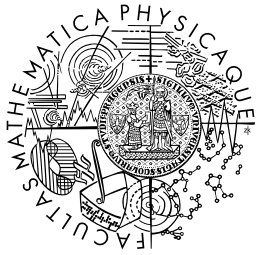

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Michal Lehončák

Analysis of Inferred Social Networks

Department of Software Engineering

Supervisor of the master thesis: doc. RNDr. Irena Holubová, Ph.D.
Study programme: Computer Science
Study branch: Software and Data Engineering

Prague 2021

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

First, I would like to thank my supervisor doc. RNDr. Irena Holubová, Ph.D., who excited me for this topic and provided me guidance on the long path of this research. I would also like to thank my consultant RNDr. Martin Svoboda, Ph.D. and the whole team working on the project for introduction into topic and helpful discussions. I am also very grateful to my parents and my girlfriend for endless support and understanding.
Once again, thank you!

Title: Analysis of Inferred Social Networks

Author: Bc. Michal Lehončák

Department: Department of Software Engineering

Supervisor: doc. RNDr. Irena Holubová, Ph.D., Department of Software Engineering

Consultant: RNDr. Martin Svoboda, Ph.D., Department of Software Engineering

Abstract: While the social network analysis (SNA) is not a new science branch, thanks to the boom of social media platforms in recent years new methods and approaches appear with increasing frequency. However, not all datasets have network structure visible at first glance. We believe that every reasonable interconnected system of data hides a social network, which can be inferred using specific methods. In this thesis we examine such social network, inferred from the real-world data of a smaller bank. We also review some of the most commonly used methods in SNA and then apply them on our complex network, expecting to find structures typical for traditional social networks.

Keywords: social network, social network analysis, graph data management

Názov: Analýza odvodených sociálnych sietí

Autor: Bc. Michal Lehončák

Katedra: Katedra softwarového inženýrství

Vedúci práce: doc. RNDr. Irena Holubová, Ph.D., Katedra softwarového inženýrství

Konzultant: RNDr. Martin Svoboda, Ph.D., Katedra softwarového inženýrství

Abstrakt: Aj keď analýza socialnych sietí nie je nová vedecká disciplína, vďaka prudkému rastu sociálnych médií v posledných rokoch pribúda stále viac nových metód a postupov. Nie všetky datové sady majú sieťovú štruktúru viditeľnú na prvý pohľad. My veríme, že každý rozumný vzájomne prepojený systém dát v sebe skrýva sociálnu sieť, ktorá môže byť odvodená využitím špecifických metód. V tejto práci preskúmame takúto sociálnu sieť odvodenú zo skutočných dát menšej banky. Tiež predstavíme prehľad často používaných metód v analýze sociálnych sietí a následne ich aplikujeme na našu komplexnú sieť, pričom očakávame prítomnosť štruktúr typických pre sociálne siete.

Kľúčové slová: sociálna sieť, analýza sociálnych sietí, správa grafových dát

Contents

1. Introduction

The term **social network** has become a buzz-word in last few years. Many people consider social media and social network as a synonym. However, social media are only a carriers of social networks, as they provide their users a way how to interact with others.

Formally, a social network is defined as "a set of actors and the ties among them" [Wasserman and Faust, 1994]. An actor can be any entity, such as an individual or an organization, with its attributes (name, age, etc.) and ties are social interactions, usually binary, but can be also ternary or composed of even larger groups of connected nodes. Every social network is a graph, but not every graph is a social network, it has to contain observable social structures to satisfy the definition.

The social network analysis (SNA) is a relatively new science discipline with rapid development in past years. Having a social network, with today's tools we are able to analyse not only what happened, but also how and why. We can identify things, which require our attention or even predict future evolution.

We all are parts of multiple social networks throughout all day. Our families, neighbourhoods or workplaces can be examples of such networks. Various collaboration networks (for instance co-acting of actors or mutual collaboration of scientists) are often presented as good examples of social networks because of well-observable social structures. Medicine also often relies on social networks analysis in disease prevention. Recently people around the world observed epidemiologists as they had to make decisions about restrictions or vaccination strategies during ongoing SARS-Cov-2 pandemic based on pandemic development models. Social networks are not only connected to behaviour of humans, animals also create groups with social ties among them, whereas scientists often study community structures and key players of these groups.

Possession of structured data, which often have tabular structure and lack any graph properties, may not imply useful analysis of a "hidden" social network. Many traditional approaches of data mining, concluded e.g. in the book by Han et al. [2011], but also in many others, would be certainly the first choice of every data analyst. However, data of almost any structure can be converted to a graph. Entities are converted into nodes, interactions between them into relationships. Some nodes and relationships can be inferred from properties of original entities, either trivially or after a complex statistical processing.

Constructing such inferred network can lead to discovery of relationships, which are easily observable in an interconnected network, in contrast to tabular or unstructured data, where revelation through multiple joins would be difficult. Graph structure also leads to straightforward visualization, where many intuitive solutions are available. However, an inferred social network is not omnipotent. The process of inference must be careful, as the original data may not be of a high quality, missing values are often nearly impossible to extrapolate or guess, or worse, existing values may be false.

The main goal of this thesis is to analyse given real-world tabular data from the financial domain, namely clients of a small bank and their financial transactions, and to describe the whole process of a social network inferred from them. The

inference itself was conducted as a part of project of the Faculty of Mathematics and Physics and a private company and it is not a part of this thesis.

We will first explore the input data, especially their format and quality, and try to propose suggestions to create actors and ties. We believe that a graph created artificially from strict relational data contains such social structure if the data capture real-world behaviour of social actors. We will then describe these social structures, both in general and at a microscopic level.

We will review various methods and approaches of SNA and discuss their usability on inferred social networks. We will then apply these methods on our network and examine the results, eventually discuss possible improvements.

We will also focus on common problems of a such bank, which would like primarily to maximize its profit and secondarily to gain new clients and bring comfort to its current clients. We will examine many aspects of bank business, from clients and their characteristics, through money transfers and card payments to ATMs and their usage. Some of the problems were originally proposed by Holubová et al. [2019]. We will explore some of them and also propose new ones.

Outline

In Chapter 2 we will introduce our use case, original data and also examples of network inference from this data. Chapter 3 reviews approaches and methods for analysis of a social network, Chapter 4 then offers a short review of the software providing analytical support for the SNA. In next chapters we use this knowledge to examine properties of our model network. In Chapter 5 we break down properties of basic domain entities in our data. Chapter 6 provides deeper analysis of money spending through various media. In contrast to this chapter, the following Chapter 7 is devoted to analysis of income. We also attempt to introduce client characteristics according to his income and identify some natural communities. As evolution is an important part of every social network, we try to describe it in Chapter 8. The final chapter concludes the discussion and offers ideas for future work.

2. Bank Domain and Input Data

Our domain is well-known to everyone, who has ever opened an account in a bank. **Client** is the main entity, (s)he actively creates new relationships, which makes him an actor. Every client must provide basic data, such as date of birth, gender, etc., to the bank during registration. Client interacts with other entities by financial actions, e.g. withdrawals from **ATMs**, shopping in **stores** and **e-shops** or transferring money to another accounts, either internal or external. If the counterparty is internal, we know its owner and money of transaction can be tracked, however, if it is external, we can only guess who is on the other side.

Client can also apply for a **loan** – apart from the fact that a loan produces new transactions between a client and a bank, this application is very important because it requires additional data, which must be provided to a bank in order to judge client's financial abilities. This data, such as the social or education status and number of children, can be useful in large number of different analyses.

