CUSTOMER VALUE ESTIMATION CONSIDERING SOCIAL NETWORK RELATIONSHIPS

HINRICH HARMS

Studienarbeit

Department of Knowledge and Language Engineering
Faculty of Computer Science
Otto-von-Guericke University Magdeburg
December 2009

INF FAKULTÄT FÜR INFORMATIK

SUPERVISORS:
Prof. Rudolf Kruse
Georg Ruß
Dymitr Ruta

ABSTRACT

It is of crucial importance for telecommunication companies to know the exact value of a customer. This allows them to detect the most profitable customers, that need to be retained. The value also depends on the influence of other customers in the network: for example, a customer might stimulate others to make more phone calls. The impact is measured by analysing the neighbours in the social network of customers, that recently changed to another company. Does the likelihood, that the neighbours change, rise as well? Does their call time decline? These questions will be answered in this paper. A measurement indicating the amount of impact on the neighbours is proposed. Furthermore, the expected time a customer stays with a company before changing to a competitor is determined. This is used to calculate a *Customer Lifetime Value*, the expected total revenue a person generates during the time he is customer with that company. To assist the analysis of the network in general and of a specific customer, a software tool is presented.

ZUSAMMENFASSUNG

Für Telekommunikationsunternehmen ist es von entscheidender Bedeutung, den konkreten Wert eines Kunden zu kennen. Er ermöglicht es festzustellen, welche Kunden besonders profitabel sind und unbedingt gehalten werden sollten. Dieser Wert wird auch durch Einflüsse auf andere Kunden im Netzwerk bestimmt: So kann z. B. ein Kunde andere zum Telefonieren animieren. Dieser Einfluss wird bestimmt, indem die Nachbarn im sozialen Netzwerk von Kunden, die ihren Telefonanbieter gewechselt haben, untersucht werden. Steigt bei ihnen die Wahrscheinlichkeit, dass sie auch wechseln? Sinkt die Anzahl ihrer Telefongespräche? Dazu wird ein geeignetes Maß angegeben, das Aussagen über die möglichen Auswirkungen auf Nachbarn treffen lässt. Weiterhin wird die durchschnittliche Zeit ermittelt, die ein Kunde bei einem Unternehmen bleibt, bevor er kündigt oder zu einem Konkurrenten wechselt. Damit wird der sogenannte *Customer Lifetime Value* eines Kunden berechnet, der den erwarteten Gesamtumsatz bis zur Kündigung angibt. Zur Unterstützung der Netzwerkanalyse und zur Untersuchung einzelner Kunden wird schließlich ein dazu entwickeltes Programm vorgestellt.

CONTENTS

LIST OF FIGURES

Due to data confidentiality issues, the numbers on some of the figures were removed.

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

Do telecommunication customers influence each other in their calling behaviour? This is the key question examined in this paper from a telephone company's point of view. It is of special interest for them to know the true value of a customer, because it can give them an advantage over competitors. Which customers are the most valuable? To prevent them from changing to another company, it might be profitable to grant those customers special benefits. Studies have shown, that it is about five to eight times more expensive to acquire a new customer than to retain an existing one [12].

This value depends on many factors. The most obvious one is his telephone bill. Another factor, that is more complicated to measure, is the impact on other customers, which will be the focus of this paper. Different methods will be introduced to gain an understanding of this influence by looking at structural changes in the social network of the customers. This network is a representation of the interactions of one customer with another, for example via voice calls, SMS or email. Each interaction creates a connection between two persons, one being the initiator and the other one the receiver. All persons a customer has interacted with are called his neighbourhood. A strong connection usually results from a relationship between two customers like friendship, family or business relations. It is very likely that those people influence each other. For example, if a customer decides to cancel his contract, he might convince others in his neighbourhood to do the same.

These customers, also called *churners* (from *to churn*: to change the service provider; mainly used in the context of telecommunication or broadcasting companies), provide valuable information. By analysing the network before and after churn, characteristic effects can be deduced, which can be used to detect churn beforehand or to calculate the real loss that occurs upon churn.

Furthermore, the possibility to classify customers as churners and non-churners, using only data of telephone calls, is analysed. This is not done with the intention to replace existing churn prediction systems, but to clarify which information can be gained from the available data.

As a final result, a total customer value is presented in special consideration of influences of other customers and the potential neighbourhood value loss in case of churn.

1.2 PROBLEM OUTLINE

The goal is to detect influences between customers by analysing temporal changes in the social network before and after churn and estimate their real value. In order to achieve this, a large amount of call data has to be analysed. It consists of more than 76 million calls and a call duration of 440 years in total. This evokes the need for algorithms that scale well and can handle large tables stored in a database. The network at the time of churn provides crucial information for analysis,

as the structure changes and therefore the calling behaviour of other customers might also change.

If successful, this provides valuable information to enhance existing customer retention systems by specifying a value for each customer considering a network added value. As the source material consists just of the information about the caller, called person, length in seconds and the date, it is necessary to approximate the incurred costs for the customer using this data. This also implies that the network is made up of only call data and all other means of communication have to be ignored. This makes it more difficult to detect the impact of churn on other customers, because less information is used, but this data is available to most telecommunication companies and can be analysed by the methods investigated here.

Additionally, the possibility to predict future churners using call data is analysed. Are there any characteristic features in the network, that indicate whether a customer might cancel his contract soon?

As a main result, a customer lifetime value will be calculated. It depends on his expected lifetime and the customer value. This figure, describing a customer's total revenue, is of high importance for marketing, because it gives them the information about which customers to target.

1.3 SOCIAL NETWORKS

The term *Social Network* refers to structures made up of actors, tied together by connections representing social relationships of various kind between them [11]. In graph theory, the actors are represented by a set of vertices and the connections are directed or undirected edges joining two vertices [1].

In the data analysed here, vertices are a telephone number representing individual customers and edges are telephone calls from one customer to another. Edges are directed with the origin at the caller. They are weighted by the total call time from one customer to another, but other metrics such as the number of calls are possible as well. This defines the strength of the connection.

Two customers connected by an edge of an arbitrary direction are called neighbours and all neighbours of a customer are called his neighbourhood.

1.4 STRUCTURE

The rest of this paper is structured as follows. The next chapter discusses the special problems that arise from real world data and how undesired effects can be compensated. Chapter 3 describes statistical methods to determine the expected time a customer stays with a company until he terminates his contract, also called customer lifetime. After that follows a chapter presenting the calculation of the customer value considering neighbourhood effects and the combination of this with the result of the previous chapter to determine a lifetime value. Chapter 5 presents a software tool implementing the discovered methods and visualising a customer's neighbourhood. Finally, a summary of the results and an outlook on future work is given.

2 NETWORK ANALYSIS

The analysed network consists of phone calls made by customers of a large telecommunication company over a period of three months in a certain area. More than 76 million calls were made during that time which consume several gigabytes of database storage. Additionally, the data contains information about the customer type (residential or business), the acquisition date and cessation date (churn date) if appropriate.

The social network was built using the customers, represented by their telephone number, as nodes and the duration of all calls between two customers as the connection strength (edges). Edges are directed, leading from the caller to the called person. The total duration of calls made was used as a measure for connection strength, because it relates closely to the actual revenue a customer generates.

2.1 DATA PREPARATION

The original data is organised in an Oracle database using a separate table for each day. The large amount of phone calls requires special attention to the running time of the queries that were initiated from a Java application. To increase the speed, indexes were created on the caller's and the called person's telephone numbers. For queries involving all available days, a snapshot ("Materialized View") of the table union of all call data and an index were created to reduces the query time. The complete index made it possible to look for all data set of a specific customer in reasonable time. This also simplifies further queries, because only one table has to be used instead of using a separate table for each day.

