$$

and $\delta_{s.}(i)$ can be written as

$$\begin{aligned} \delta_{s.}(i) = & \sum_{j|i\in P_s(j)} \sum_{t\in V} \delta_{st}(i,\{i,j\}) \\ = & \sum_{j|i\in P_s(j)} \frac{n_{si}}{n_{sj}} + \sum_{t\in V/\{j\}} \frac{n_{si}}{n_{sj}} \frac{n_{st}(j)}{n_{st}} \\ = & \sum_{i|j\in P_i(v)} \frac{\sigma_{ij}}{\sigma_{iv}} (1 + \delta_{s.}(v)). \end{aligned} \tag{3.35}$$

The dependency for a single vertex can be computed in $O(m)$ whenever the shortest path from a vertex i to all others is known. To compute the shortest paths for all vertices using the breath first algorithm costs $O(N + m)$. Thus, the total algorithm to compute the betweenness for all vertices has an operation time of $O(N(N + m))$.

For some networks, beside the ties connecting vertices, the vertices also have extra information assigned to them. For example, the vertices might be a set of people and as a characteristic (or class) the age of the people might be given. This is of importance for a network diagnostic as it can be observed occasionally that vertices have more ties with vertices having the same characteristic which is called *homophily* or *assortative mixing*. The opposite, that is vertices rather connect to vertices with an unlike characteristic, is called *disassortative mixing*. In mathematical terms, the characteristics of one class can be written as a set, i.e. let $c_i \in \{b_1, b_2, ..., b_n\}$ be a characteristic of vertex i where b_s is an element of the class of characteristics. The modularity, Q, measures whether vertices were more inclined to connect to other vertices of the same class or tend to show disassortative mixing.

To derive the modularity, first consider the total number of edges that exist between vertices of the same type, that is

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\delta(c_i, c_j), \tag{3.36}$$

where $\delta(c_i, c_j)$ is the Kronecker symbol, $1/2$ needs to be multiplied because edges are counted twice. The total expected number of edges joining vertices of the same type is the sum of all edges of the same type that might form a link with the edges of other vertices of the same type, i.e.

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{k_i k_j}{2m}\delta(c_i, c_j). \tag{3.37}$$

Now, Q is defined to be the difference between the actual number and the expected number of edges that join vertices of the same class divided by the total number of edges in the network:

$$Q = \frac{1}{2m}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} - \frac{k_i k_j}{2m}\delta(c_i, c_j). \tag{3.38}$$

The modularity, Q, is the diagnostic of whether vertices are connected to other vertices with the same class. Positive values of Q are related to assortative mixing by the given class whereas negative ones indicate a disassortative mixing. Newman developed a method to find Q for two groups. To determine Q, first consider the case when there are just two groups, say, b_1 and b_2. Then, we can set $b_1 = 1$ and $b_2 = -1$. Thus, $c_i = 1$ or $c_i = -1$ for all $i \in V$ such that $\delta(c_i, c_j) = 1/2(c_i c_j + 1)$. Hence, we can write

$$Q = \tfrac{1}{2m}\textstyle\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} - \tfrac{k_i k_j}{2m}\left(\tfrac{1}{2}c_i c_j + 1\right). \tag{3.39}$$

Note that

$$\begin{aligned} & \textstyle\sum_{j=1}^{N} a_{ij} - \frac{k_i}{2m}\sum_{j=1}^{N} k_j \\ = \quad & k_i - \tfrac{k_i}{2m}2m \\ = \quad & 0. \end{aligned} \tag{3.40}$$

Say, $d_{ij} = a_{ij} - (k_i k_j)/(2m)$ then

$$\begin{aligned} Q = \quad & \tfrac{1}{2m}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}(c_i c_j + 1) \\ = \quad & \tfrac{1}{4m}\tfrac{1}{2m}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} c_i c_j, \end{aligned} \tag{3.41}$$

or in matrix form

$$Q = \frac{1}{4m} c^T D c. \tag{3.42}$$

3.3 Computation of Network Diagnostics

There are different ways of constructing networks out of the data given by FirstGiving. An obvious way is to use bipartite networks. The vertices are the donors and the classes to which they belonged are formed from the charities. That is, whenever a donation from a person i to a charity j was recorded, we set $b_{ij} = 1$. Because of computation time and also memory space, we had to aggregate the data. We decided that an interesting way would be to look at different time slots and see how the network structure developed over time.

Thus, the entire donor-charity network is aggregated in monthly and daily donations. For some computations of network diagnostics, the size for monthly aggregated network was still quite large. Thus, for this case, we further aggregated the network for a day only. In particular, we will look at the donation behaviour in November, and the first four Mondays and Saturdays of November of each year starting from 2003 until 2010.

The degree is an important basis for other network measures. But also, from the degree, we can tell a bit about the structure of the network. Figures 3.2 and Figure 3.4 show the mean degree of the donor vertices and the charity vertices, respectively.

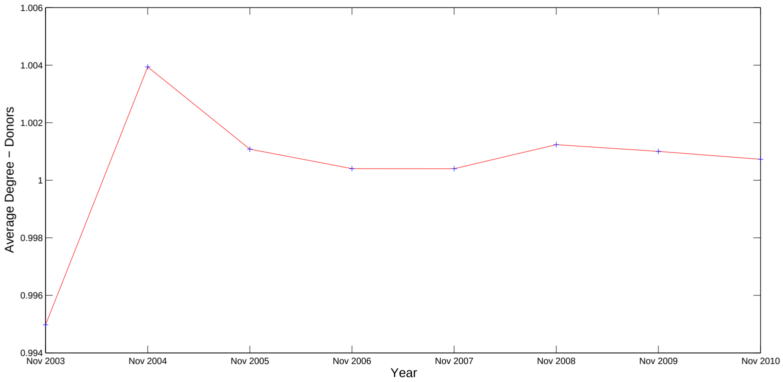

Figure 3.2: Mean degree of donors in November.

2003 can be disregarded, as that was the setup year of FirstGiving. Interestingly, both the mean degree for the set of charities and the set of donors stay approximately the same from 2004 to 2010. Thus, the charities have a mean degree of 21.7114 and the donors have a mean degree of 1.0005.

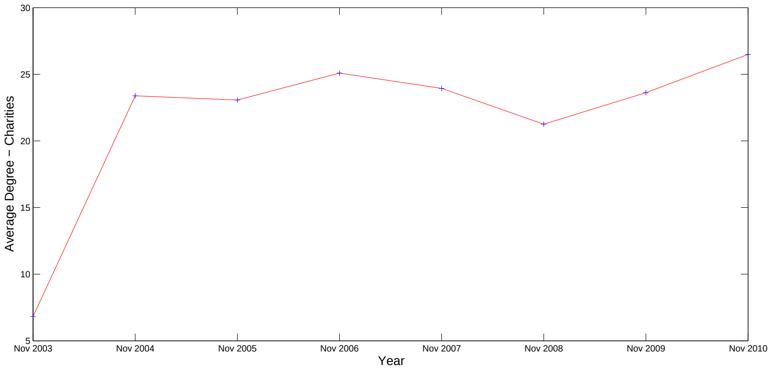

Figure 3.3: Mean degree of charities in November.

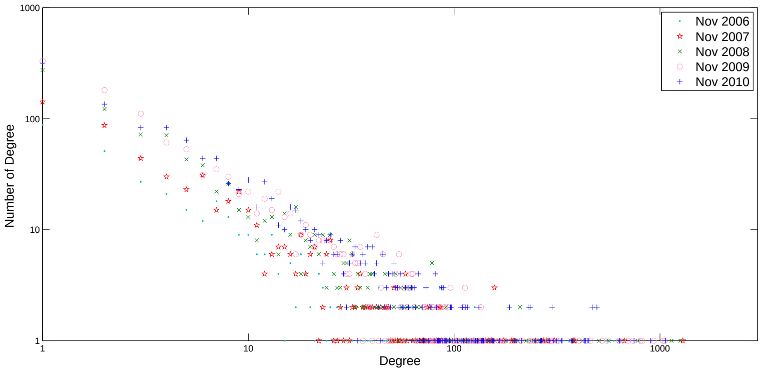

Figure 3.4: Mean degree of donors in November of various years.