Every client also provides his/her address of residency and contact address. The second one appears to be more useful as clients decide not to change residency after they move from parent's house or other long-term residence, while contact address usually changes to current client's location.

2.1 Original Data

Our original data were collected by the bank in a relational database. Timestamps of events and transactions in data range from January 2015 to February 2019. However, bank's lifetime exceeds this period and access to earlier data is considerably limited. Data have the structure of a relational database with many interconnected tables with millions of rows. Database tables are summarized in Table 2.1.

For usage in our analysis, all data are anonymized. Full or partial name of a client is not available, instead, there is a hash of a full name and two clients can be at least compared to each other by this hash. Equality of a part of surname between two clients can imply family relationship between them, along with the same addresses. All addresses, however, are also anonymized, the level of detail of an address reaches only to street, building number is not known. On the other hand, the owner of data could provide similar analysis also with sensitive data legally, resulting in even more accurate results.

Most of the personality attributes are filled in by a client through a form and hence they are quite understandable, e.g., education or marital status. There are also attributes, which are added by the bank, such as risk class or risk score. These attributes, which are supposed to rate client's ability to successfully repay a loan, are evaluated in every bank in its custom way. For instance, in our bank the risk class is a subjective rating by a bank employee, who considers inputs from the mentioned form and possibly other documents delivered by a client, while the risk score is taken from the Czech Credit Bureau[1], which stores data about client's loan history.

[1] https://www.crif.cz/

Name	Count	Attributes	Description
clients	416,254	type, birth year, gender, hashed full name	a client of the bank
client personalities	11,295,369	risk class, risk and behavioural score, education status, social status, marital status, number of household members, number of household children, phone tariff type, row creation date	snapshot of all listed attributes in a point in time
client address history	984,094	type of address, country code, city, zip code, part of city, date	snapshot of address in a point in time
product accounts	971,884	state, type, currency, various dates (sign, start and end, optional repricing and maturity dates for mortgages), number of instalments	all bank products are represented as an account, this table contains loans, mortgages, current and savings account
card transactions	96,387,215	account and owning client, timestamp, terminal info (name, address, merchant code), amount with currency, various type descriptions	all transactions conducted by card - ATM withdrawals and shop payments
money transfers	60,249,552	account and owning client, timestamp, counterparty bank account, amount and currency, symbols, direction	incoming and outgoing money transfers
instalment transactions	4,329,149	account and owning client, timestamp, amount in CZK, direction, type and description of instalment	loan or mortgage payments

Table 2.1: Overview of input database entities

2.2 Preprocessing

The original format and values of data was not suitable to be directly translated into a network. Many irregularities are caused by historical development of the bank, as it was merged with an other bank with its own systems and data having a different structure. Also in the context of one bank, the amount of data collected changes over time, some data may not be available for earlier clients or products. Also, bank's products change over time, some older might be transformed into similar ones with a different name. We created an intermediate database which attempts to include all data available. We will mention at least some of the issues and their solution.

Lack of Normalization

Some tables were meant to cover multiple similar things, which then results into tables with large number of columns, of which majority is nullable and defined only for specific rows. For instance, only approximately 3.5% of client's personality records contain values for all columns, whereas the rest of them, which is over 10.5 million of records, contain only value for client's behavioural score with other values undefined. The reason is that while behavioural score is computed for every client on the last day of every month, all other values are filled in by client only during an application for a loan or a mortgage. Another example is general "product accounts" table, which contains many specialized columns for loans and mortgages, even though the majority of records in this table are current or savings accounts.

New tables were created in the intermediate database, specifically **personalities** and **scores** to resolve issues with original **client personalities**. **Product accounts** were also separated into **loan accounts** and **deposit accounts** with more specific columns for these types of products, resulting in fewer **NULL** values, and an additional common **accounts** table with a relationship to owning client.

Enumeration Values

Another issue is the absence of enumerations normalization. The original data does not contain any enumeration tables, all columns with final and predefined set of possible values were represented by strings, some values with the same meaning changed slightly over time, making simple grouping is not sufficient.

New tables with enumerations were created and every table which contained an enumeration column was remapped to use a reference to these tables instead of original string values. These tables are listed in Table 2.2. Along with proprietary enumerations related to bank's products, there were also general enumerations, such as addresses and merchant category codes, added into the schema.

As for addresses, the bank changed their representation from simple textual data filled in by clients to the official format RUIAN introduced by the ČUZK[2] and therefore standardized and matching to a large number of other systems. Older addresses are therefore less accurate, some of them even unusable. **Mer-**

[2]Czech Office for Surveying, Mapping and Cadastre, www.cuzk.cz

chant Category Codes (MCC) are standardized by ISO 18245[3] and specify an exact area in which a merchant operates. These codes are available for every card transaction and hence can be very helpful in understanding client's expenditure. For simplification of analysis (there are over 500 codes present in our data) and to overcome the fact, that MCCs are not balanced (e.g., while there is one code for supermarkets and grocery stores, every airline has its own MCC), we created a generalized **Merchant Category Class**, which merges several codes into classes with a sufficient level of detail.

Name	Count
proprietary	
insurance types	9
product types	140
transaction channels	17
transaction subtypes	47
transaction types	9
general	
merchant codes	602

Table 2.2: Enumeration tables created from original category column values

Another standardized enums of great importance are **Currency** and **Country Codes**. Although they are also standardized by ISO[4] [5] and the majority of systems nowadays use them, in our dataset we found also some two-letter country codes for older records, which implies either a change of the standard over time or ambiguous use by the two banks before merging.

Missing Entity Tables

This absence of normalized tables, which would contain entities originally present only in transactional tables, does not hold only for categorical columns, but also for domain entities. There exist no tables with all ATMs or stores with additional information, these entities are only represented as subjects of event tables, such as card transactions. To be able to identify every entity with unique ID instead of terminal signature or bank number of a counterparty, entity tables were added, containing both our integral identification and additional data, which can be added by a supervisor with a knowledge of the domain.

2.2.1 Descriptors

In addition to remapped original data mostly composed of events (transactions, product applications, etc.), a new set of statistical tables was created. They contain various aggregations over time for all domain entities. Each type of aggregation can bring some new information, either directly through new relationships in the network, or indirectly, as we can measure the importance of nodes and those with only a small significance may be excluded from network.

[3] https://www.iso.org/standard/33365.html
[4] https://www.iso.org/iso-4217-currency-codes.html
[5] https://www.iso.org/iso-3166-country-codes.html

Overall Summarization Descriptors

These descriptors summarize overall activity of an entity over the whole observed period. As it is the easiest aggregation to compute, it can be the first computed also for a table with a large number of rows and its results can provide hints for detection of less important entities. An example of such descriptor is introduced in Table 2.3. As we can see, for each client with at least one debit transaction there is a row with total number and amount (sum, average, maximum) of debit transactions over whole observed period.

client	txCount	amountSum	amountAvg	amountMax
1	170	1,225,522.05	7,208.95	285,000.00
2	537	1,895,588.00	3,529.96	70,000.00
3	1	840.00	840.00	840.00

Table 2.3: Example of overall aggregation descriptor

If we identify a client, whose overall number of card transactions or money transfers is less than a predefined threshold, or an ATM, for which we have only a few withdrawals, they can be claimed as not important and removed from the further analysis.