As with many real-world data, there is always the possibility to have errors in it. That makes it necessary to check the sanity of the data. Examining the given call data revealed several issues that had to be dealt with:

- Some entries were longer than the maximum telephone number size of eleven digits. This might be caused by the way the data was captured and people pressing more numbers after the connection is made (e. g. if using an automatic phone system). Because the numbers can have a variable length (ten or eleven digits), it has to be ensured that all entries are trimmed to the correct size. This was done by mapping the entries of the call data to a list of customer phone numbers. To prevent the need of having to change the original data, a new table was created with that mapping.
- There were also entries with too short telephone numbers. These had a call duration of zero, which indicates that no valid connection was made and were ignored.
- Some calls were initiated by the operator number "0", which were also left out of the further analysis.

- For some telephone numbers, there were two or more cessation dates in the database. If this was the case, only the latest cessation date was used.
- After this adjustment, there still existed entries in the call data initiated by an already cancelled phone number. Using the information of the calls, the cessation date was corrected according to the last call made.

All these clean-up steps needed to be taken in order to ensure the data is in a consistent state.

2.2 TRENDS AND NORMALISATION

An analysis of all calls was carried out to see an up- or downward trend in the data. The results show (see Figure 1), that there is a variation in the number of calls made over time. The linear regression shows a decline in the total number of calls. To eliminate the effects of this trend in the further analysis, all results were normalised according to the total number of calls made in the corresponding period (e. g. a week).

To achieve the normalisation, the calculations of all following analyses were conducted two times for every period: One time for the examined subgroup (the churners) and another time for all customers. The results for the churners r_c were then divided by the results for all customers r_{all} in the according period, and multiplied with an average over all data records to get a meaningful numeric value:

$$r'_{c,p} = \frac{r_{c,p}}{r_{all,p}} \cdot \frac{1}{p_{max}} \sum_{n=1}^{p_{max}} (r_{all,n})$$

The normalised result r' depends on the analysed period p, the subgroup of examined customers c, the highest period number p_{max} and the specified results. This enabled comparing the changes between different periods to each other and compensated the effect of the general decrease in calls.

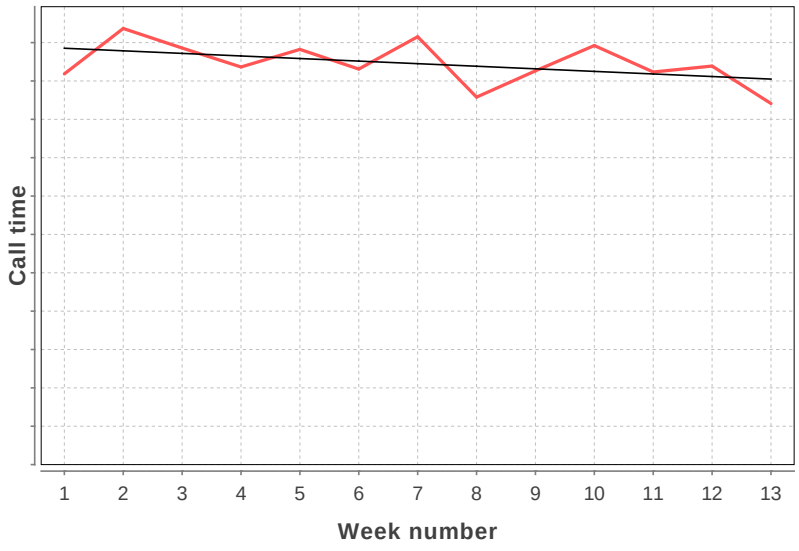


Figure 1: The average call time of all customers including a linear regression line visualising the declining trend.

3 CUSTOMER LIFETIME

One aspect of customer value calculation is to know the expected time between acquisition and cessation date, also called lifetime, of a customer. Together with the periodic revenue, this can be used to calculate the total customer lifetime value. This chapter describes a statistical analysis of customer lifetime and looks into available options for prediction by inspecting the informational value of the original data. Does the data allow to classify customers in churners and non-churners?

3.1 STATISTICAL ANALYSIS

The information of customer acquisition and their cessation within the last 100 years was used to determine the distribution of lifetimes. As a first step, different standard distributions were fitted to the data. The histogram (Figure 2) suggests an asymmetric distribution with a starting point of zero (because negative lifetimes are impossible) and an asymptotic decline towards the x-axis at infinity. The Exponential, Weibull and Gamma distributions have these properties and will be analysed in-depth.

These three distributions were fitted to the data using the maximum likelihood estimation. The results are shown in Figure 2. It can be seen, that the Weibull as well as the Gamma distribution follow the probability distribution of the data very closely, whereas the Exponential distribution seems to be inappropriate for approximating customer lifetimes. The Q-Q plots in Figure 3 help to distinguish further: The closer the data values, drawn against the tested distribution, are to a line with a slope of 1, the better the theoretical distribution agrees with the data. The data points for the Weibull distribution are on the line only for small lifetimes. The longer the lifetimes get, the farther away the points are from the line and are therefore only a good approximation for short lifetimes. If using the Gamma distribution, the data points

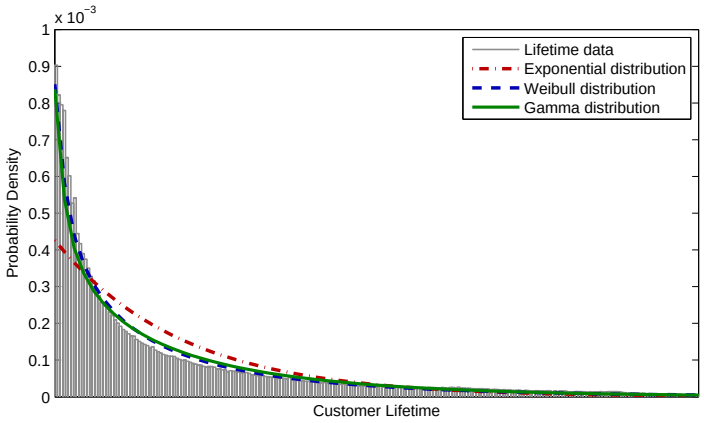
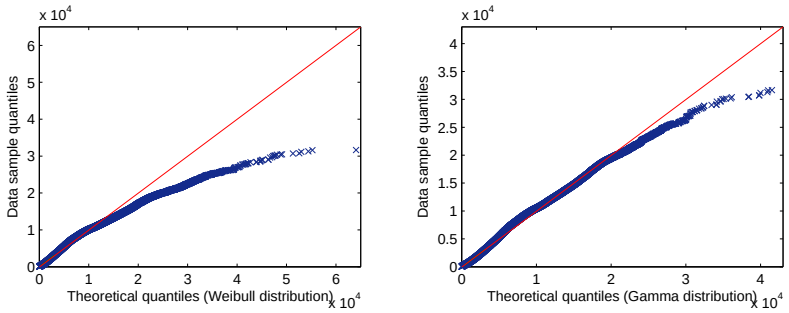


Figure 2: Probability density of the customer lifetime data and fitted distributions.



(a) Weibull distribution (b) Gamma distribution

Figure 3: Q-Q plots of the customer lifetime data drawn against the quantiles of a fitted distribution.

are much closer to the ideal line. Only very long lifetimes (longer than 60 years) produce significant outliers and seem to follow a different distribution. But for all other lifetimes shorter than 60 years, which make up more than 99.9 % of the data, the Gamma distribution is an adequate approximation and the best-fitting distribution found. It is also widely used in other lifetime modelling applications [2].