However, when looking at each year individually (Figure 3.4), the degree distribution might indicated a power law distribution. (Figure 3.4 is has both axes in log scale. Thus, for a power law distribution we would expect a straight line.)

We applied the algorithm from Clauset *et al.* as explained above to the yearly values and also performed a fit and p-value test for exponential and Gaussian distribution using built-in Matlab functions[2]. The results can be seen in Table (3.1) and Table (3.2).

[2] *normfit* and *expfit*

	Nov 2006	Nov 2007	Nov 2008	Nov 009	Nov 2010
p_G	0.000	0.0000	0.0000	0.0000	0.0000
p_E	0.000	0.0000	0.0000	0.0000	0.0000
p_{PL}	0.0124	0.0000	0.9004	0.7444	0.0500

Table 3.1: p-value for power-law (PL), exponential (E) and Gaussian distribution (G) determined from 2500 trials.

	Nov 2006	Nov 2007	Nov 2008	2Nov 009	Nov 2010
α	1.7400	1.6700	2.2500	2.2500	2.2500
g	396	626	991	1269	1298
σ	0.0372	0.0268	0.0397	0.0351	0.0339
x_{min}	7	5	49	51	50

Table 3.2: Power - law exponent (α), size of the subset of charity vertices (g), the standard derivation (σ) and x_{min}.

Clearly, the values indicate that the degree distribution of the charities cannot be model with the exponential or the Gaussian distribution. For the power-law distribution, November 2006 and November 2007 have low p-values. For November 2008, November 2009 and November 2010, the power-law distribution cannot be ruled out. Also note that the α values for these years are about the same with a similar value for k_{min}. In order to say more, i.e. whether it reached a saturation state. However, the p-values are not enough to ensure that the distribution really follows the power-law but we can tell from these values that the distribution is not Gaussian or exponential.

The degree distribution of the donors is shown in Table (3.3). Table (3.3) shows the fraction of donors with $p(k = 1)/m$ and $p_{(k \geq 2)}/m$. Certainly, the majority of the donors donated once, and only a couple of people donated more than once in a month.

	Nov 2006	Nov 2007	Nov 2008	Nov 2009	Nov 2010
$p_{(k=1)}/m$	0.9995	0.9996	0.9990	0.9992	0.9992
$p_{(k \geq 2)}/m$	7.0451e-004	5.3383e-004	9.0308e-004	8.0139e-004	7.2757e-004

Table 3.3: Fraction of donors with a degree greater or equal to 2.

We also determined the shortest path between all vertices in the system. As the operation time for the shortest path is $O(m + g + n)$, we decided to aggregate the network once more, and choose as the set of vertices the donors and charities active Mondays in November of each year. Then, we determined the closeness diagnostic and determined the mean for the first four Mondays of each year seperatly to, finally, determine the mean shortest path for the first four Mondays of each year. The mean closeness diagnostic for each year can be seen in Table 3.4. We can see that the highest value for $\langle c \rangle$ for both donors and charities occur in November 2008.

	Nov 2006	Nov 2007	Nov 2008	Nov 2009	Nov 2010
$\langle c_{\text{Charity}} \rangle$	0.0072	0.0052	0.0888	0.0027	0.0512
$\langle c_{\text{Donor}} \rangle$	0.0129	0.0105	0.1506	0.0095	0.0544

Table 3.4: The mean closeness centrality of the first four Mondays in November 2006 - 2010 for charities and donors.

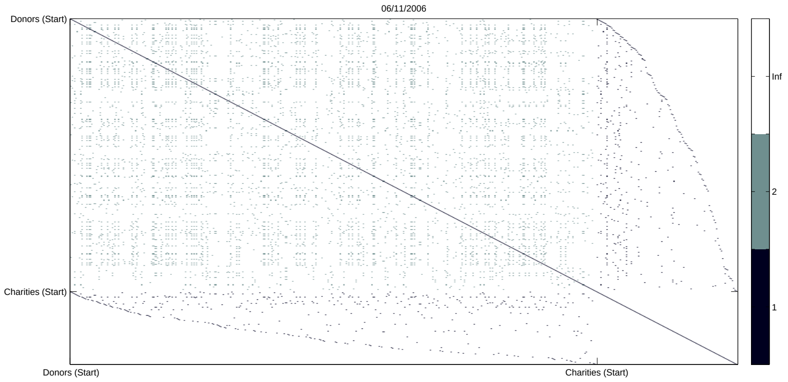

Figure 3.5: Distance matrix of 06/11/2006.

To further investigate this, we plotted the distance matrix P for indiviual days. Figure 3.5 shows the distance matrix P for 06/11/2006. It has a diameter of 2. That is none of the donors donated twice. Therefore, the donors either have a path of zero, two or ∞. However, charities have a degree of zero or one. Thus, for a subgraph, such that we consider one charity and the donors donating to this charity only, the structure is star graph with the centre being the charity vertex. This also explains the low value for the closeness measure for both the charities and the donors in 2006. This structure is depicted in Figure 3.6.

We can see a rather high value for the closeness for charities and donors in 2008. Figure 3.7 shows the distance matrix for 03/11/2008. The diameter here is 14. Two donors donated to three different charities and 944 donors donated to two different charities. The original bipartite network therefore has a mean degree for the donor

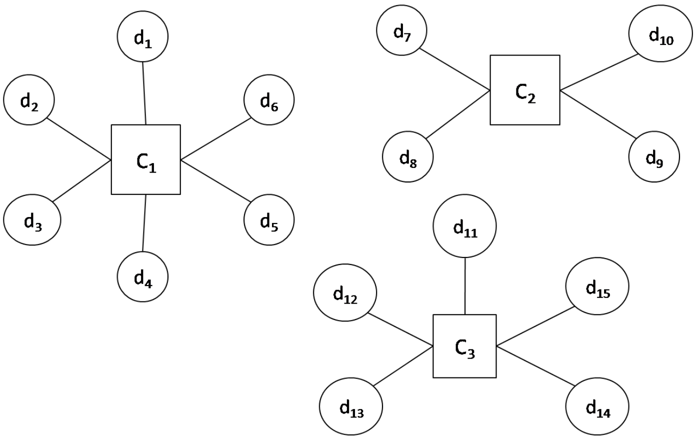

Figure 3.6: Schematic diagram of star-graph structure where C_i indicates charity elements and d_i donor vertices.

vertices of 1.8103. Therefore, out of $982081 = (n+g)^2$ possible entries of the distance matrix only 1449 entries of P are equal to infinity. The change in the network structure is caused by the donors who donated more then once. These donors form bridges connecting people donating to one charity. This structure is depicted in Figure 3.8 where d_5 and d_6 form bridges.

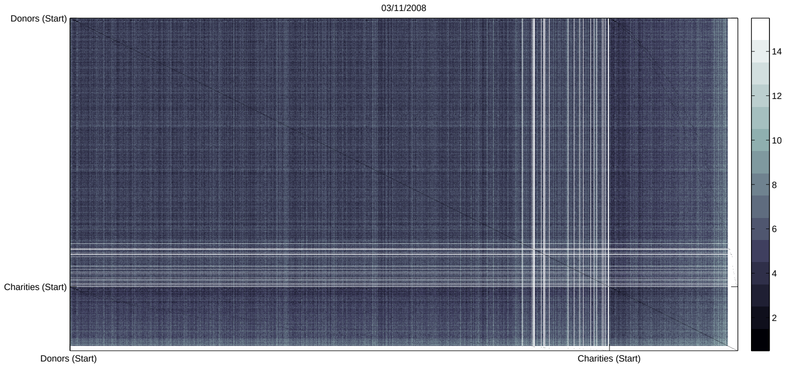

Figure 3.7: Distance matrix of 03/11/2008.