Periodical Summarization Descriptors

If there is a grouping by temporal data, we can trace an evolution of data over time. Any granularity can be used, it often depends on the size of the data and a given task. An example is shown in Table 2.4. For each ATM and month, in which there exists a withdrawal from an ATM, there is a number and amount (sum, average, maximum)

ATM	month	txCount	amountSum	amountAvg	amountMax
1	3	1	2,700.00	2,700.00	2,700.00
2	2	5	33,000.00	6,600.00	15,000.00
2	3	4	12,000.00	3,000.00	7,000.00

Table 2.4: Example of a descriptor with aggregation by month

Histograms

The same information, but in a different format is present in histograms. A temporal value is not stored in the value of one column, but there is a column for every value. For instance, there are 50 value columns in tables aggregating by months of the observed period.

cntParty	m1	m2	m3	m4	m5	m6
1	3,857.00	4,250.00	4,250.00	4,120.00	4,250.00	4,120.00
2	*NULL*	*NULL*	*NULL*	86,300.00	*NULL*	12,500.00
3	*NULL*	6,700.00	*NULL*	1,447.00	1,253.00	*NULL*

Table 2.5: Example of histogram descriptor

In an example in Table 2.5 we can see a smaller histogram for a period of 6 months. For each counterparty (internal and external) and month in the observed period (here 6 months) there is a table with nullable columns, where there are average amounts of money transferred by counterparty in particular months.

For a more detailed temporal dimension there would be too many columns, which would be impractical. However, we can use the effect of periodicity and overlap intervals in which actors behave similarly. For instance, if we use weeks as the temporal value and years as the period, we obtain 52 value columns, with a value for a week being the average of all analogical week's values throughout all observed years – this table would be useful for week comparison within a year (e.g., summer vs. winter weeks). Similarly, another such histogram, which proved to be useful, was an aggregative histogram by hours of a day (24 columns) or a week (168 columns). Although such aggregations are useless for mortgage or savings sector, usage for card transactions is straightforward – some ATMs can be then characterized as "office ATMs", if the amount of money is greater in lunch time of workdays, and some as "friday night ATMs", if there are significant peaks in these hours.

2.3 Inferred Network

All tables mentioned in the previous section can be used for inference of the network. Apart from direct factual data about clients and other bank-specific entities and transactional data, we can also use third-party data from various available sources (e.g., demographic data, public company registries, etc.).

2.3.1 Nodes

The main entities of the domain can be directly translated into nodes. As they are a part of our original or enriched intermediate databases, the translation is quite natural – for every row with its attributes there is a node with properties. Client, counterparty and account node types are examples of such cases. Some of the enumeration properties may be detached and form separate node types. The possession of the specific property value would result in relationship with the node representing the value. For instance, we can create an address node type, or even better, a separate node type for every item of an address. A client would then be connected to the country, city and street of his/her residence and particular analysis would pick its relevant detail.

On the other hand, artificial entities can be also created. The **employer** entity is a good example – supervised analysis of counterparties and their transactions showed that some counterparties transfer money to clients on a regular basis and that these transactions have often the variable or constant symbol. Although we can claim that these counterparties are companies employing our clients, we do not know anything else about them. Another example of artificially created entities are ATMs, e-shops and stores, which are created from terminal data of card transactions – the MCC, name and address of terminal provide hints and supervised analysis can distinct the type of terminal.

2.3.2 Relationships

Three types of sources can be used to infer relationships. *Rule-based* relationships are created directly from input data, either factual, or descriptors made by data aggregation. We predefine rules for extraction and apply them on data, which produce various kinds of relationships. They can differ in their temporal validity, as some of them may be definitive (e.g., family relations) and some may have a limited validity (e.g., an address of residence, social status, etc.). Reliability of these relationships also differs – while some of them can be almost certain, as a client's year of birth or ownership of an account, some may be questionable, for instance the fact that a client is employed in a company. Reliability is also dependent on the quality of the source data – older data without the use of standardized inputs and then transformed are less reliable than recent data, where client's options were limited by a list of options to choose from, e.g., when filling a form.

The second type of extraction is *similarity-based*. New relationships are created for two entities, which behave similarly. The similarity is measured by similarity of histograms introduced in the previous section. These relationships are indirect and therefore less reliable, as they are only an approximation of reality.

The third and the least reliable are *probability* relationships. They assume the real-world relationship from the behaviour of two nodes, which in fact may not be true. All such relationships have a probability assigned, which is computed by specialized algorithms analysing certain aspect of behaviour. For instance, there is a **household** relationship between two clients, which are believed to live together, inferred from the gender, address, age and mutual money transfers. Another example, which we use in larger extent in our analysis, is a **friendship** relationship, based on card transactions at stores and ATMs conducted at the same time – the more transactions two clients have in common, the more probable their friendship is. However, to achieve a good reliability of this relationship, we have set a high threshold for its probability, resulting in more reliable, but also fewer, "friendships" and "households".

2.3.3 Grouping

Granularity on the lowest level with individual clients or ATMs may be helpful for a detailed analysis, but not suitable for overall network discovery. Above the basic structure of nodes and relationships we can create groups of nodes, which will bear some characteristics common for all its members. For instance, we can group together all clients between 25 and 30 years old or all ATMs in Prague. Descriptors may be used to divide employers according to their average salary or number of employees. These predefined *buckets* can make some analyses simpler and more efficient, as we deal with fewer numbers of relationships to clusters instead of a detailed network with original relationships. In other words, we can work at different levels of detail.

2.4 Bank World Tasks

Having this conclusion, we can propose some real-world questions, which are likely to be asked by the analysts of a bank. We can even go further and expand these proposals to sensitive data, which are not available for our purpose, but bank analysts could use them.

Income Analysis

Banks often try to characterize client's income and its development in time to predict client's behaviour. Usually, they attempt to rate client's income with a number or a class, but it is quite difficult. We will try to create groups of clients with similar income over time, so that membership of this group will describe a client similar to other clients within the same group. A client with no income may be transferring his/her main account to another bank and our bank probably want to change his/her mind, a client with low income might be interested into loan and finally, a client whose profits gradually grow may want to create an investment portfolio.

Card Payments and Their Purpose

More and more people use card payment instead of cash nowadays. Exploration of client's expenses in these categories and its development in time can bring new view on client's behaviour. The bank may provide specific vouchers for shopping in stores, in which a client may be interested and receive small fees from every transaction.

Card Transactions in Foreign Countries

Besides having assigned its category, every terminal also provides information about its geographical location. Through observation of payments and ATM withdrawals client's movement both inside and outside the domestic country can be tracked. Bank then can simply identify, e.g., people, who travel every summer on vacation by the sea, but also those, who travel often on business trips. Additional products, as account in foreign currency or travel insurance, may be then targeted for chosen people.

3. Social Network Analysis

Since an analysis of networks requires knowledge from many other branches, such as graph theory, statistics, sociology and others, a whole new branch of science, social network analysis (SNA), emerged. It attempts to characterize the structure of a network, to explain its evolution in past and also to predict future development.

3.1 Overall Network Structure

Some properties of the network can be determined only from the structure of its nodes and edges. Distribution of edges can give hints about network's behaviour in certain situations.

Random Networks

The simplest networks that can be found are **random networks**. The idea is to distribute relationships randomly among nodes of a network. Random networks were studied and defined as a model by Erdos and Rényi [1959].

All nodes have the same probability p of attaching an edge to another node. Probability, that a node has exactly k edges can be expressed as:

$$p_k = \binom{N}{k} p^k (1-p)^{N-k-1}$$

where N is total number of nodes, this probability follows Binomial distribution. For larger and sparser networks this distribution can be approximated with Poisson distribution.