3.1.1 *Expected Customer Lifetime*

Since the Gamma distribution has been established as the best-fitting distribution, it will be utilised for all further customer lifetime analysis tasks. It is described by the probability density function

$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)}$$

where k is the shape and θ the scale parameter. Γ is the Gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. The cumulative distribution function is obtained by integration of the probability density function. It can also be expressed using the lower incomplete gamma function $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$:

$$F(x;k,\theta) = \frac{\gamma(k,x/\theta)}{\Gamma(k)}$$

This function calculates the probability for a newly acquired customer to churn within a period of the length x. To obtain the probability of a customer to stay longer than that period, the survival function can be used, which is calculated by

$$S(x;k,\theta) = 1 - F(x;k,\theta)$$

The expected remaining lifetime of a customer at an instance x_0 after acquisition, also called lifetime function or expected future lifetime, can be calculated by integrating the survival function starting from x_0 and re-normalisation [8]:

$$L(x_0;k,\theta) = \frac{1}{S(x_0;k,\theta)}\int_{x_0}^{\infty} S(x;k,\theta)dx$$

If the lifetime function stated above is used directly at acquisition, at $x_0 = 0$, the expected future lifetime simplifies to the mean or expected

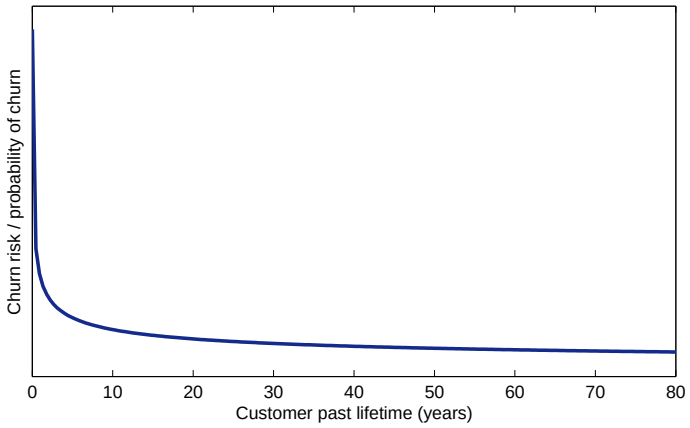
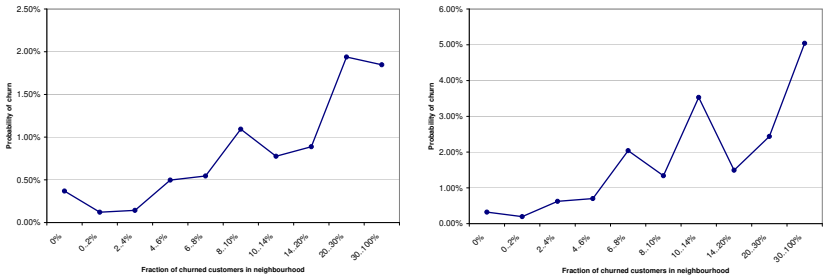


Figure 4: Development of the churn risk over time within a period of 3 months.



(a) Residential customers (b) Business customers

Figure 5: Percentage of churners in dependence on the numbers of churners in their neighbourhood.

value of the gamma distribution $\mu = k \cdot \theta$. Another interesting aspect is the development of the churn risk over the customer's lifetime in a fixed period τ. This is calculated using the survival function at the customer age x_0 minus its value at the length of the period ahead and normalisation:

$$H_\tau(x_0) = \frac{S(x_0) - S(x_0 + \tau)}{S(x_0)}$$

Figure 4 shows the churn risk within a period of three months. It can be clearly seen, that the churn risk for newly acquired customers is considerably higher than for customers who have already stayed for a longer period of time. This behaviour is to be expected, because long-standing customers tend to be more loyal, and new customers are often customers, that frequently change their providers.

3.2 STIMULATION OF OTHER CUSTOMERS TO CHURN

Do churned customers make it more likely for another customer in their (former) neighbourhood to churn? To investigate this question, the churn rates in neighbourhoods with a certain number of churners were examined. To lessen the effect of the neighbourhood size, the fractions of churners in each neighbourhood are compared.

Figure 5a clearly illustrates the dependence between the probability of a residential customer to churn and how many customers in

Fraction of churned neighbours between...	Churns	Total customers	Probability of churn
0%	1401	380123	0.37%
0..2%	1	825	0.12%
2..4%	3	2105	0.14%
4..6%	8	1610	0.50%
6..8%	6	1100	0.55%
8..10%	11	1006	1.09%
10..14%	5	644	0.78%
14..20%	8	901	0.89%
20..30%	5	258	1.94%
30..100%	7	379	1.85%

Table 1: Number of neighbourhoods with a certain fraction of churners and churned residential customers.

his neighbourhood churned the month before: The more of a customer's neighbours have churned, the more often that customer has also churned. Due to the fact, that the probability of churn in such a narrow time frame of three months is relatively low, there were only very few examples of churned customers with other churners in their neighbourhood (see Table 1), resulting in a high variance of the measurements. To get results with a higher significance, the same tests have to be conducted within a much larger time frame or on data of more customers with a higher amount of samples that can be analysed.

Looking at business customers (Figure 5b) shows a similar result, but here churners even have a higher impact on others than between residential customers. This is in contrast to the results of Section 4.1, where the call time of business customers was almost independent from churners in the neighbourhood. The high impact might result from affiliated or subsidiary companies, who are using the same telecommunication infrastructure and are very likely to change the telecommunication provider together. This implies that preventing a business customer from churning is more important than it might seem by looking only at this one customer's revenue.

3.3 CHURN PREDICTION

Knowing which customers are likely to churn gives a company a competitive edge over its competitors, because those customers can be targeted by marketing campaigns convincing them to stay. This can be done by giving customers special rewards or discounts. The question of customer value loss in case of churn is addressed in Chapter 4. This section focuses on detecting which customers are most likely to churn.

There are many commercial solutions for churn management and prediction using various kind of data. This thesis is not trying to

introduce a complete replacement for these solutions, but rather to determine which information can be gained using just call data. The results could be integrated into existing churn prediction concepts.

3.3.1 *Attributes Used for Prediction*

The first step is to extract special properties for every customer from the raw call data that can be used for prediction. The prediction is made by comparing these properties of customers that are known to churn to other customers. The used information includes:

- Number, total duration and average length of in- and outbound phone calls
- Fraction of inbound call time
- Neighbourhood size (number of other subscribers talked to in the observed data)
- Fraction of neighbours that have already churned

Furthermore, the same information was collected for every neighbour and stored as a combined value for every customer. There are some customers in the data which have an extremely high number of neighbours and a disproportionately high call time. These customers lead to an asymmetrical distribution and a skewed mean value. To counteract the effects of these outliers, the median was used for number of calls, the total call time and the neighbourhood size. The fractions of inbound calls and churned neighbours were averaged.

The information if the customer has churned in the next month was saved as a binary value. This indicates the customer's class, if he is a churner or not, which will be used to train the classifiers.

3.3.2 *Classification*

The problem to be solved is the correct classification of customers in churners and non-churners using this data. A good framework to achieve this is the data mining software WEKA [4]. As a first step, linear regression was used to reveal simple dependencies between some of the properties of the original data and the resulting value. The generated model shows no dependence for most of the used attributes (see above, 3.3.1), but there was a small positive correlation between the fraction of inbound call time (0.0021), the fraction of inbound call time of the neighbours (0.0015) and the fraction of neighbours having churned the month before (0.0343). This confirms the results from Section 3.2, namely that churners can stimulate other customers to churn as well.

WEKA provides various kinds of classifiers, of which a Radial Basis Function Network, a Decision Tree, and a Bayesian Network were used. The results are verified using 10-fold cross-validation, which means that 9/10 of the data were used for training and the other 1/10 were used for validation. This was repeated ten times for every part of the data. The result is compared to the expected value of correctly identified churners by merely guessing using a known probability of churn.