To summerize, Table 3.5 shows the mean path length of each network containing the networks formed from the donation information of the first four Mondays in November. We can see that if the maximal degree for donors is greater then 1, then the diameter of the network increases and the mean path length gets smaller. Certainly, we can see from the data collected from the different network structure extracted from different Mondays in November that slight changes in the structure lead to big

	06/11/2006	13/11/2006	20/11/2006	27/11/2006
$< p >$ - donors	71.1988	91.1894	89.1072	64.7641
$< p >$ - charities	147.4896	131.5063	154.1202	125.0000
diameter	2	2	2	2
k_{max} - donor	1	1	1	1
	05/11/2007	12/11/2007	19/11/2007	26/11/2007
$< p >$ - donors	126.4989	112.1553	89.1072	75.0637
$< p >$ - charities	220.7381	181.1511	154.1202	181.6006
diameter	2	2	2	2
k_{max} - donor	1	1	1	1
	03/11/2008	10/11/2008	17/11/2008	24/11/2008
$< p >$ - donors	5.0341	99.6535	120.4093	5.1249
$< p >$ - charities	5.6841	283.1490	293.4885	5.8076
diameter	14	2	4	12
k_{max} - donor	3	2	4	2
	02/11/2009	09/11/2009	16/11/2009	23/11/2009
$< p >$ - donors	102.3093	66.3854	141.3674	171.0805
$< p >$ - charities	431.7321	378.8514	385.3848	369.7956
diameter	2	2	2	2
k_{max} - donor	2	1	1	1
	01/11/2010	08/11/2010	15/11/2010	22/11/2010
$< p >$ - donors	4.0144	169.2461	169.2461	192.9037
$< p >$ - charities	4.0684	468.7028	423.6522	293.4885
diameter	2	2	2	2
k_{max} - donor	2	1	1	1

Table 3.5: Mean path length, $< p >$, diameter and maximal degree, $k_{\max}$ for the aggregated, bipartite networks of various years.

differences in the mean path length and therefore closeness centrality for the single vertices.

We also calculated the mean betweenness for the first four Mondays of November from $2006 - 2010$ (Table 3.6). Not surprisingly, in 2008 both the average betweenness for donors and charities is the highest as in 2008 the bipartite charity-donor network has the most multiple donors over the years. Note, that for the 2006, 2007 and 2009 the betweenness centrality for the star subgraphs with the centre being a charity vertex and the donor vertices attached around the centre, for this subgraph the charity vertices have a betweenness value of one. For the entire network, this clearly differs as then we have to take more vertices into account.

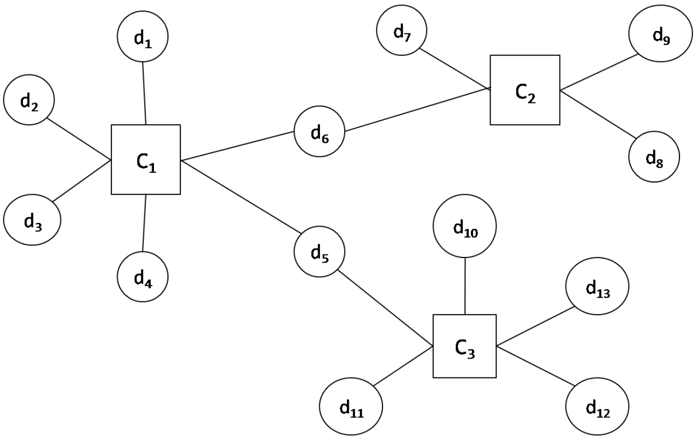

Figure 3.8: Schematic diagram of star-graph structure with bridges where C_i indicates charity elements and d_i donor vertices. Further, d_5 and d_6 form the bridges in this network.

	Nov 2006	Nov 2007	Nov 2008	Nov 2009	Nov 2010
$\langle b_{\text{Charity}} \rangle$	0.0000	0.0000	0.0888	0.0000	0.0002
$\langle b_{\text{Donor}} \rangle$	0.0029	0.0025	0.1140	0.0014	0.0525

Table 3.6: The mean betweenness centrality of the first four Mondays in November 2006 - 2010 for charities and donors.

Finally, we determined the modularity for age, household income and state residence for the donors in November from 2003 until 2007. As it is of interest to see whether the donors have similar characteristics, we wanted to determine the modularity for donors donating to the same charity. For this, we first computed the one-mode projection for the donor vertices from the bipartite network. However, when computing the modularity, we ignored the diagonal elements because these elements indicate the degree for a donor. But also then the one-mode projection can be regarded as an adjacency matrix where the vertices are the donors and ties exists when two donors donated to the same charity. Table 3.7 shows the results. Note, we did not differ between a person from living inside the US or any other country. Further, for some donors no personal details where known. Thus, we disregarded these donors. As one can see in Table 3.7, there is a tendency for the donors that donated to the same charity to also have similar characteristics. That is people in the same age group also donate to the same charity. For the states, we see that people tend to donate to the same charity, if they also are residence of the same state. With the highest value in November 2003. The modularity for the state decreases for the next years until 2007. From 2007 onwards, we can see that the $Q_{\text{state}} \approx 0.02$. We already mentioned that 2003 was the founding year of FirstGiving. Thus, it can to be expected that the people setting up the company started with advertising it in areas close by such as Massachusetts. Note, that in Figure 2.2, we saw that Massachusetts has one of the highest fractions of its population donating via FirstGiving. Still, this is just a hypothesis and needs further investigation. The same holds for the modularity values for the age and household income. These level around 0.015 for Q_{age} and 0.045 for Q_H.

	Nov 2003	Nov 2004	Nov 2005	Nov 2006	Nov 2007	Nov 2008	Nov 2009	Nov 2010
Q_{age}	0.0165	0.0147	0.0138	0.0147	0.0179	0.0164	0.0164	0.0239
Q_{state}	0.0563	0.0269	0.0168	0.0138	0.0239	0.0239	0.0202	0.0205
Q_{H}	0.0397	0.0419	0.0464	0.0430	0.0482	0.0467	0.0497	0.0510

Table 3.7: Modularity for one-mode charity matrix for age (Q_{age}), state (Q_{state}) and the household income (Q_{H}).

In conclusion, we have seen from the degree distributions that donors are more inclined to donate to charities once than multiple times in a time period. Further, the distribution of charities most likely does not follow a power law distribution but leans in the direction of a power-law distribution. The closeness and betweenness centralities also indicate that donors donate only once. However, both have highly varying values whenever donors donate more than once and the network structure changes from having many star subgraphs with charities in the centre. The modularity values for state residency, age and household income suggests that donors with similar characteristics also donate to the same charity.

Chapter 4

Collective-Behaviour Model

4.1 The SI Model

In this chapter, we model the donation behaviour of a set of people. For this, we will use a similar approach that has been used to model epidemics spreading through a population where the members of a population again represent nodes. These models are called *agent-based models.* For these kind of models, one considers the interaction of each agent (i.e. in the case of the epidemics, the agents are people) of the system separately and the model is a simulation imitating the interaction between the autonomous agents [3].

We are interested in the change of configuration of the single agents, the elements of a system V. For a set of states $\{s_1, s_2, ..., s_n\}$ an agent can be in, where n is the number of states. Further, we denote $\rho_i \in \{s_1, s_2, ..., s_n\}$ to be the configuration of $i \in V$, i.e. the state assigned to $i \in V$. The total configuration of a system is then $\rho(t) = (\rho_1(t)), \rho_2(t), ..., \rho_N(t)$ where N denotes the number of agents in the system. The master equation is used to predict the likelihood of a change from one configuration ρ' to another ρ using the transition probabilities $W(\rho' \to \rho)$ for the various states, i.e. [3]:

$$\frac{\mathrm{d}P(\rho, t)}{\mathrm{d}t} = \sum_{\rho'} [P(\rho', t)W(\rho' \to \rho) - P(\rho)W(\rho \to \rho')], \tag{4.1}$$

where the sum runs over all possible configurations ρ' a system can be in and $W(\rho' \to \rho) = \prod_i \omega(\rho'_i \to \rho_i)$, i.e. we consider the probability for each node i to change its configuration from ρ'_i to ρ_i individually as the agents are independent of each other. Usually this equation cannot be solved and simplifications need to be made. A typical approximation of the Master equation is the mean-field assumption. That is, all elements i in the system have the same properties and the likelihood to change from one state to another is the same for all elements, i.e. independent of i.