Networks, which we usually see in practice, however do not follow this random network theory.

Scale-free Networks

Many scientists on the edge of millennium, such as Barabási and Albert [1999] and others, proposed, that real-world networks does not have Poisson distribution of degrees, but distribution following the power law. In context of the degrees it means, that vast majority of nodes has a small degree and as the degree of a node grows, number of nodes with this degree significantly decreases.

Networks with this property are called *scale-free* networks. The main difference between random and scale-free networks is the presence of nodes with high degrees, called *hubs*. While they form the core of scale-free networks and contain most of the flow in it, hubs are very rare in random networks, where most of nodes have a comparable degree close to the mean of distribution.

The scale-free property gives us hints about the network itself, and about its future development as well. The network with this property is more resistant to disconnection of random nodes, as shown by Callaway et al. [2000], and remains connected, on the other hand, breakdown of network is caused by failure of only a few nodes, if chosen wisely.

However, the idea that most real-world networks are scale-free was questioned recently by Broido and Clauset [2019], as requirements for the scale-free property were not confirmed in majority of real-world networks.

In general, determining of node degree distribution can reveal possible presence of scale-free property in network.

3.2 Key Players

As networks contain many nodes and relationships, it is important to be able to differentiate their importance in network flow. Identifying important nodes can be crucial when performing actions, as we can concentrate on this nodes minimizing overall cost. For example, during an epidemic situation scientists can slow the spread of a disease by distributing protective gear to people, which have the highest probability to infect others. Or, luxury clothes brands distribute their products to important individuals, e.g. celebrities, which would make people following them to want and buy the product.

Centrality Measures

There are more views on node importance. The most straightforward, *degree centrality*, is based on a simple metric from the graph theory, degree of node, in directed networks often separating in-degree and out-degree. If the weight of edges is measured, then the degree is equal to the sum of weights of edges. If not, the degree of a node is simply the number of its edges. This centrality can provide basic view on the close neighbourhood of a node, for example how many nodes are directly affected.

From the aspect of a network flow we want to detect nodes, which are important for data transfer in a network by participating in shortest paths between nodes. This amount of importance is measured by the *betweenness centrality*, computed for a node as a proportion of shortest paths in the network containing the node to all shortest paths in the network. The node itself does not have to have many connections, but the amount of data flows can make it important. This centrality can be measured also for edges, determining the importance of an edge for flow in the network.

The measure of reach is captured by *closeness centrality*. Higher values indicate that the average shortest path to other nodes is short. This measure can be useful when trying to spread information throughout the whole network, for example when trying to publish a message through media to every person.

3.3 Structural Node Equivalence and Similarity

In Section 2.3.2 we introduced the similarity of value histograms as a method of the social network inference. In addition, there is the *structural similarity*, which is based on comparison of relationships between a node and its neighbourhood.

The core idea behind structural similarity is a question, whether two nodes have so much in common, that they are practically interchangeable. For example, if all members of a certain group of clients shops in two grocery stores, these stores

are interchangeable (we can swap their names) and behaviour of the clients will stay the same. This idea leads to the definition of *structural equivalence* [Hanneman and Riddle, 2005], according to which "two nodes are structurally equivalent, if they have the same relationships to all other nodes". As this assumption is often hard to achieve in real-world networks, another extensions have been introduced.

To relax this assumption authors also proposed to change the nature of equivalence from binary to nominal – it allows us to ask not only whether two nodes are structurally equivalent, but also "how much are they equivalent", leading to structural similarity.

There are several measures evaluating structural similarity, although the common idea for all of them is that two nodes are more similar, if they are both connected to some third node of arbitrary type, and similarly, their similarity decreases, if the first one is connected to a third node and the second one is not – this third node creates the difference of the neighbourhoods.

For unweighted graphs there exists a very simple approach of obtaining such similarity, as a ratio of number of nodes, which have the same tie to our evaluated nodes – both present or missing – and the total number of all nodes in the network. However, this measure performs poorly for very large sparse networks, as all nodes with possibility of a tie must be taken into account.

Another very popular similarity measure used in social network analysis is *Jaccard index* Jaccard [1912]. It is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A and B are neighbourhoods of two evaluated nodes. Since it ignores nodes, which are connected to neither of nodes, it yields good results on large sparse networks, as absence of a tie is often irrelevant and is not a sign of similarity.

Considering also weights of relationships, a straightforward idea of distance is often used, as relationships of a node with all other nodes can be represented by a vector of real numbers and we can compute their distance in Cartesian coordinates, e.g. the *Manhattan*[1] or the *Euclidean*[2] distance.

Various generalizations of the Jaccard index have been introduced. For example, Ružička [1958] proposed an index in which one-complement is known to be a distance function, given by

$$Ru(A, B) = \sum \frac{\min(a_i, b_i)}{\max(a_i, b_i)},$$

where A and B are vectors with real $a_i, b_i > 0$. The value of a_i then determine how much does item i correspond to group represented by the vector A, unweighted Jaccard index corresponds to binary representation of a set and the membership in it.

[1]Paul E. Black, "Manhattan distance", Dictionary of Algorithms and Data Structures, ed. 11 February 2019, https://www.nist.gov/dads/HTML/manhattanDistance.html

[2]Paul E. Black, "Euclidean distance", Dictionary of Algorithms and Data Structures, ed. 17 December 2004, https://www.nist.gov/dads/HTML/euclidndstnc.html

3.4 Small Worlds

Small-world phenomena was first introduced by Milgram [1967] as a sociological problem of connectivity in large networks. In a network having this property an average length of the shortest path between two randomly selected nodes is quite short. There are two conditions, that small-world network must fulfil: high clustering coefficient and low average shortest path.

Clustering coefficient denotes how many nodes tend to create clusters. It was originally proposed by [Luce and Perry, 1949] as global characteristics given by

$$C = \frac{\#of\ closed\ triplets}{\#of\ all\ triplets},$$

where triplet is a group of three connected nodes, which are considered as open triplet if they form a path and closed if they form a triangle. Another definition was introduced by [Watts and Strogatz, 1998], who defined *local clustering coefficient* as the measure of connectivity for a node and its neighbourhood, given by

$$C_i = \frac{|\{e_{jk} : \exists v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)},$$

where i is the measured node, N_i is its neighbourhood and E represents all relationships in the network. In other words, it is equal to the number of pairs from node's neighbourhood, which are also connected. Watts and Strogatz enhanced their definition to a global coefficient given as an average of all local coefficients.

3.4.1 Extension for Multi-mode Networks

Our model network contains multiple node and relationship types. The basic definition of a small-world property, however, applies only to one-mode networks, which are those containing only one node type. In a two-mode network, there are two possible approaches - either to project a two-mode network to one-mode by contraction of heterogeneous relationships to homogeneous (if two nodes are connected to a node of different type, we will connect them in the contracted network), or to apply some enhanced definition of clustering coefficient.

3.5 Community Detection

Community detection is another important part of social network analysis. It helps us to understand the structure of nodes in a network. It is analogous to cluster analysis from the graph theory and also uses similar detection techniques.

3.5.1 Community

Community in SNA is usually defined as a group of nodes that have higher probability to connect to another node in the same community than to one from an other community [Barabasi, 2015]. Relationships amongst nodes in one community are usually stronger than those connecting two different communities. Relationships between communities are called *bridges*.