The results from Table 2 show the general problem most classifiers have with the data: Because there are many non-churners and no significant indicators for churn, classifiers like Decision Trees (tested

Classifier	True positive	False positive
Guessing	0.37 %	0.37 %
Decision Tree	0 %	0 %
Radial Basis Function	0 %	0 %
Bayesian Network	0.96 %	0.32 %

Table 2: Results of the analysed classifiers predicting churn and the expected results by guessing.

with J48 [6], Alternating and LogitBoost Alternating Decision Trees [3]) or Radial Basis Function networks tend to classify all customers as non-churners. This gives them a high percentage of correctly classified customers, but makes the results insignificant. The Bayesian Network shows a more balanced, and with about 1 % of correctly classified churners (in average) a slightly better result than guessing, while maintaining the rate of false positives (non-churners classified as churners).

This shows the difficulty of predicting churn using only call data. There seems to be no significant evidence in the raw data that can be used to reliably predict churn. The main problems are the small amount of observed samples and the unbalanced data resulting from the low probability of churn within the analysed three months. Although no reliable prediction could be made by using just the available data, the number of churners in the neighbourhood as well as the ratio between in- and outbound call time could be valuable parameters to be integrated into existing churn prediction systems, that use many more parameters, as additional inputs.

4 VALUE ESTIMATION

An interesting part of the data are the customers who cancelled their contract. These customers, also called churners, are of particular concern for telecommunication companies or companies with other subscription-based business models, because all the revenue the customer generates is lost [7]. In addition to that, there might even be an influence on other customers [9]. The direct loss that occurs can be easily estimated according to the customer's bill, but the effect on others is more difficult to assess. This chapter focuses on the latter issue and tries to calculate a value for the expected neighbourhood loss in the case of churn for every customer. As a final result, the customer lifetime value is calculated by using the expected lifetimes from the previous chapter.

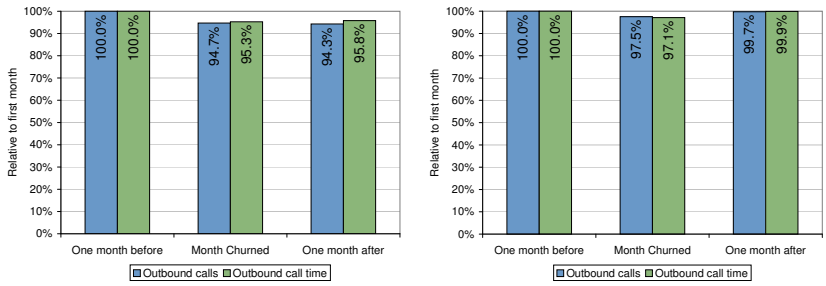
4.1 NEIGHBOURHOOD VALUE LOSS AFTER CHURN

To examine this influence, the neighbourhoods of churners one month before and one after the churn were analysed.

As can be seen in Figure 6a, a churning residential customer has a significant impact on his neighbours. The total amount of his neighbours' calls has decreased by 5.7 % on average for 2411 residential churners and the call time has decreased by 4.2 %.

The business customers (Figure 6b) show a different picture: In the month after the churn the amount of calls has been restored to 99.7 % of the calls in the initial month, and the value for the call time went back to 99.9 %.

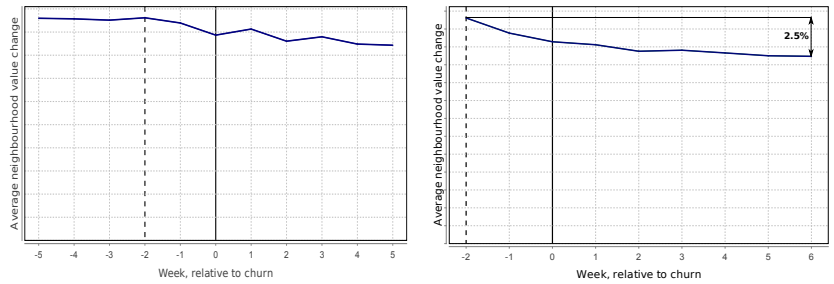
These results show that residential customers have an impact on their neighbours if they churn. Further analysis of this is carried out in Section 4.2, where more parameters are included in the calculation. Business customers do not seem to have a significant impact on each other, so the main focus will be on the residential customers.



(a) Between residential customers, average of 2411 churners

(b) Between business customers, average of 1989 churners

Figure 6: Change in the value of neighbourhoods of churners, one month before and after churn.



(a) Residential customers with five weeks prior to and after churn

(b) Residential customers with two weeks prior to churn and six weeks after

Figure 7: Neighbourhood value change of churners, normalised to the week of churn (week number zero).

4.2 IMPACT OF CHURNERS ON THEIR NEIGHBOURHOOD

For a more in-depth view of the data, the call time was analysed on a weekly basis to calculate a more exact measurement of the neighbourhood value loss. The week of churn is represented by the week with the number zero. To get an overview, all residential customers that churned in the middle of the analysed time span were examined, with at least five weeks of data available before and after churn. As can be seen in Figure 7a, the neighbourhood value stays steady prior to about two weeks before churn and begins to decline from that point on.

The steady phase of the neighbourhood value was omitted in order to be able to include more customers in the analysis. The result is shown in Figure 7b. The available data suggests a convergence to a constant value if the time after churn increases. After about four to six weeks, the neighbourhood value stays at the same level. Due to the lack of data over a longer period of time, it will be assumed that this doesn't change. A verification could be carried out in the same manner using data of e. g. a year.

All calculations used the normalised data according to Section 2.2. This implies that the neighbourhood value of churners has dropped more than the one of non-churning customers, namely by 2.5 % less call time in a period of only three to four weeks.

4.3 SOCIAL PARTICIPATION VALUE

This decrease in neighbourhood value has to be further differentiated to gain a better understanding how big a loss might be. A neighbourhood with many calls but only a few to or from the analysed customer is probably hardly influenced by his churn. A measurement called *Social Participation Value* by the author has shown a correlation. It describes the involvement of a customer with his neighbours to determine the impact on them in case of churn. The value is confined by 0 and 1. The lower bound is excluded, as it implies no contact with neighbours, which would result in having no neighbours at all. Nevertheless, small values are very likely. The upper bound represents a neighbourhood, that is only interacting with the analysed customer and no other customers

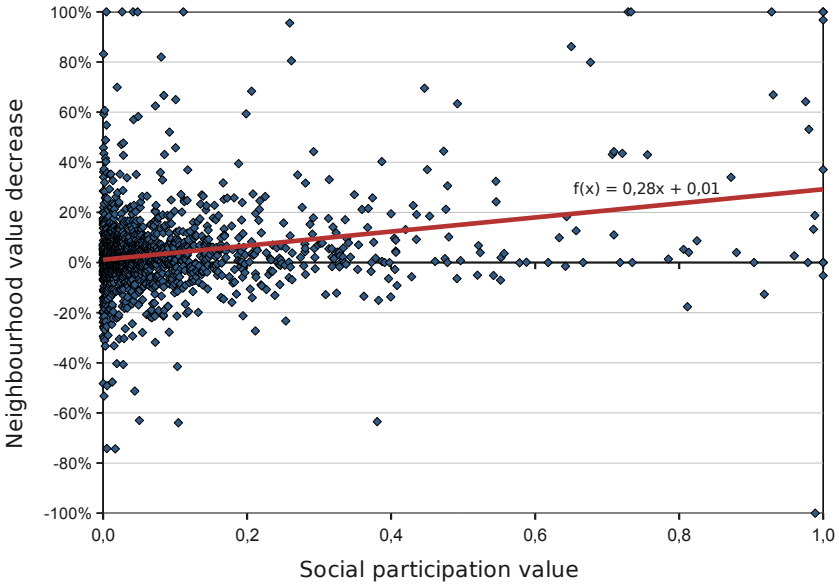


Figure 8: Scatter plot of the neighbourhood value decrease of single churners.

at all. It is calculated by dividing the call time of the customer by the length of all outbound calls of that customer and his neighbours:

$$SPV(x) = \frac{C_{out}(x) + C_{in}(x)}{\sum_{n \in N}(C_{out}(n)) + C_{out}(x)}$$

Where N is the set of all neighbours, x is the analysed customer and $C_{in/out}(x)$ the in- or respectively outbound call time of customer x. Instead of call time, a similar measure (e. g. number of phone calls) would be possible, too.