This approach is also used in epidemic models. The simplest model is the susceptible/infected (SI) model. The SI model simulates the epidemic spread when a virus is transmitted between hosts that are infected by a disease, and susceptible people that can gain the disease by meeting any infected [15]. Thus, we have two states in which each member of the population can be: *susceptible* and *infected.* In the first instance, we consider the fully mixed case ignoring any underlying network structure but consider the case whenever the chance of individuals to get the disease, per unit time, is the same for all individuals. If we have a population V of N members, than let $S(t)$ be the number of members that are susceptible and $X(t)$ the number of people that are infected at time t (following the notation used by Newman [15]).

We then assume, using the mean-field approximation, that the number of people that individuals meet per unit time is the same for all members of the population. Say, this number is β. The average probability to meet someone in the susceptible state is S/N. Thus, we can say that an infected person has on average contact with $\beta S/N$ people per unit time. We also know that in the total population there are X infected people. Thus, the average rate for new infected people can be written as

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \frac{\beta SX}{N}, \tag{4.2}$$

and the change of susceptible people in the population is

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta SX}{N}. \tag{4.3}$$

Further, $s(t) = S(t)/N$ is the normalised number of susceptible and $x(t) = X(t)/N$ is the normalised number of infected. Note, that $s(t)$ and $x(t)$ can also be seen as the probability of an individual to be in the susceptible and infected state, respectively. So the simplified approximation to the master equation for the probability of a member of the populations to change into the infected state is then given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \beta sx, \tag{4.4}$$

and for the susceptible state:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\beta sx. \tag{4.5}$$

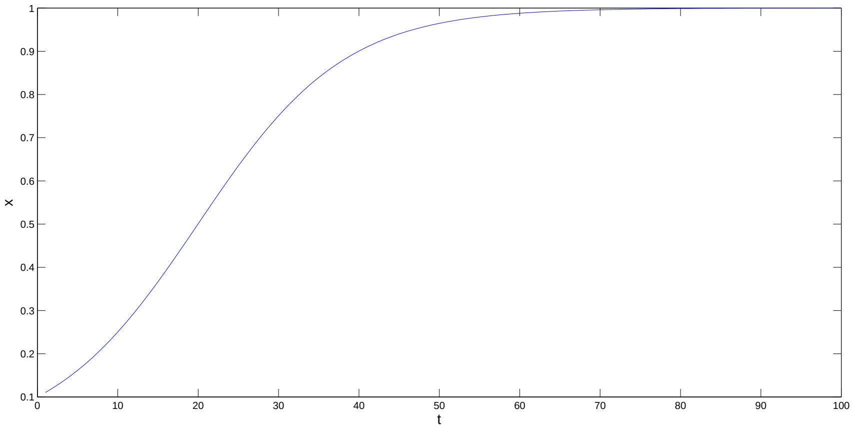

Figure 4.1: The logistic equation with $\beta = 0.11$, $x_0 = 0.1$.

Finally, note that $s + x = 1$. Thus, we can express equation (4.4) in the following way:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \beta(1-x)x. \tag{4.6}$$

For x_0 at $t = 0$, the solution of the above is then

$$x(t) = \frac{x_0 e^{\beta t}}{1 - x_0 + x_0 e^{\beta t}}, \tag{4.7}$$

that is the well-studied logistic growth equation. As seen in Figure (4.1), the graph of the logistic equation is of S-shaped form.

However, a person does not meet everybody in the world with the same probability but rather has a circle of contacts. This circle of contacts can be represented using a friendship network, with an adjacency matrix $A = \{a_{ij}\}_{i=\{1,\ldots,N\},j=\{1,\ldots,N\}}$. Then, the probability for each person to get infected depends on the people which whom the person comes into contact. Losing the approximation that everybody has the same chance of getting infected implies that we now have to consider the state for each individual separately. Thus, say, if a person i is still susceptible then $s_i = 1$ otherwise $s_i = 0$. Similar, if a person i is infected then $x_i = 1$ otherwise $x_i = 0$.

We need to incorporate the ties leading to a vertex symbolizing a person to other vertices in a friendship network. The ties represent the circle of contacts a person is able to meet. That is, the probability for a vertex i to meet someone who is infected is $\sum_{a_{ij} \in V_i} a_{ij} x_j$ where V_i is the subset of vertices to which i is connected. Note, that we still use the approximation that the transition probability β is the same for all vertices ignoring effects such as differences in the immune system, etc. Thus, the change in probability of being susceptible in time for a vertex i is given by

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\beta s_i \sum_{a_{ij} \in V_i} a_{ij} x_j, \tag{4.8}$$

i.e. the rate of change β followed by the probability of finding i in the state s_i and the probability of one of i's neighbouring vertices to be infected. Similarly, the probability change of a vertex i being infected is then

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \beta s_i \sum_{a_{ij} \in V_i} a_{ij} x_j. \tag{4.9}$$

4.2 A Threshold Model

The concept of contagion can be extended beyond epidemics but also to any spread of a dynamical process in social interaction such as collective-behaviour [3]. For the collective-behaviour simulation, we also included a *threshold model*. People might consider the "cost" and "benefits" an action would persuade against each other [8, 12, 18]. For certain situations, the benefits and costs of one person depends on the behaviour of others. Hence, one might include a threshold to model the behaviour when a person considers to commit an action incorperating the idea of whether an action is beneficial or harming. As such situations, Granovetter includes, amoung others, "diffusion of innovations", "rumours and disease", "strikes", "voting", and "educational attainment". However, also social norms and movements belong to the list of "social contagion" [8]. In general, in all the different cases, the decision to use a new product, believe a rumour, attend a strike, vote for a certain party, or attend college depend on the decisions of one's peers. That is, one is more likely to use a new innovation whenever others are using it as well, or believe in a rumour if there exists more than one source [8, 18].

For example, Rogers [18] points to a study conducted in the late 1950s and early 1960s that observed the introduction of a new maths syllabus called "New Mathematics" at schools in Pennsylvania and West Virginia. He notes that six school superintendents were needed until other schools in the area followed to introduce the "New Mathematics" syllabus (however that the syllabus changed again at the end of the 1960s [18].)

Another example Rogers mentions is the study of the introduction of birth control in the 1960s and 1970s. The data indicates that the threshold for a Korean woman to participate in birth control depended on their husband's opinion but also on her age, education, and status in the villages [18].

There are two different cases for thresholds, as Centola [8] points out: In the first case, the threshold depends on the number of people committing an action. This was introduced by Grenovetter. However, Watts [23] introduced another concept. Here, the threshold depends on a fraction of the population. Thus, the threshold does not change with increasing or decreasing size of the population. For the *fraction* threshold, Centola explains that all non-participators are taking into account as well. Whereas, as the *number* threshold, only the participants play an active role. Which threshold notation is better to use depends on the problem. As an example for the *fraction* threshold contagion, Centola lists the example of littering in a neighbourhood and disease spread. For the littering, it might stop with increasing neighbourhood size even if the same amount of people place their rubbish in the environment. However, for a highly infectious disease where the threshold is one person, the transition behaviour will not change by increasing population - still only one person is needed to distribute the disease to the next one [8].