Community can be *explicit*, when an actor acknowledges membership in it, e.g. a community consisting of colleagues in one company or of former high school classmates, and *implicit*, e.g. a community of people similar in their financial transactions or food preference.

In real-life networks one node can be a member of many communities. For example, a person in social media can belong to a group containing members of his neighbourhood, a group of his classmates and many others. However, in community detection techniques we usually want to create a partition of network, where one node belongs exactly to one community.

3.5.2 Detection Techniques

Several approaches have been used for detection of communities in social networks, each of them relying on different principles. Although some of them can yield good results for smaller networks, they are impractical for usage in larger or real-life networks.

As all networks are graphs, we can use algorithms developed for graph analysis in networks too. Despite they do not yield optimal results, they can be really helpful in an early analysis of a network.

Connected Components

One of the simplest solutions is to identify connected components, which are groups of nodes, where there exists a path from a node to any other node. For directed graphs we can consider only directed paths between nodes, then we obtain *strongly connected components*. If we ignore direction of relationships, we obtain *weakly connected components*.

This simple and straightforward algorithm can divide the whole network into smaller subnetworks with no edges between them. We can then discard nodes or groups of nodes, which are not connected to the largest connected component of the network, and hence we do not have to consider them in further community analysis, as they would be a part of small disconnected communities.

Hierarchical Clustering

As the concept of clusters in graphs is very similar to the concept of communities, we can use well-known clustering algorithms. There are many approaches to divide nodes of a graph into clusters, but most of them are based on positions of nodes in a space. However, in our network we do not have positions of nodes in a space, only connections between them.

Only *hierarchical clustering* [Tan et al., 2019] allows us to use relationships between nodes for separation into communities. However, this method requires a metric function, which would give us distances between nodes. We can try to use inverted edge weights as distances - for strong ties we have closer nodes, for weaker ties nodes, which are far from each other and zero similarity for no tie at all. A problem is that this distance measure is not metric, because it does not satisfy triangle inequality.

Girvan-Newman Algorithm

As clustering is developed primarily for graphs, where nodes are placed in some space and any distance metric in this space efficiently creates clusters, it does not suit for our case. We have to use similarity function adjusted for our problem. The algorithm, proposed by Girvan and Newman [2002], combines an approach of this algorithm with the idea of centrality. The core idea uses top-down hierarchical clustering, where in the beginning all nodes are a part of one community and in every step, one community is separated into two smaller. The difference is that weights of all edges are in each step replaced by their betweenness. Edges with higher betweenness tend to be community bridges, because a lot of flow between these two communities runs through them.

However, in every step we have to compute edge betweenness for all edges affected by the previous step, which leads to high computational complexity (worst case $O(m^2n)$) and allows one to use this algorithm only on small or medium-sized networks.

3.5.3 Modularity

Another category of algorithms for community detection evaluate the measure of "how good are communities detected so far" and try to optimize this measure by small changes in assignments of nodes to communities. The most common measure of community partition quality is *modularity* (Newman and Girvan [2004]). It is defined as the number of edges falling within communities minus the expected number of edges if they were placed at random:

$$Q = \sum_{i=0}^{k} (e_{ii} - (\sum_{j=0}^{k} e_{ij})^2),$$

where k is the number of communities in current division and e_{ij} is fraction of edges going from community i to community j. This measure lies in range $\langle -0.5, 1 \rangle$. Positive values indicate possible presence of community structure in network, authors state that most networks with community structure have modularity in range from 0.3 to 0.7, higher values are rare. Zero and negative values on the other hand ensure no presence of communities.

Optimization Methods

There are various methods of modularity optimization. As this task is an optimization of a function, any generic technique for approximating the global optimum can be used. For instance, *simulated annealing* [Kirkpatrick et al., 1983] can be used to search the state space of all assignments of nodes into communities, combining exploration and exploitation to ensure yielding at least sub-optimal solution. One of the approaches using simulated annealing was proposed by Liu and Liu [2010], which, in combination with the *k-means clustering algorithm* [Tan et al., 2019], can yield very good results with great efficiency.

On the other hand, there are algorithms, which aim to find better solutions with the knowledge of this particular problem. The simplest one is the greedy optimization introduced by Clauset et al. [2004]. Although some optimizations and improvements have been proposed, the basic principle stays the same. In the

initialization phase, every node is placed in its own community. Then, iteratively, we identify two communities, whose merging would bring the biggest increase in modularity of network. However, not only this algorithm can yield solution with lower similarity than those produced by simulated annealing, its computational complexity is not suitable for large-scale networks. Ineffectiveness of this algorithm has been discussed by Wakita and Tsurumi [2007] and crucial parts have been improved.

Louvain Algorithm

Another method has been proposed by Blondel et al. [2008]. The initialization phase is the same as in the greedy one, every node is assigned its own community. The core of algorithm is different: it comprises of two phases, which are iteratively repeated. In the first phase, for every node and all its neighbours the possible *gain in modularity*, if the node would be removed from its community and placed in community of the neighbour, is computed. If this gain is positive, the node is moved. If there is no positive gain, the node stays in his current community. After all nodes are processed, the second phase is executed. In this phase all communities are merged into nodes of a new network – for all edges between nodes of two communities an edge in the new network with a weight equal to the sum of weights of these edges is created (edges between nodes in one community result in self-loop edges in the new network). This new network is then used in the next iteration of the algorithm.

As the algorithm is iterative, we have some intermediate results after each iteration – the collection of communities represented by nodes created in the second phase. As all communities created in the first phase are optimal, the algorithm provides multiple layers of community structure in the network.

The biggest advantage of this algorithm is its speed, as it happens to be linear in the number of edges on both dense and sparse networks. As shown by the authors, the algorithm outperformed all modularity optimization techniques in performance and the modularity of optimal community structure.

4. Prerequisites

As we aim to build heterogeneous network with millions of nodes and relationships, simple solutions working with the whole network in memory do not suffice. We have to use one of available NoSQL databases with support for effective graph-like data storage.

There are several graph databases available with suitable licensing, as we do not need a production-ready application, we do not require capability of sharding, horizontal scaling or replication. We chose Neo4j[1] after functionality considerations and multiple comparisons, e.g. [Fernandes and Bernardino, 2018] (since 4.0 Neo4j, sharding is also supported).

Neo4j

Neo4j is a single-model graph database, which can be more optimized on graph-like data, because of storing values directly as nodes and relationships and extensive usage of indexing. It comes with powerful and intuitive query language Cypher[2] and it has a flexible data model, which is useful for storing intermediate results as stand-alone nodes.

Graph Data Science Library

Although Neo4j provides a large number of basic mathematical and statistical functions, to perform extensive network analysis we have to use specialized methods introduced in the previous chapter. The authors of Neo4j provide a rich, well-documented plugin containing various algorithms, the Graph Data Science library (GDS)[3], available in open source Community Edition.

This plugin contains most SNA algorithms introduced in Chapter 3. It offers several community detection algorithms (Louvain, Girvan-Newman, etc.), various methods for computing similarity of two nodes (Jaccard, Cosine, etc.) and many others.

All algorithms are computed in-memory, which in combination with good memory management ensure fast execution of most algorithms. On the other hand, the limitation on the size of input may cause that not all input data are accepted and must be trimmed.