The results within residential customers, where the highest impact can be expected, are shown in Figure 8. The scatter graph depicts for all churning customers their social participation value and the occurred neighbourhood value loss. For low participation values, the value change is distributed equally around zero. For high participation values it can be clearly seen, that the neighbourhood value of almost all churners has decreased. A linear regression line supports this with a starting point of 0.01 (almost no impact) and a slope of 0.28 which represents a correlation between a high social participation value and a neighbourhood value loss.

4.4 CUSTOMER LIFETIME VALUE

Finally, a total customer value is calculated. It depends on the expected customer lifetime and the customer's revenue in a fixed time segment (e. g. a day) as well as the expected lifetime of his neighbours, the neighbours' revenue and the customer's impact on his neighbours.

The expected lifetimes are calculated using the lifetime function L from 3.1.1 and the time in days the customer has already stayed with the company. For the neighbourhood, their average "age" is used. The lifetime function calculates the area of the survival function S starting at the customer's age in days to infinity and re-normalises it.

A good estimate for a customer's revenue is his outbound call time. To get the daily revenue, this has to be divided by the length of the

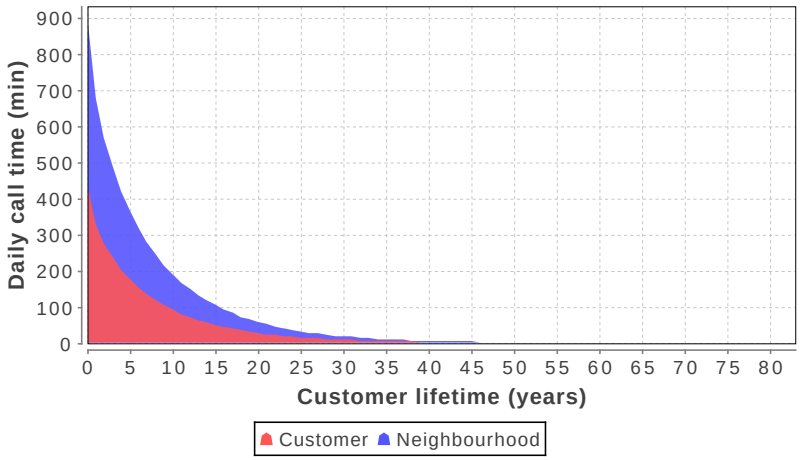


Figure 9: Chart of the customer value over time (example).

observed time period $T_{analysed}$. Multiplied with his expected lifetime, this results in the value directly generated by the customer:

$$v_C(x) = \frac{C_{out}(x)}{T_{analysed}} \cdot L(age(x); k, \theta)$$

Additionally, the potential neighbourhood value loss in case of churn has to be considered. This is estimated using the social participation value from the previous section and the parameters m (slope) and b (y-intercept) of the linear regression. This fraction, representing the amount of impact on the neighbourhood, is multiplied with the outbound call time of all neighbours N. This is again divided by the length of the analysed time period in days and multiplied with the neighbours expected lifetime using their average age.

$$v_N(x) = \frac{\sum_{n \in N} C_{out}(n)}{T_{analysed}} \cdot L(\overline{age(N)}; k, \theta) \cdot (m \cdot SPV(x) + b)$$

The sum of those two numbers, the customer value without considering neighbours v_C and the expected neighbourhood value loss in case of churn v_N, results in the expected total value of the revenue the customer generates until he churns, also called *Customer Lifetime Value*:

$$CLV(x) = v_C(x) + v_N(x)$$

The underlying function for both of the values, the survival function, can be seen in Figure 9. It depicts the expected development of the customer value and the potential neighbourhood value loss over time, stacked on top of each other.

5 SOFTWARE TOOL

The methods presented in Chapters 2, 3 and 4 were implemented in a tool for social network analysis. It enables the user to visualise the neighbourhood, extract statistics of a customer and to obtain various kinds of information about the network as a whole. This chapter describes the functionality and gives an overview of the components.

5.1 DATA STORAGE

The data is stored in an Oracle database. The phone calls are in a separate tables for each day with information for caller, callee, time and call duration.

To enable the user to "browse" fluently through the network and view statistics of an arbitrary customer, it is necessary to let the program generate certain tables that contain aggregational statistics. The time span used for analysis can be of any arbitrary resolution. The examples in this paper mainly use a seven-day time span, but on other data different lengths might be more adequate. Tables with the following suffixes are generated by the program:

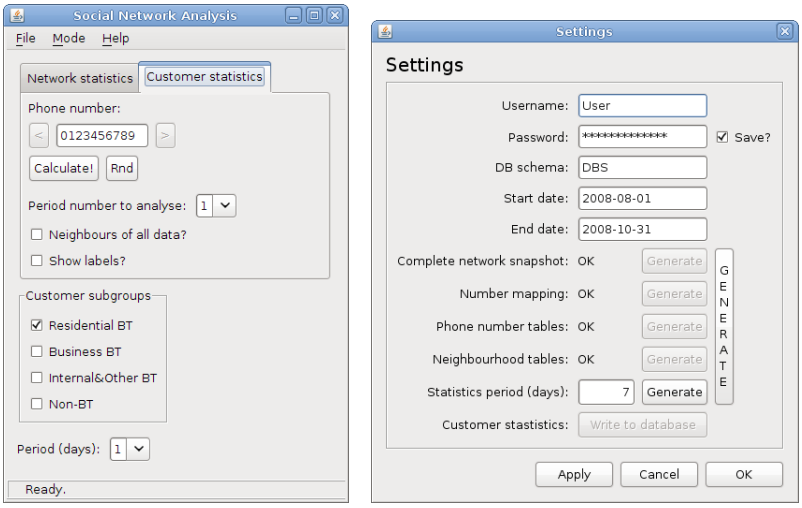
- ALL: View of the calls of all days for queries on the complete stored data
- NEIGHBOURS: Neighbourhood relations for every customer within their subgroup, constructed from all data
- TELS_<SUBGROUP>: Telephone numbers belonging to that certain subgroup (e. g. residential or business customers)
- STAT_<PERIOD>_<NUM>: Aggregational statistics of a period with the given length
- CUSTOMER_STATS: Statistical results for every single customer; this table can be used to import the data in other programs

After these tables are generated, the program can be efficiently used to view network statistics or explore the calling behaviour of a customer.

5.2 SOFTWARE ARCHITECTURE

The program consists of mainly three parts: The database access, the statistical back-end and the graphical user interface.

The database access is abstracted by the classes in the package *sna.backend.database* as shown in A.1. The class *DatabaseCon* manages the connection to the database and provides easy to use functions to establish or close the connection, to execute SQL statements and retrieve the results. If the total size of the result is rather small (less rows than a configurable limit), all rows are fetched from the database and stored as a *SQLResults* class. This result is also saved in a buffer on the hard disc, which utilises a least recently used cache to determine which results to delete if the buffer exceeds a certain size. If the same query is executed



(a) Main window (b) Settings and aggregational statistics

Figure 10: Screenshots of the Social Network Analysis Tool.

again, the appropriate cached result is used instead of actually sending the query to the database. The *SQLResult* class behaves in the same way as a *java.sql.ResultSet*, which enables the programmer to call the methods of *DatabaseCon* always in the same way and benefit of the speed advantage if there is already a cached result present.

The second part of the program is the statistical back-end. The UML diagram of the classes of the package *sna.backend.statistics* can be seen in A.2. It contains methods to calculate overall statistics as well as customer specific ones. The logic for creating aggregational statistics is also stored in this package.