Finally, we will adopt Centalo's notation of for the difference whenever there is only one person needed to spread the contagion, i.e. the threshold is one. Then, he calls the process *simple contagion.* However, if more than just one person or more than a fraction is needed to trigger contagion than he calls this is *complex contagion.*

4.3 Collective-Behaviour Model for Charity Donation

As outlined in the Section 3.3, anybody wanting to use FirstGiving creates a web page with a cause and then distributes the address to acquaintances, friends and family. Thus, we have to differentiate between fundraisers that used FirstGiving and donors that used the web page they received from a fundraiser. For the fundraiser, the idea of a complex contagion might apply as FirstGiving distributes a product, namely, the tool to construct web pages and the simplification of the money transfer to a charity. The knowledge of the new product seems to follow a similar line as outlined by Granovetter in his threshold model of collective behaviour [12]. FirstGiving introduced a new product that customers in the form of fundraisers and donors might need recommendation before actually using it. For people to use the tools provided by FirstGiving, we therefore consider that people also need to have knowledge about the company via acquaintances, family and friends before they would first become fundraisers themselves. We further assume that complex contagaion changes to simple contagaion, as we assume, people then have their own experience, and do not rely on others opinions. Thus, they might choose to collect money via FirstGiving again despite the opinion of others. Further, we will also use a fraction threshold model as the size of the set of contacts varies for each person. This seems to be a plausible choice as each donor has a varying circle of contacts. Also, for reasons of simplicity, we like to make the rough assumption that the treshold is the same for each person. Further, a person with a low number of friends might not be able to gain a treshold bigger than a certain number (for the case of the number threshold). Therefore, we decided to use the fraction treshold.

Again as outlined in Chapter 2, FirstGiving aims at people that have a "private" cause to fundraise money. That is, people can choose if they want to donate because of a special day in their lives, i.e. rather than a birthday or wedding present they would like friends to donate money to charity. This suggests that it is a reasonable to assume that people donating via a fundraiser do so because they know this person and otherwise would not have chosen to use the web page set up by the fundraiser to donate money to their charity of choice. Note that this assumption excludes whenever people use the search function to donate money via FirstGiving at their web page. Ignoring people donating via FirstGiving directly, we again have a form of a threshold model. However, this time we do not want to use the fraction treshold but rather we consider that donating money via a fundraiser might be an act of kindness towards the fundraiser, and rather it is seen as a favour from the donor to the fundraiser. Hence, we we consider simple contagion as just one person triggers a possible donation.

We also take into consideration that a donor or fundraiser most likely needs some time before he donates money again. We call this period "recovery" time. In epidemic models, this is the time a person is immune against a disease. It is usually denoted as the period of recovery.

Thus, we say, a person i can be in five different states. Either, he or she never heard of FirstGiving and is susceptible (n_i). Next, he can be in a fundraiser state (f_i), or a donor state (d_i). That state is followed by a recovery time whereby we denote the state by r_i. Finally, as the person i after the recovery does not rely on the opinions of others anymore to use FirstGiving again, we also introduced another susceptible state: s_i.

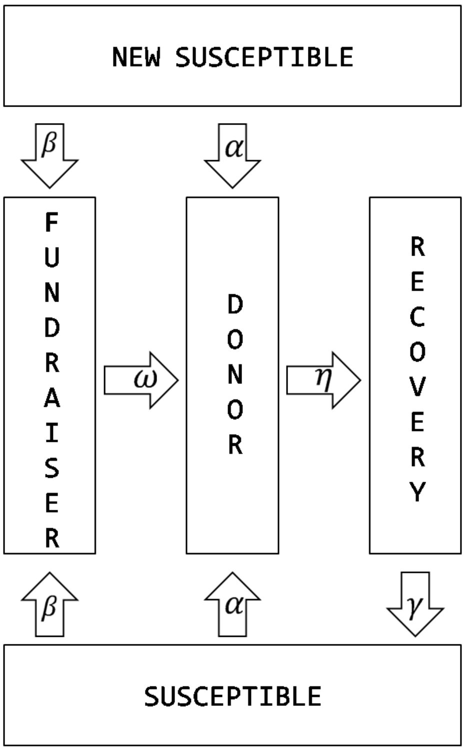

Figure 4.2: Schematic showing the transition from one state to another.

We also have to introduce the following transition probabilities from one state to another: β is the transition probability to become a fundraiser, f, from either the new susceptible state, n, the susceptible state where people have not donated via FirstGiving yet, or susceptbile state, s, where people have donated at least once. Further, α is the transition probability to become a donor from either n, or s, η is said to be the transition probability for people to change into the recovery state, r, while before being a donor, and γ is the transition probability to become a susceptible, s, from being in the recovery state, r. Finally, we have to define a transition probability ω for the change from fundraiser, f, to donor, d. Note it certainly is possible that a fundraiser does not donate money to a charity. We ignore this case mainly as it seems to be plausible that if someone makes the effort to generate a page this person also donates money to that charity.

We need to introduce a threshold function, F_i for a person i for complex contagion.

$$F_i(f_j, d_j, \epsilon) = \begin{cases} \frac{\sum_j a_{ij}(f_j+d_j)}{k_i} & \text{if } \frac{\sum_{a_{ij}\in V_i} a_{ij}(f_j+d_j)}{k_i} > \epsilon \\ 0 & \text{otherwise} \end{cases}, \tag{4.10}$$

where V_i is the set of neighbours of i, k_i is the degree of i, and ϵ is the threshold. Note, that this function also includes the varying probability depending on the fraction of neighbours donating. That is, not only does one need to have a certain fraction of friends before one is willing to donate but also the bigger the fraction of friends is above the threshold limit donate or fundraise the more likely does one become a fundraiser for the first time oneself. This might be because the more positive "reviews" one has about a new product the more inclined one is to use it as well.

The next threshold function is for simple contagion:

$$G_i(f_j) = \begin{cases} 1 & \text{if } \sum_{a_{ij}\in V_i} a_{ij} f_j > 1 \\ 0 & \text{otherwise} \end{cases}, \tag{4.11}$$

Here, one person is enough to trigger a potential donation for reasons explained above.

In the model, we further assume that the number of people in the population does not change. We call this number N. Further, as mentioned before, there is no difference between people. That is, we will ignore any characteristics that might also influence the donation behaviour such as income, age, and varying thresholds, etc. but say that all people are equal (except of the number of contacts). Thus, we end up with the following five ODEs:

$$\begin{aligned} \frac{\mathrm{d}n_i}{\mathrm{d}t} &= -\alpha n_i \mathrm{G}_i(f_j) - \beta n_i \mathrm{F}_i(f_j, d_j, \epsilon), \\ \frac{\mathrm{d}f_i}{\mathrm{d}t} &= -\omega f_i + \beta s_i + \beta n_i \mathrm{F}_i(f_j, d_j, \epsilon), \\ \frac{\mathrm{d}d_i}{\mathrm{d}t} &= \alpha(n_i + s_i)\mathrm{G}_i(f_j) + \omega f_i - \eta d_i, \\ \frac{\mathrm{d}r_i}{\mathrm{d}t} &= \eta d_i - \gamma r_i, \\ \frac{\mathrm{d}s_i}{\mathrm{d}t} &= \gamma r_i - \alpha s_i \mathrm{G}_i(f_j) - \beta s_i, \end{aligned} \tag{4.12}$$

for $i = 1, 2, ..., N$ and, note, that $n_i + f_i + d_i + r_i + s_i = 1$. This is a set of five ODEs with five unknown parameters and six unknown variables.