APOC

Another plugin, which provides great extensibility to the Neo4j database, is APOC[4]. It was created to serve as a standard library providing various utilities, which are not distributed with the Neo4j. APOC offers wide range of tools for efficient network manipulation, e.g., **periodic.commit** or **periodic.iterate**, which are able to split the large graph manipulation into batches. It also contains

[1] https://neo4j.com/
[2] https://neo4j.com/developer/cypher/
[3] https://neo4j.com/docs/graph-data-science
[4] https://neo4j.com/developer/neo4j-apoc/

useful data structures and tools for the import and export of data. Due to the complexity of the network this plugin is essential for graph structure manipulation.

Powerlaw Library

As GDS, version 1.5, does not contain any methods for recognizing the degree distribution of nodes, we have to use a different solution. Our goal was mainly to verify the power law because of the several subnetworks with the long-tailed distribution. **Powerlaw** library, created by Alstott et al. [2014] for Python, is able to compute optimal parameters for multiple long-tailed distributions. Optimal fits can be then compared and we can say whether one distribution is significantly better than other. Apart from the comparison it also yields parameters of the possible distributions.

5. Overall Network Properties

Before we start with a more complex analysis of the network, we have to inspect the basic structures of entities in our network. We will examine the demographics or educational properties for the **client** entity and also various products of the bank and their usage over time. Then we will be able to study deviations present in certain parts of the network.

5.1 Clients

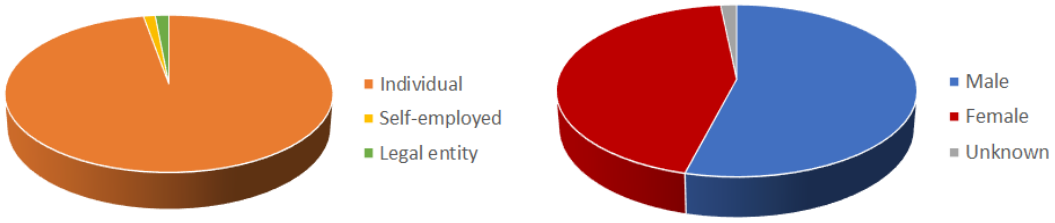

Figure 5.1: Structure of clients by type (left) and by gender (right)

The bank had at the time of capturing the data about 406,521 clients. Each client has to fill in at least some information, regardless the purchased product. The required data include the full name, the date of birth and the gender. However, due to the anonymization of the data the name and the exact date of birth are not known. The only property based on the purchased product is the client's type – individuals have an access to retail products and legal entities or self-employed ones to commercial products.

In Figure 5.1 we can see the distribution of the type and gender. The vast majority of clients are individuals, there is only small amount of self-employed (5,284 $\sim$ 1.2%) and legal entities (6,179 $\sim$ 1.5%). Although some individuals may be in fact self-employed, if they use only retail products, they are classified as individuals. From those who filled in their gender (all clients except legal entities and a few individuals) there are more male than female clients (54% vs. 44%).

As we can see in the structure by gender and age in Figure 5.2, the biggest differences between genders are among the clients of age between 25 and 50, the other age groups are more balanced. The age structure mostly copies the overall structure of the Czech Republic[1], with exceptions in young age (only adults can be owners of a bank product) and old age (bank's products are more oriented on clients in young and productive age).

In Figure 5.3 we can see regional distribution of clients by addresses, which they filled in during the registration process. The number of clients is normalized by the total number of clients living in the district[2]. As we can see, the bank

[1] https://www.czso.cz/staticke/animgraf/cz/index.html

[2] Czech Statistical Office, https://www.czso.cz/csu/czso/population-of-municipalities-1-january-2019

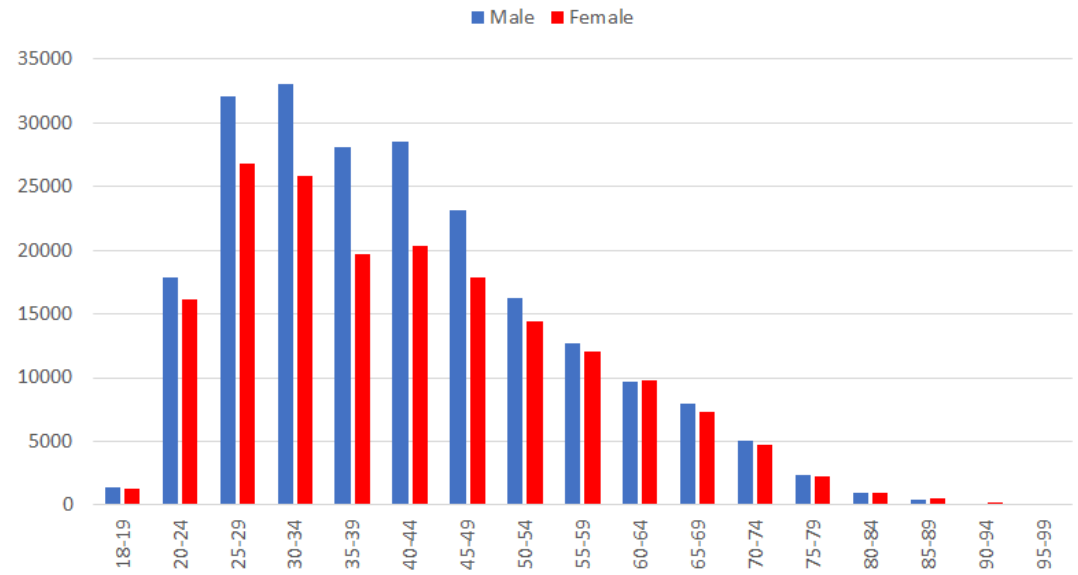

Figure 5.2: Structure of clients according to their age and gender

has the most clients among the population in the district of Mlada Boleslav (nearly 6%) followed by large cities (Brno 5.6%, Ostrava 5.5%, Prague 4.9%), and districts in Silesia (Karviná 4.8%, Opava 4.1%). On the other hand, the bank is less popular in the most of the countryside, the lowest ratios of inhabitants of the district are in Cheb (1.3%), Písek and Strakonice (both 1.5%).

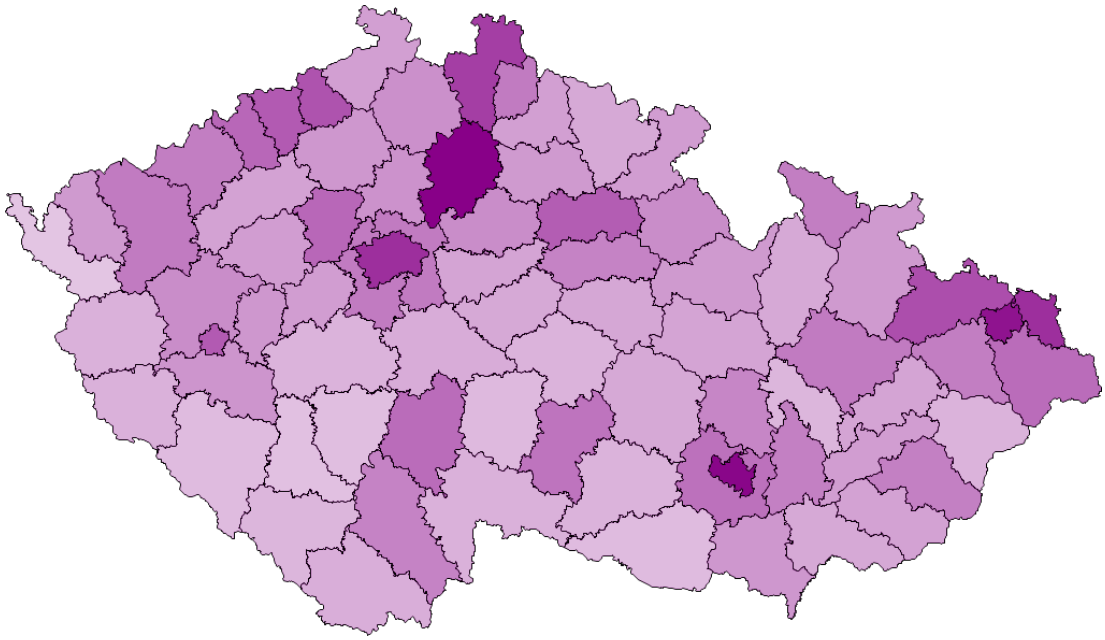

Figure 5.3: Map of districts of the Czech Republic with their colour being the ratio of the number of clients and the total number of people living in them

Categories

Client's personality profile, filled in during the application for a loan or a mortgage, can bring new view on socio-economic structure of clients. However, these

structures may not be a reliable description of all clients, as they are only available for clients, who applied for the mentioned products. Personalities of other 65% of clients remain hidden. Therefore, it is in bank's interests to convince these clients to take a loan or mortgage – to make a profit and to gain new data.