The last part is the user interface, with the package *sna.gui* and the corresponding UML diagram in A.3. The main window is divided into separate parts:

- Controls for network specific statistics (*GuiCtrlNetwork*)
- Controls for customer specific statistics *(GuiCtrlCustomer*)
- Settings dialogue (*SettingsDialog*)
- Graphical representation of a customer's neighbourhood (*NeighbourGraph*)
- Various charts to present the results

The network and customer statistic components are placed on separate tabs. Depending on which one is selected, the corresponding result charts are shown. This is implemented using the docking framework *FlexDock* [10], which enables the user to detach the charts from the main window and place them separately at any position on the screen.

5.3 NETWORK STATISTICS

The first tab of the main window provides options to calculate general statistics about the network. This includes the average call time per customer and its change over time, as well as the possibility to display the highest ranking customers in any of the calculated statistics.

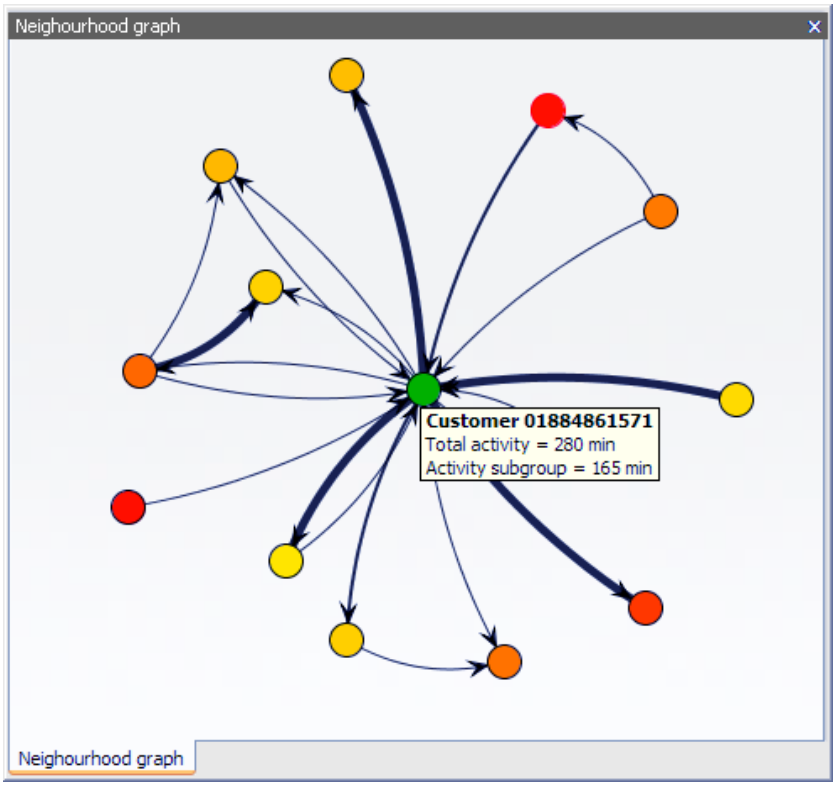


Figure 11: Screenshot of a neighbourhood graph.

Furthermore, the development of churning customers before and after churn can be analysed and displayed as a line chart plotted against time with the point of churn represented by the period with the number zero. The available analytical methods are the neighbourhood value in terms of total call time minutes, which can be normalised to compensate the general trend as described in Section 2.2, the weighted neighbourhood value to equalise the influence of all customers and the average percentage of neighbourhood value change before and after churn between all customers or only between customers of the same type, as described in Section 4.2. This information is valuable to detect the influence churners have on the calling behaviour of their neighbours.

5.4 CUSTOMER ANALYSIS

The controls shown after activating the second tab of the main window (see Figure 10a) enable the user to display a visual representation of the neighbourhood and statistical information about a specific customer.

5.4.1 *Neighbourhood Visualisation*

The neighbourhood graph shows all customers the analysed customer interacted with in the current period. The "Java Universal Network/ Graph Framework" (JUNG, [5]) is used for visualisation of the interaction strength and activity of neighbours. It shows the current customer in the middle as a green dot and arranges the telecommunication partners around him (see Figure 11). The thickness of the edges represents

Statistics	Value
Phone number	0123456789
Customer type	Residential
Acquisition date	1956-12-14
Expected remaining lifetime (years)	9,26
Outbound call time, per day (min)	13,17
Inbound call time, per day (min)	3,35
Total call time, per day (min)	16,52
Fraction of inbound call time of all	0,20
Number of neighbours	11
Neighbour outbound call time, per day (min)	351,73
Average neighbour age (years)	37,01
Social participation	0,05
Customer value, without neighbours (min)	44481,17
Total neighbourhood value (min)	1171649,85
Neighbourhood value loss in case of churn (min)	21491,20
Total customer value (min)	65972,37

Table 3: An example of the calculated customer statistics.

the connection strength measured in terms of call time to or from that neighbour, and the colour of the nodes indicates the relative customer value of his neighbours approximated using their outbound call time. Neighbours with a low value are drawn in yellow, the ones with a high value are drawn in red. A red border indicates customers who churned at some time of the complete analysed time span.

This feature could allow e. g. customer service to gain an instant overview of someones calling behaviour and its change over time. To explore the network, a mouse click on a neighbour displays his neighbourhood. The back and forward buttons help to browse through the network.

5.4.2 *Customer Statistics*

Additionally to the graph, the program calculates various statistics about every customer and displays them in a table (see Table 3). This includes:

- Basic information like acquisition and churn date, telephone number and customer type.
- An expected remaining lifetime is calculated using information about customers that have stayed already the same amount of time.
- The average number of daily in- and outbound phone calls and their duration for interaction with the whole network or just within customer of the same type and the relation of in- to outbound call time.
- Neighbourhood statistics like number of neighbours, number of churners, the outbound call time of the neighbours, average time since acquisition and a value indicating the degree of social participation of a customer with his neighbours.

Using this data, more advanced statistics can be calculated. This includes a total value for a customers outbound call time during his expected lifetime and a similar value for his neighbours. Using the social participation value and the data of already churned customers, the possible impact on his neighbours in case of churn is calculated. This value, also called the expected neighbourhood value loss in case of churn, and the total outbound call time sum up to the total customer lifetime value, as described in Section 4.4. This is presented as a pie chart (see Figure 12) to visualise the proportion between the customers self induced value (the phone calls he initiates) and the network added value that would be lost if the customer churns. The chart at the bottom shows the survival function of the customer value considering only his own calls (red) and his neighbourhood value (blue) to get a better understanding of the expected development of the customer and neighbourhood value over time.

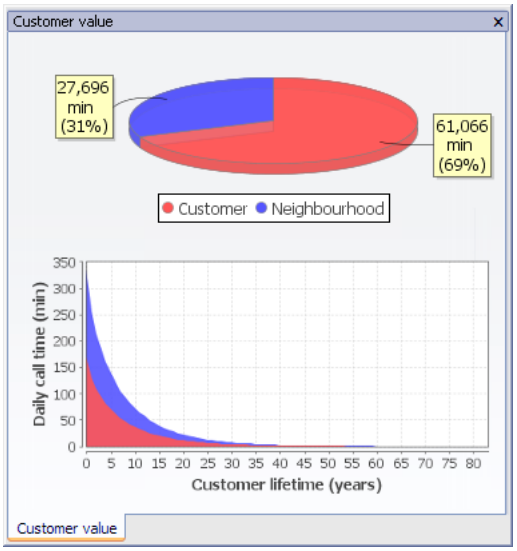


Figure 12: Screenshot of a customer value chart.

6 CONCLUSIONS

6.1 SUMMARY

The preceding chapters presented methods to calculate the value of a customer. It has been shown that network effects have to be taken into consideration if one wants to know his real value.