First, we will non-dimensionalize the above system of equations where we use that $\hat{t} = \beta t$. Then equations (4.12) can be written:

$$\begin{aligned}
\tfrac{\mathrm{d}n_i}{\mathrm{d}\hat{t}} &= -\hat{\alpha} n_i \mathrm{G}_i(f_j) - n_i \mathrm{F}_i(f_j, d_j, \epsilon), \\
\tfrac{\mathrm{d}f_i}{\mathrm{d}\hat{t}} &= -\hat{\omega} f_i + s_i + n_i \mathrm{F}_i(f_j, d_j, \epsilon), \\
\tfrac{\mathrm{d}d_i}{\mathrm{d}\hat{t}} &= \hat{\alpha}(n_i + s_i)\mathrm{G}_i(f_j) + \hat{\omega} f_i - \hat{\eta} d_i, \\
\tfrac{\mathrm{d}r_i}{\mathrm{d}\hat{t}} &= \hat{\eta} d_i - \hat{\gamma} r_i, \\
\tfrac{\mathrm{d}s_i}{\mathrm{d}\hat{t}} &= \hat{\gamma} r_i - \hat{\alpha} s_i \mathrm{G}_i(f_j) - s_i,
\end{aligned} \tag{4.13}$$

where $\hat{\alpha} = \alpha/\beta$, $\hat{\gamma} = \gamma/\beta$, and $\hat{\eta} = \eta/\beta$ [14]. Dropping the hats, we determine the equilibrium points for equations (4.13). That is, we set $\mathrm{d}n_i/\mathrm{d}t = \mathrm{d}f_i/\mathrm{d}t = \mathrm{d}d_i/\mathrm{d}t = \mathrm{d}r_i/\mathrm{d}t = \mathrm{d}s_i/\mathrm{d}t = 0$. Then, we can re-arrange the first equation of (4.13) to

$$0 = (-\alpha x - y)n_i, \tag{4.14}$$

for all $i \in V$, where $x = G_i(f_j)$ and $y = F_i(f_j, d_j, \epsilon)$. Thus, assuming $\alpha \neq 0$ which implies that either, $n_i = 0$ or $x = 0 = y$. For the latter, this implies that $f_j = 0$ for $j \in V_i$. This is valid for all $i \in V$. Thus, $f_i = 0$. This implies that the third equation of (4.13) becomes

$$0 = -\eta d_i. \tag{4.15}$$

Assuming $\eta \neq 0$, which implies that $d_i = 0$. Similarly, $r_i = 0 = s_i$. Thus, one equilibirum point is $p_0 = (1, 0, 0, 0, 0)$.

Now, consider $n_i = 0$. Then, the system of equations (4.13) can be re-arranged such that the equilibrium point $\mathbf{p}_* = (n_i^*, f_i^*, d_i^*, r_i^*, s_i^*)$ can be expressed by

$$\begin{aligned}
n_i^* &= 0, \\
f_i^* &= \tfrac{s_i^*}{\omega}, \\
d_i^* &= \tfrac{\alpha x + 1}{\eta} s_i^*, \\
r_i^* &= \tfrac{\alpha x + 1}{\gamma} s_i^*, \\
s_i^* &= \tfrac{\omega \eta \gamma}{\eta \gamma + (\alpha x + 1)\omega \gamma + (\alpha x + 1)\omega \eta}.
\end{aligned} \tag{4.16}$$

To determine the stability, we are interested in the behaviour in the neighbourhood of the equilibrium point. Thus, we consider a close distrubence, say, δ. Then, $p' = p + \delta$. Using a Taylor series around the equilirium, we get

$$\begin{bmatrix} \dot{p}'_1 \\ \dot{p}'_2 \\ \dot{p}'_3 \\ \dot{p}'_4 \\ \dot{p}'_5 \end{bmatrix} \approx \begin{bmatrix} -\alpha x - y & 0 & 0 & 0 & 0 \\ y & -\omega & 0 & 0 & 0 \\ \alpha x & & -\eta & 0 & \alpha x \\ 0 & 0 & \eta & -\gamma & 0 \\ 0 & 0 & 0 & \gamma & -(\alpha+1) \end{bmatrix} \begin{bmatrix} p'_1 \\ p'_2 \\ p'_3 \\ p'_4 \\ p'_5 \end{bmatrix}, \tag{4.17}$$

where $\dot{p} = \mathrm{d}p/\mathrm{d}t$ and the 5×5 Jacobian matrix is denoted by J. In order to determine the stability, we are interested in the signs of the real part of the eigenvalues, λ, of $J|_{p'}$. If there is an eigenvalue where the real part is positive then the equilibrium point is unstable. However, if all eigenvalues have a negative real part, then the equlibrium is stable [14]. For $p' = p_0 + \delta$, assuming $\alpha, \beta, \omega, \eta$ and γ bigger than zero, the eigenvalues are

$$\begin{aligned} \lambda_{p_0 1} &= \omega, \\ \lambda_{p_0 2} &= \eta, \\ \lambda_{p_0 3} &= 0, \\ \lambda_{p_0 4} &= -\gamma, \\ \lambda_{p_0 5} &= -\alpha - 1. \end{aligned} \tag{4.18}$$

This indicates an unstable equilibrium point (a saddle). For $p' = p_* + \delta$, the first two eigenvalues are

$$\begin{aligned} \lambda_{\mathbf{p}_* 1} &= -\omega \\ \lambda_{\mathbf{p}_* 2} &= -\alpha x - y. \end{aligned} \tag{4.19}$$

However, for the three eigenvalues we need to solve

$$\lambda^3_{p_* 3,4,5} + (\gamma + 1 + \alpha + \eta)\lambda^2_{p_* 3,4,5} + (\alpha\gamma + \gamma\eta + \gamma + \eta + \alpha\eta)\lambda_{p_* 3,4,5} - \gamma\eta\alpha x + \alpha\gamma\eta + \gamma\eta = 0 \tag{4.20}$$

For this, we used Maple's *solve* function, and get that

$$\begin{aligned} \lambda_{p_* 3} &= \tfrac{1}{3}(-\alpha\eta - \gamma - 1) + O(\alpha^2, \eta^2, \gamma^2), \\ \mathrm{Re}(\lambda_{p_* 4,5}) &= \tfrac{1}{3}(-\alpha\eta - \gamma - 1) + O(\alpha^2, \eta^2, \gamma^2). \end{aligned} \tag{4.21}$$

For, $0 < \alpha, \eta, \gamma < 0$ the real parts of the eigenvalues are negative. For, $\alpha = \gamma = \eta = 1$ and $x = 0$, we that that:

$$\begin{aligned} \lambda_{p_* 3} &= 0, \\ \mathrm{Re}(\lambda_{p_* 4,5}) &= -2, \end{aligned} \tag{4.22}$$

and for $\alpha = \gamma = \eta = 1$ and $x = 1$

$$\begin{aligned} &\lambda_{p_*3} \approx -0.2451, \\ &\mathrm{Re}(\lambda_{p_*4,5}) \approx -1.8774. \end{aligned} \tag{4.23}$$

Thus, the p_* is for $0 < \alpha, \gamma, \eta \leq 1$ stable. Note, that $1 < \alpha, \gamma, \eta$ is possible but the analysis was inconclusive. Trials with the ODEs suggest that the equilibirum point is still stable.

To test this model, we constructed a simulation incorporating the ideas of equations (4.12). We checked the state of each agent individually and depending on the current state the agent, we computed a random variable. Whenever the variable was less than the total probability to change from the current state to the next one, the agents state moved, otherwise the agent stayed in the same state. Also, the simulation includes a distribution of donors to charities. That is, in the simulation, we have M charities that whenever a fundraiser is selected the charity gains a "donation" from the fundraiser, but also from the donors that are attracted via the fundraiser donate to the selected charity. Note, that again the charities are considered to be chosen with the same probability disregarding any difference such as popularity, or a particular cause the charities fundraises.

Further, we needed to choose a social network that represents people as vertices and the edges as the social interactions between these people. We choose to use a network from Facebook data of members of Vermont University of the Facebook100 data set [22] as the underlying social network. Facebook is a social networking site that was founded in 2004. It started as a university project in Harvard but by 2005 it allowed people to register with an email address with an ".edu" ending. The Facebook100 dataset contains the data of Facebook members of 100 American universities at a certain day in September 2005 [22]. In particular, the file contains 100 adjacency matrices whereby the vertices are the members of Facebook of each individual university. An edge connects two vertices whenever they are "Facebook friends". A "Facebook friendship" starts when one person invites the other to become friends and the other accepts. This leads to an undirected network. In a Special Topic [24], we already discussed the network for Vermont University. Note, that online social networks represent real world social networks just to a limited degree [5] as accepting a "Facebook friendship" involves much less maintenance then to keep a real world friendship. However, certainly, Facebook and other sorts of online social networking sites as well as emails might to be used to distribute the address of the fundraiser webpage as it seems to be an easy way to attracted people with a link of the page. Still, further research needs to go into this question to be completely sure whether social online networks play an influential role in the distribution of fundraiser pages.