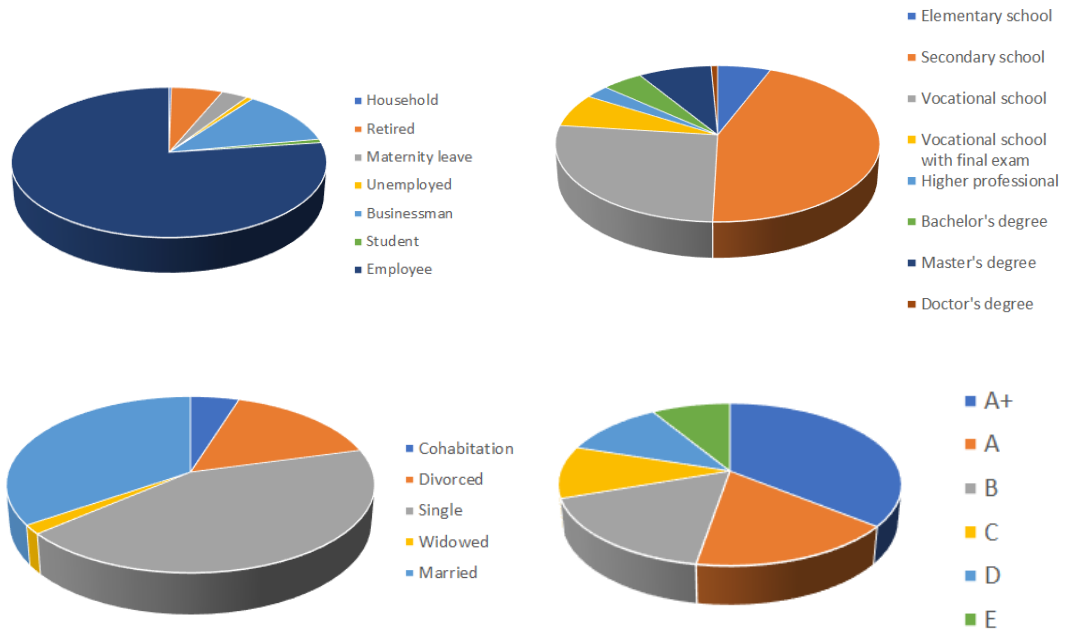

Figure 5.4: Structures of clients, who provided personality details purchasing a product – social status (top left), education status (top right), marital status (bottom left) and risk class (bottom right)

5.2 Bank Products

The bank offers a big variety of products. Some of them are more profitable, like loans or mortgages, others have more supplementary role, like savings account. Mortgages and loans, on the other hand, create the risk, that the client will not be able to pay the instalments, which the bank has to take.

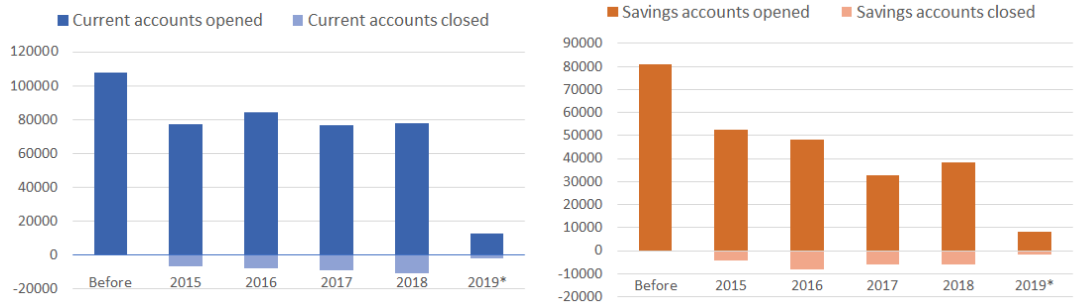

Figure 5.5: Overview of the numbers of current and savings accounts, which were opened or closed in particular year (only first two months of 2019 are included)

A current account is an elementary product of every bank. While it may not generate a large profit, it serves as a gate to other bank's products. All

individual clients, but also companies, have to open an account nowadays. The big advantage of the current account is that it provides behavioural information about client's life, via either transfer or credit card transaction data. The savings account, on the other hand, does not usually contribute to client's behavioural profile, although it may increase his/her reliability through regular money saving. The bank offers unlimited saving with optional termination at any point, but also term deposits – accounts with fixed date of maturity. The term deposit is also more likely to prevent the client from leaving the bank.

Figure 5.5 shows the total number of opened and closed deposit accounts in span from 2015 to February of 2019, we can also see the number of accounts opened before the observed period. During the observed years, the bank has a stable increase of the number of opened current accounts every year. Although the number of opened savings accounts decreased in 2017, it again increased in 2018 and a simple projection for 2019 suggests an increase to values in the strongest years.

We can also see the amount of closed accounts in the negative values. The amount of closed current accounts is about 8% of opened accounts in 2015, but reaches up to 13% in 2018. Similarly, the number of closed savings account starts at 8% percent in 2015 and rises to 16-17% in the following years. However, the number of closed savings accounts is affected also by the maturity of term deposits, when the account is closed naturally. If we ignore these deposits, the ratio of closed savings accounts drops to 10%.

Closing of the client's account is certainly one of main concerns of a bank. If it could recognize the signs, which precedes the closing, it could try to prevent it. We will analyse client's behaviour before closing of his/her account in further chapters.

Credit Products

As for loans and mortgages, it is even more important that the corresponding account is not closed preliminary. When such product is closed before its maturity date, either the client defaulted and the bank's risk has not paid off, or (s)he paid it all before the maturity date and the bank will not get the interest money for the remaining period. It is crucial that the bank predicts at least the first case or at least recognizes the behaviour which it preceded. It can then identify similar clients.

Although a mortgage is only a special case of a loan, it is specific in many ways – the amounts are significantly higher, the maturity periods longer and it is always bound to a property. Therefore, the demand for mortgages is in general lower than the demand for loans, which are short-term and can be used for any purpose. Figure 5.6 shows the comparison in newly signed loans by their type. Since the bank, which was merged into our bank, specialized more on mortgages, the number of mortgages signed before the observed period is higher compared to the number of loans. We can also see that while the number of newly signed mortgages stayed nearly the same, the number of new general loans increased every year.

The bank also provides overdrafts – small loans, with which a client is enabled to have the negative value at disposal on the current account. The client then

has to set the parameters, like the maximum amount or the term, in which client has to align the account with the bank before, and use it whenever needed.