At first, the time a customer stays with a company was estimated. Statistical analysis of historical data has shown, that these lifetimes, especially values smaller than 60 years, which applies to more than 99.9 % of all customers, follow a gamma distribution very closely.

Additionally to the customers own revenue, the neighbourhood value loss after churn was calculated. Analysis revealed a dependence between an introduced measure called *Social Participation Value* and the value change in the neighbourhood. This measure can be used to estimate the possible impact on his neighbours if a customer leaves the company – a high participation value lead to a large decline in neighbourhood value. This loss, occurred for whatever reason, has to be accounted to that customer's value. Reasons might include a stimulation of others (to place more telephone calls) while being a customer, convincing neighbours to also cancel their contract if churning, moving to another place or even death. Some of the reasons obviously cannot be influenced, but others, like changing the telephone company, could possibly be prevented by giving them special benefits.

Furthermore, the possibility to predict churners using the available call data was analysed. It became apparent that the data didn't seem to provide sufficient information for a reliable classifier. Nevertheless, two values were identified to have an impact: the number of already churned customers in the neighbourhood and the ratio between in- and outbound call time. These values provide valuable information to enhance existing churn prediction systems.

During the work on this paper, a software tool was developed to assist the temporal analysis of telephone call data in special consideration of customer churn. It provides various methods to extract information from a network in general as well as a single customer. The customer analysis gives an overview of his neighbourhood and calculates different figures like his lifetime value. These statistics can be exported to a database table for use by other programs.

Conventional methods, analysing only a customer for itself, underestimate his value. The network-aware value is considerably higher and gives better models of customers, which is a requirement for good customer retention management systems.

6.2 FUTURE WORK

While working on this paper, different areas that need further investigation have emerged.

Analysis suggested, that the probability of customers to churn is higher if there is already a churner in the neighbourhood, but due to the low churn rate, there were only very few samples. To get more

significant results, data with more churners has to be analysed using the presented methods.

The limited time span of the data of three months made it necessary to estimate the development of the neighbourhood value afterwards. As described in Section 4.2, the neighbourhood value decreases to a constant level from four to six weeks after churn. This suggests that the value converges, but further investigation is needed to proof if this is true after the analysed period of time.

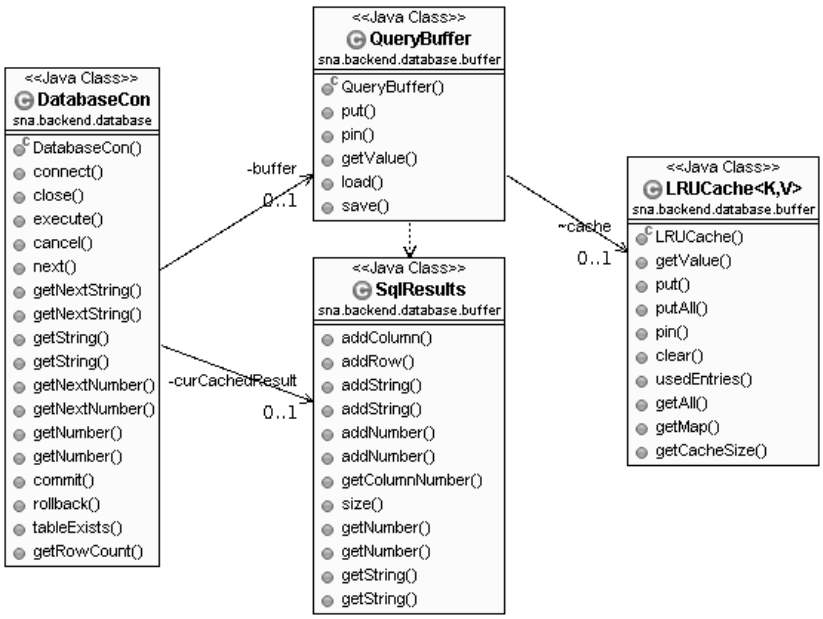
The revenue was estimated by just using data of the amount and length of calls. This is a reasonable if charged per minute, but for more detailed results it is necessary to also take different tariffs, standard charges or free minutes into account.

Finally, using more information about customers than just the call data, could considerably improve the results of the presented algorithms for churn prediction. This might include gender, age or profession, information about billing records, customer service enquiries or any other available data.