To give an overview of the network of the Vermont University [17] Facebook network, we will summarize some main results of the Special Topic [24]. 7324 students and faculty members were part of Facebook in September 2005, and 382442 friendships existed between the Vermont University Facebook members. The network is a connected network. The vertices in the network have a mean degree of 52.22. Further, the mean distance for the shortest path is 2.13. Vertices have a normalised closeness value of 0.3673. Thus, we can say that students and faculty members at Vermont University in the Facebook data were closely connected.

To solve the system of ODEs, we used the forward Euler method [11, 21]. We do not think the model is really accurate in the sense that the model predicts the behaviour of individuals accurately. However, it can be seen as a guidance for the excepted behaviour over time. This is because we use quite a lot of assumptions and neglect other important aspects (i.e. characteristics of the donors, a growing network of customers, using a Facebook network as the underlying social network, etc.). Therefore, we do not need to use an accurate method to solve the ODEs so decided to use a method that is fast to implement.

We first checked whether the simulation and the system of ODEs show the same long-term behaviour. As an initial starting condition, we used that 500 fundraiser and 6824 "new" susceptible existed initially. Figure 4.3 shows the time evolution of the normalized total number of people in a state. We run the simulation 100 times and took the mean of all the outcomes of the simulation. The result is shown in the red graphs and the solution to the ODEs is represented by the blue graphs.

$\|\|n_s - n_{ODE}\|\|$	$\|\|f_s - f_{ODE}\|\|$	$\|\|d_s - d_{ODE}\|\|$	$\|\|r_s - r_{ODE}\|\|$	$\|\|s_s - s_{ODE}\|\|$
0.0127	0.0464	0.0467	0.0533	0.0209

Table 4.1: The difference between simulation and ODE values in the Euclidean norm.

There are 5 unknown parameters. Unfortunately, the time limit set on this thesis

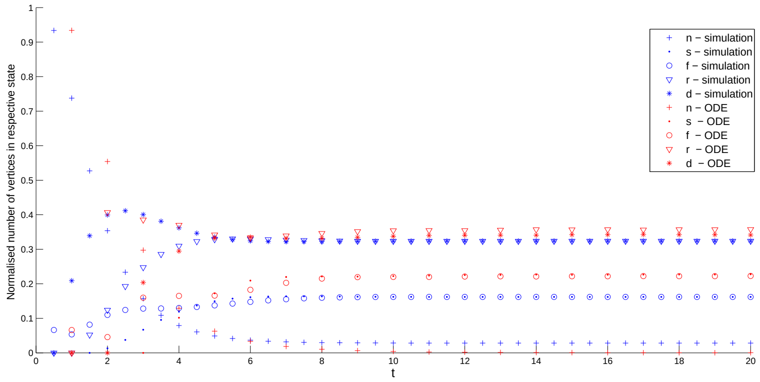

Figure 4.3: The states n, f, s, d, and r represent the number of vertices in the states normalized by the entire population of the Vermont University network. The red graphs indicate the time evolution computed by the simulation whereas the blue graphs show the solution of the ODEs. Initially there were 500 fundraiser and 6824 new susceptible present. Further, $\alpha = \omega = \eta = \gamma = 0.5$, $\epsilon = 0.1$, and $\delta t = 0.5$.

did not allow an experiment in form of a questionnaire[1] or other means to get further inside into the specific ranges for the rates of change of the different states. Thus, to see whether the system of ODEs shows similar behaviour as the simulation, we said that $\alpha = \omega = \eta = \gamma = 0.5$. The error between the simulation and the ODE values can be found in Table 4.1. The error of the Euler method is δt [11, 21]. This is not really accruate. However, in Figure 4.3, we see that the behaviour of the model and the simulation is similar.

The next step is to check whether we can see similar behaviour for different parameters as we observed in Section (3.3). First, we looked at the degree distribution for the donors and charities for varying parameters. The bipartite network we studied earlier is a network that evolves from an underlying friendship network. In particular, the bipartite network has one set of vertices of donors and the other set is has elements that represent the charities. To test the simulation for a wider range of parameters, we varied α, η, γ and ω. Table 4.2 shows the values for each varying parameter. Note that we kept the non-varying parameters at 0.5. Thus, this still shows a small percentage of possible combinations of parameters. For future work, we would attempt to test the parameter space with a Monte-Carlo algorithm and choose parameters more arbitrarily. However, it becomes clear that the maximal degree for the donors is one.

[1] Note, that FirstGiving already asked the donors about the likelihood for them to become donors or fundraisers. This information can be found in the file containing the demographics.

Also note that the number of donors donating to a charity for α, ω and η increases when α, ω and η increase. The parameter values α and η also do not seem to have a big influence on the number of donors in the system as the number of active donors hardly changes. Whereas for increasing γ the number of donors decreases. This can be explained by refereing to equations (4.16). The system is in steady therefore equations (4.16) apply. To have a large number of active donors, we require γ to be small whereas for γ large the number of active donors decreases. The opposite is valid for ω.

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
α	p(k = 0)	5577	5636	5656	5698	5750	5746	5807	5843	5851	5862
	p(k = 1)	1747	1688	1670	1626	1574	1703	1804	1814	1899	1936
γ	p(k = 0)	6642	6255	5999	5866	5699	5621	5520	5510	5425	5388
	p(k = 1)	682	1069	1325	1458	1625	1703	1804	1703	1610	1941
ω	p(k = 0)	3981	4109	5383	5837	5922	5696	5563	5500	5456	3366
	p(k = 1)	4344	2418	1941	1610	1402	1628	1761	1824	1868	1958
η	p(k = 0)	6672	6610	5987	5812	5655	5693	5527	5509	5426	5495
	p(k = 1)	1609	1689	1337	1487	1669	1631	1797	1815	1815	1854

Table 4.2: Number of degrees for varying parameters α, η, γ and ω for 10 trials.

For the degree distribution for the charities we omited to have a look at different parameters but used $\omega = \eta = \alpha = \gamma = 0.5$, again. Figure 4.4 shows that we get a distribution for the number of degrees for the charities that does not resemble the distribution observed in Section 3.3. Thus, the model is insufficient to imitate the degree distribution for the charities observed in Section 3.3.

Finally, we also can say something about the betweenness centrality and the shortest paths from the degree measures of the vertices of the bipartite network. Because the active donors all have a degree of one and have not donated to more then one

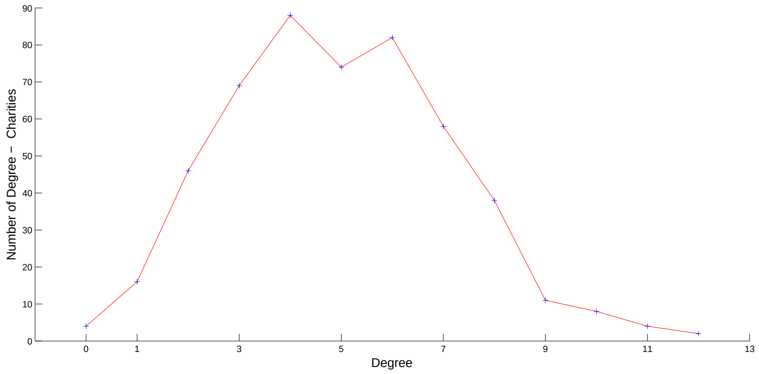

Figure 4.4: Degree distribution of donor-charity network for varying β where $\omega = \eta = \alpha = \gamma = 0.5$ at $t = 90$ with 500 charities present.

charity, the shortest path for most of the donors to other donors who donated to the same charity is two and to the charity itself one. Additionally, all charities have a shortest path of one to the donors donating to them. For the betweenness centrality, the charities form the centre of star sub-graphs similar as explained in Section 3.3. Thus, on these subgraphs, all charities have a betweenness centrality value of one and the donors have all a betweenness centrality value of zero.

This concludes the analysis of the collective behaviour model for a newly introduced product. We have seen that it is important to have the fundraisers active for some time to attract new customers in the beginning. An option to improve the model would be to add a term that artificially increases the attention of new customers, for example, via advertisement. Further, we also saw that the model is still insufficient to explain the degree distribution of the charities that was observed in Section 3.3. However, the model seems to lead to a similar degree distribution, betweenness value, and shortest path as observed in Section 3.3.

Chapter 5

Conclusions

We started this thesis by introducing FirstGiving. In particular, we drew attention to the process of creating a fundraiser web page that was later incorporated in the collective behaviour model. Then we started the study of the data set by extracting some demographics from the data set containing information about the users of FirstGiving. We suspect that the mean user is middle aged and earns above the 2010 average income of the US.

From the data set containing the transaction between donors and charities, the number of donors increase per year steadily, further we saw that April and September are prefered month for donation. The data seems also to suggest that, on Mondays, more donors partcipate via FirstGiving then on Saturdays. Finally, assuming that the time whenever the transaction was made was saved in East Coast time, donations peak in the afternoon and continue during the night.

The chapter about the demographics was followed by outlining the definitions and network measures used in this thesis. In particular, we looked at the degree, highlighting the power-law degree distribution. The concept of a path between vertices lead to the closeness centrality and the betweenness centrality. For the betweenness centrality, we looked at the derivation of Brandes' algorithm to weight vertices depending on their occurence in the shortest paths from other vertices and modularity.

This was followed by a chapter where we applied the network diagnostics. We constructed the networks from the transaction data set where the vertices were formed from the donors and charities. We said, there exist an edge between a donor and a charity whenever a donor made a donation to a charity. This lead to the structure of a bipartite network. To save some computation time and memory space, we aggregated the entire network into smaller once by only considering contributions made to charities in November of each year. We further aggregated the networks by only using active vertices of the first four Mondays in November.

The degree distribution for the charities indicate that it might follow a power-law distribution. However, the p-test was inconclusive and, although we cannot definitely say that the charity degree distribution follows a power-law, it is still valuable to say that it is leaning towards it. For the donor degree distribution, it is fair to say that most donated once during a month/day with a couple of people using First Giving multiple times in that time period. The closeness and betweenness centralities for the days indicated something similar. That is, most of the donors had a betweenness centrality of zero and low closeness values. However, we saw a big change in the topology of the network when just a couple of people donated more then once. This caused the diameters of the networks to be significantly smaller and increased the closeness value, as more vertices were available to have a path from one vertex. Finally, we looked at the modularity of the donors who donated to the same charity. It became visible that there is a leaning towards grouping of certain characteristics. Thus, donors with a similar age, household income or who live in the same state tend to donate to the same charity.

In the last chapter, we looked at a collective-behaviour model starting with the SI model that simulates epidemic spread, followed by the threshold model and finally, combining the two ideas to a model simulating the donation behaviour via FirstGiving. We used as an underlying social network: the Facebook network for Vermont University from September 2005. The goal was to write a model that leads to a similar structured network as observed in the previous chapters. Thus, after discussing the equilibrium values of the ODEs, we also wrote a simulation based on the ODEs. Part of the simulation was also a distribution of donors between charities. Finally, we compared the degree distribution, betweenness, and closeness centrality with the networks created from the data set. The simulation created similar results for the degree distribution for donors, betweenness, and closeness centrality for the donors and charities. However, the degree distribution of the charities differed from the one observed. Thus, the algorithm for distributing donors among charities is insufficient.

For future work, it might be useful to include a similar algorithm as the preferential attachment proposed by Barabasi and Albert that leads to a power-law degree distribution [1]. That is, one starts with a seed network and adds a new vertex to the network. The new vertex then distributes m edges. The edges are distributed starting from the new vertex such that the other end is more inclined to attach at a vertex that already has a higher number of edges then other vertices. This could be used to model the contribution to a charity of a donor, as it seems plausible that there are a couple of charities that are well known by a majority of people and therefore might be chosen more often then say regional charities that are known only to people living closely.

As already mentioned, another project for future work is to distribute a questionaire amoung the users of FirstGiving asking about the likelihood of becomeing a fundraiser, donating via a fundraiser, what threshold they have to start using a new product, the time they wait before donating again. Still, even when accumulating this information, a further analysis of the parameter set in form of, say, a Monte Carlo simulation might be helpful as people might give wrong information during such questionnaire. Certainly, we need to further investigate the parameter space.

The data set has not been fully explored yet. There are still characteristics about the donors that might be insightful in a further analysis, and might help to create a social map of the donnors donating via FirstGiving.

Finally, an improvement of the ODE model would be to also consider thresholds for each person individually instead of considering the threshold being the same for each person, and test if that leads to different behaviour.

Bibliography

[1] R. Albert and A.-L. Barábasi. Emergence of scaling in random networks. *Science*, 7286:509–512, Oct 1999.

[2] R. Albert and A.-L. Barábasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, January 2002.

[3] M. Barrat, A. Barthelemy and A. Vespignami. *Dynamical Processes on Complex Networks*. Cambridge : Oxford University Press, 1st edition, 2008.

[4] B. Bollabas. *Modern Graph Theroy*. Dpringer, 1st edition, 1998.

[5] D.M. Boyd and N.B. Ellison. Social network sites: Definition, history, and scholarship. *SIAM Rew.*, 2008.

[6] U. Brandes. A faster algorithm for betweenness centrality. *J Math Sociol*, 25:163–177, 2001.

[7] US Census Bureau. 2010 census. Web site: http://2010.census.gov/2010census/[Last accessed: 22/08/2011].

[8] M. Centola, D. Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 2007.

[9] C.R. Clauset, A. Shalizi and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.

[10] FirstGiving. Online fundraising website for events and causes — first giving. Web site: http://www.firstgiving.com/[Last accessed: 22/08/2011].

[11] K. Gillow. Case studies in scientific computing. —The Van der Pol equation., 2011.

[12] M. Granovetter. Threshold models of collective behavior. *J Math Sociol*, 83, 1978.

[13] T. Ito. Toward a proteinprotein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 25:1143–1147, 1999.

[14] P. Jordan, D.W. Smith. *Nonlinear Ordinary Differential Equations, An Introduction to Dynamical Systems.* Oxford University, 4th edition, 2007.

[15] M. E. J. Newman. *Networks : An Introduction.* Oxford : Oxford University Press, 1st edition, 2010.

[16] M.E.J. Newman. Power laws, Pareto distribution and Zipf's law. *Contemporary Physics*, 51:323–351, 2005.

[17] The University of Vermont Burlington. The university of vermont. Web site: http://www.uvm.edu/[Last accessed: 22/06/2011].

[18] Everett M. Rogers. *Diffusion of Innovations.* New York : Free Press, 4th edition, 1995.

[19] M. Rubinov. Brandes' algorithm. Web site: https://sites.google.com/a/brain-connectivity-toolbox.net/bct/metrics/list-of-measuresTOC-Paths-Distances-and-Cycles[Last accessed: 21/08/2011].

[20] O. Sporns. Breath first algorithm. Web site: https://sites.google.com/a/brain-connectivity-toolbox.net/bct/metrics/list-of-measuresTOC-Paths-Distances-and-Cycles[Last accessed: 21/08/2011].

[21] E. Suli. B21a numerical solution of differential equations i. typed lecture notes for weeks 1-3 [lectures 1-6], 2010.

[22] Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social Structure of Facebook Networks. arXiv:1102.2166, 2011.

[23] D. Watts. A simple model of global cascades on random networks. *PNAS*, 25:5766–5771, 2002.

[24] A. Wipprecht. Centrality measures. Special Topic., 2011.