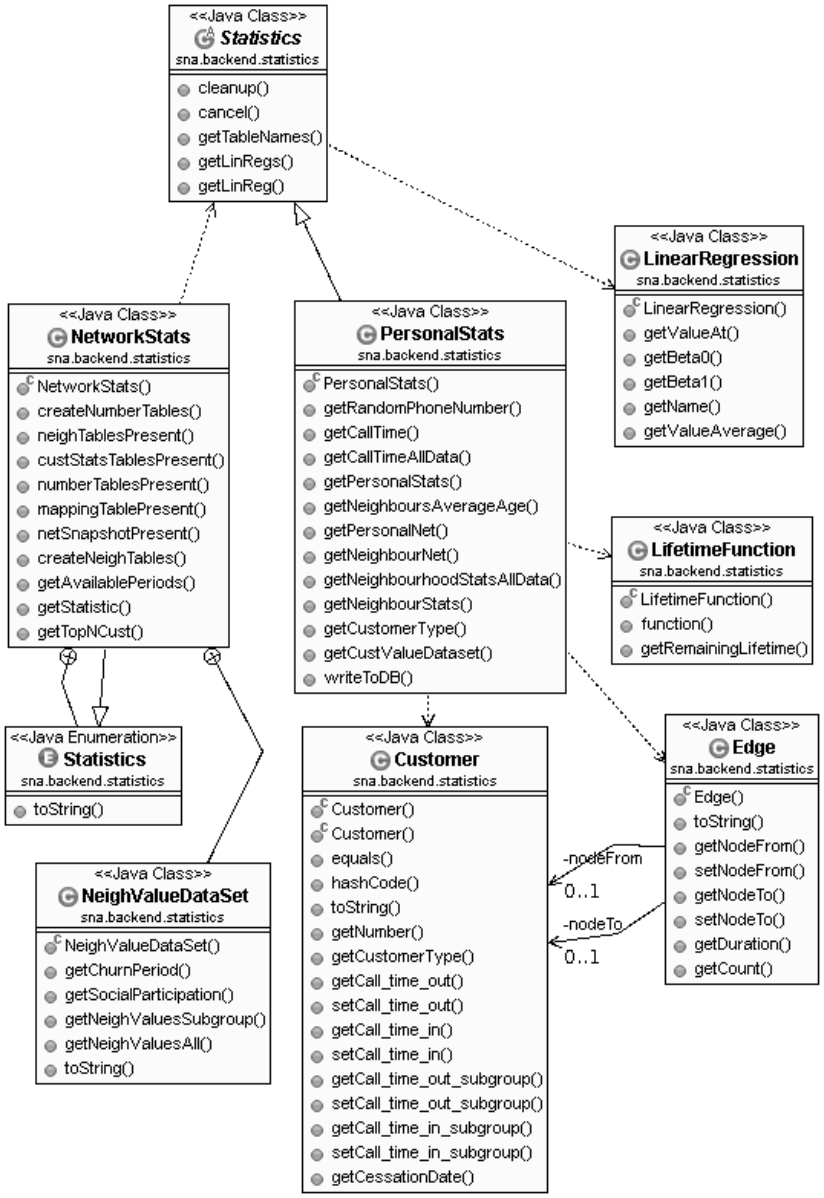
APPENDIX

A UML CLASS DIAGRAMS

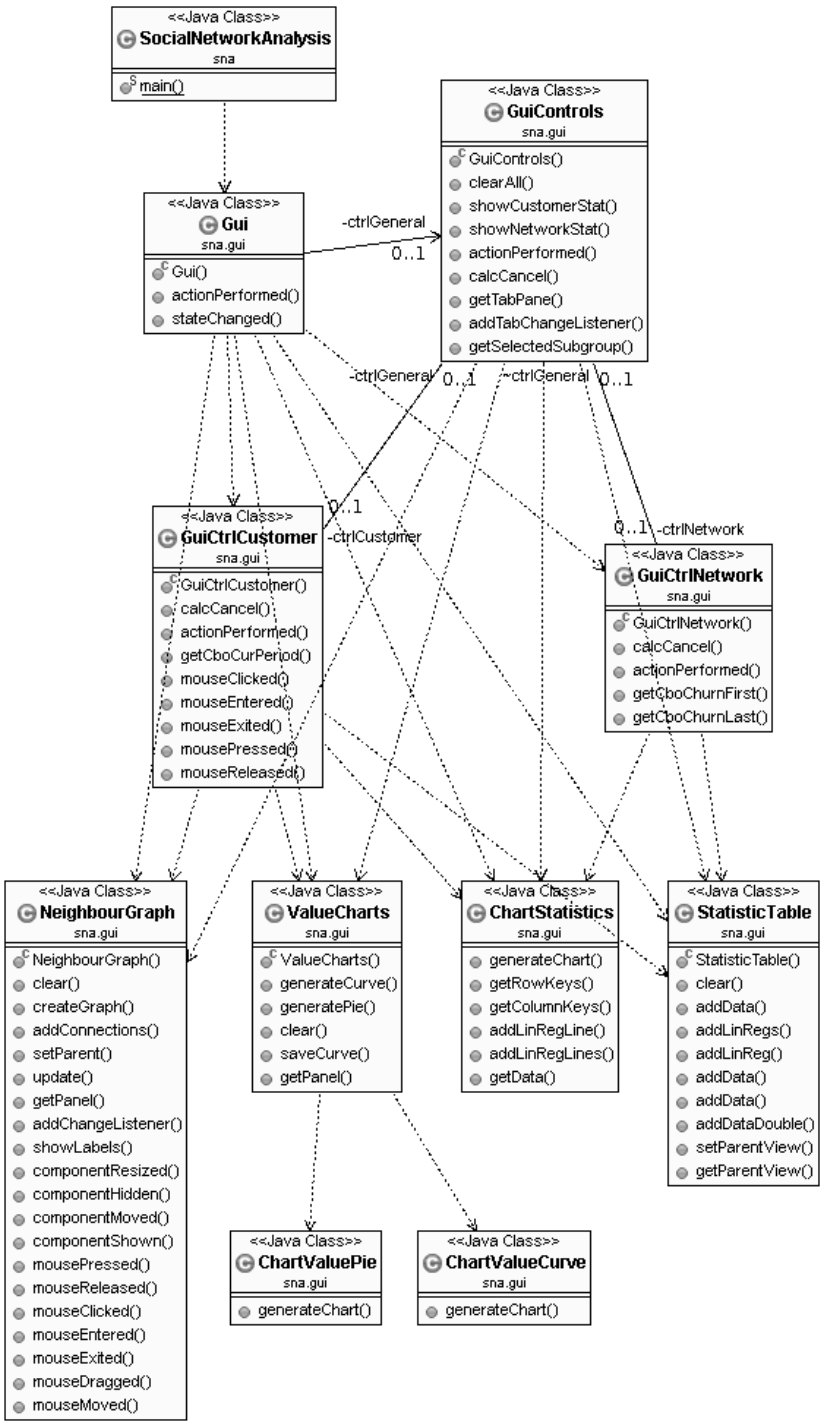
A.1 DATABASE ACCESS



A.2 BACKEND


<<Java Class>>
Statistics
sna.backend.statistics
cleanup()
cancel()
getTableNames()
getLinRegs()
getLinReg()
<<Java Class>>
LinearRegression
sna.backend.statistics
LinearRegression()
getValueAt()
getBeta0()
getBeta1()
getName()
getValueAverage()
<<Java Class>>
NetworkStats
sna.backend.statistics
NetworkStats()
createNumberTables()
neighTablesPresent()
custStatsTablesPresent()
numberTablesPresent()
mappingTablePresent()
netSnapshotPresent()
createNeighTables()
getAvailablePeriods()
getStatistic()
getTopNCust()
<<Java Class>>
PersonalStats
sna.backend.statistics
PersonalStats()
getRandomPhoneNumber()
getCallTime()
getCallTimeAllData()
getPersonalStats()
getNeighboursAverageAge()
getPersonalNet()
getNeighbourNet()
getNeighbourhoodStatsAllData()
getNeighbourStats()
getCustomerType()
getCustValueDataset()
writeToDB()
<<Java Class>>
LifetimeFunction
sna.backend.statistics
LifetimeFunction()
function()
getRemainingLifetime()
<<Java Enumeration>>
Statistics
sna.backend.statistics
toString()
<<Java Class>>
Customer
sna.backend.statistics
Customer()
Customer()
equals()
hashCode()
toString()
getNumber()
getCustomerType()
getCall_time_out()
setCall_time_out()
getCall_time_in()
setCall_time_in()
getCall_time_out_subgroup()
setCall_time_out_subgroup()
getCall_time_in_subgroup()
setCall_time_in_subgroup()
getCessationDate()
<<Java Class>>
Edge
sna.backend.statistics
Edge()
toString()
getNodeFrom()
setNodeFrom()
getNodeTo()
setNodeTo()
getDuration()
getCount()
-nodeFrom
0..1
-nodeTo
0..1
<<Java Class>>
NeighValueDataSet
sna.backend.statistics
NeighValueDataSet()
getChurnPeriod()
getSocialParticipation()
getNeighValuesSubgroup()
getNeighValuesAll()
toString()

A.3 USER INTERFACE



BIBLIOGRAPHY

[1] Ulrik Brandes and Thomas Erlebach. *Network Analysis: Methodological Foundations*. Springer-Verlag New York, Inc., 2005. ISBN 3540249796.

[2] M. S. Finkelstein. On the shape of the mean residual lifetime function. In *Applied Stochastic Models in Business and Industry*, volume 18, pages 135–146. John Wiley & Sons, Ltd., Chichester, 2 edition, 2002.

[3] Geoffrey Holmes, Bernhard Pfahringer, Richard Kirkby, Eibe Frank, and Mark Hall. Multiclass alternating decision trees. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *ECML*, volume 2430 of *Lecture Notes in Computer Science*, pages 161–172. Springer, 2002. ISBN 3540440364.

[4] The University of Waikato. Weka, 2009. `http://www.cs.waikato.ac.nz/ml/weka`.

[5] Joshua O'Madadhain, Danyel Fisher, and Scott White. Java universal network/graph framework 2, 2009. `http://jung.sourceforge.net`, BSD License.

[6] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1 edition, January 1993. ISBN 1558602380.

[7] Dymitr Ruta, Detlef Nauck, and Ben Azvine. K nearest sequence method and its application to churn prediction. In *IDEAL*, pages 207–215, 2006.

[8] Dymitr Ruta, Christoph Adl, and Detlef Nauck. Data mining strategies for churn prediction in telecom industry. unpublished, 2009.

[9] Dymitr Ruta, Przemysław Kazienko, and Piotr Bródka. Network-aware customer value in telecommunication social networks. *The 2009 International Conference on Artificial Intelligence, ICAI'09*, July 2009.

[10] Karl G. Schaefer and Scott Delap. Flexdock 0.5.1, 2007. `https://flexdock.dev.java.net`, MIT License.

[11] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. ISBN 0521387078.

[12] Lian Yan, D. J. Miller, M. C. Mozer, and R. Wolniewicz. Improving prediction of customer behavior in nonstationary environments. In *Proc. Int'l Joint Conf. on Neural Networks*, volume 3, pages 2258–2263, 2001.

DECLARATION

Hiermit versichere ich, dass ich die vorliegende Studienarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Magdeburg, Dezember 2009

Hinrich Harms