Apart from the retail products aimed for the individuals, there are also commercial products for companies and self-employed. The number of these products corresponds to the number of clients of these types. The most popular commercial loan product is certainly a loan with the closed end, although the number of these loans decreased over time, as shown in Figure 5.6. For companies the bank also offers revolving credit, which, in contrast to closed-end loan, does not have scheduled payments and it can be repaid any time during the agreed period. However, the number of newly signed loans of this type is low, on average only 10 per a year.

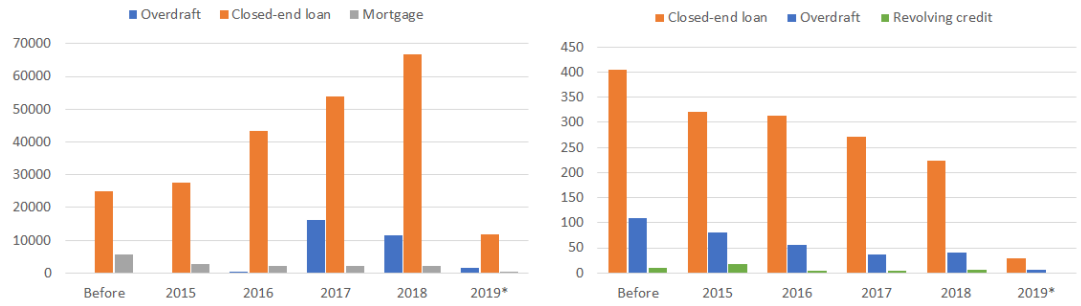

Figure 5.6: Comparison of new retail and commercial loans signed before and during the observed period

Type	Average	Maximum	Number of terminated
Commercial loans			
Closed-end loan	19,852,065	270,000,000	47 (3%)
Overdraft	4,437,330	210,000,000	–
Revolving credit	29,669,937	150,000,000	3 (6%)
Retail loans			
Closed-end loan	119,218	715,000	4386 (2%)
Overdraft	15,415	100,000	–
Mortgage	1,828,191	117,010,436	8 (0.0005%)

Table 5.1: Comparison of retail and commercial loan products according to their average and maximum amount (in CZK) and the number of terminated loans caused by client's disability to pay

Table 5.1 shows the comparison of amounts of lent money for different products as well as the number of products, which were terminated by client's inability to pay. We can see that even though the bank provided less commercial loans, the amounts of money are significantly higher than those of retail products. It also appears that commercial products are more often terminated, but because of the small total number of them, it cannot be confirmed. As we do not have more relevant data about our commercial clients, e.g., the area of expertise, we will not conduct thorough analyses of these terminations.

On the other hand, we can explore reasons behind the termination of retail loans. Although the amounts are lower, with high number of terminated loans

there should be some patterns of behaviour leading to the termination. Mortgages are not terminated so often, if there are some patterns, we should be careful, as they have probably a low support.

6. Spending Analysis

In this chapter we analyse expenses of clients. The data contain information about withdrawals from ATM, shopping in stores and e-shops. We will analyse both separated spending networks (only one type of spending) and combined networks with all kinds of spending.

As the analysis of different spending types is analogous, we will use a simplified term "pay on terminal" for all three types.

6.1 Analysed Networks

We have three types of networks for every terminal type, each with its unique properties and each revealing different features about the original real-life network:

- monthly cumulative network – a two-mode bipartite multigraph between clients and terminals with addition of the temporal dimension; we have 50 networks for 50 observed months, in each of them an edge represents that a client conducts transactions on a terminal, with information about the number of transactions and the total amount of money spent;
- total cumulative network – a two-mode bipartite graph between clients and terminals without the temporal dimension; for a particular month it contains cumulative statistics between a client and a terminal including information about the total number of transactions and the total amount of money spent over all past months;
- projected similarity network – a single-mode graph containing structural similarity of either clients according to the total cumulative relationships with terminals, or terminals according to their usage by clients.

Structural similarity of clients can be captured by the Jaccard metric introduced in Section 3.3. If we use the unweighted version of the algorithm, we only get similarities based on binary relations "a client has ever paid on a terminal / has never paid on a terminal".

However, using Ružička index and additional weights to relations (e.g., total amount, average per transaction, average per month) we can get even more accurate similarities, based on relations "a client has paid a certain amount of money on a terminal". An example of such projection is shown in Figure 6.1. Clients A and B, both withdrawing the same amount of money from ATM 1, have the maximum similarity, while client C, withdrawing a different amount of money from ATM 1 and also from ATM 2, has a lower similarity with the other clients.

6.2 Two-Mode Networks

Two-mode networks are bipartite networks, where disjoint sets are composed of nodes of one type. The first set, clients, are in the role of actors and they

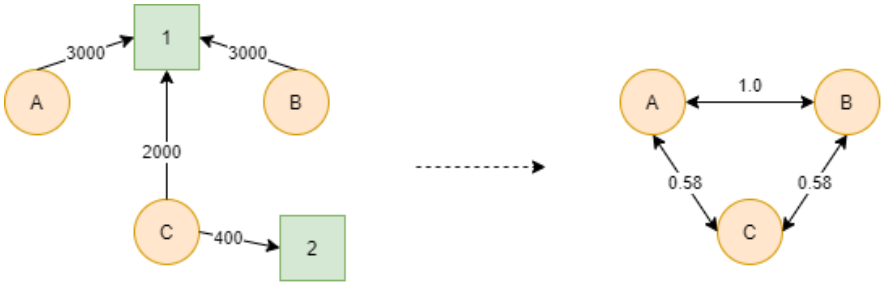

Figure 6.1: Example of converting relationships from rule-based to similarity-based using the Jaccard metric

actively create new relationships by conducting transactions on terminals, while the second set, terminals, are in the role of objects and do not interact with the structure of the network.

The analysis of multi-mode networks is more difficult, because most methods and approaches are designed for single-mode networks. We have to either perform algorithms on projected single-mode networks and backtrack results to the original two-mode world, or use advanced approaches for two-mode or bipartite networks. However, many researchers argued that a projection skews some of network properties, such as the clustering coefficient, and proposed new approaches for bipartite two-mode networks ([Opsahl, 2013], [Filho and O'Neale, 2020]). For many real-life networks it seems to be better observable in the two-mode nature.

6.2.1 Related Works

In contrast to the famous one-mode scientist collaboration network, where scientists are connected if they ever collaborated with others, also a two-mode network can be considered, where scientists are actors and scientific papers are objects. Similarly, for actor collaboration network, where in one-mode there is a relationship between movie actors, who ever collaborated together, there is a two-mode network, where movies act as objects, to which movie actors gain relationships. Evolution in such networks was described by a model proposed by [Ramasco et al., 2004]. This model, however, predicts the preferential attachment on the side of an actor (actors are preferentially picked to collaborate on a new object based on their older relationships with objects). Furthermore, objects, once they emerge, do not create new bonds with other actors, which is given by the nature of the observed data. Another model for bipartite online networks was proposed by [Zhang et al., 2013], which also allows preferential attachment on the object side and also relationships between two older nodes. This model also exhibits growth, as new objects and actors appear during evolution.

6.2.2 Our Network

In our network, the degree distribution of objects reveals more about network evolution than the degree distribution of actors, because actors actively choose which store they will visit or in which e-shop they want to shop online. We are interested in what key leads to the preference of one shop over others.

According to Figure 6.2, the degree distribution of stores has a linear shape in

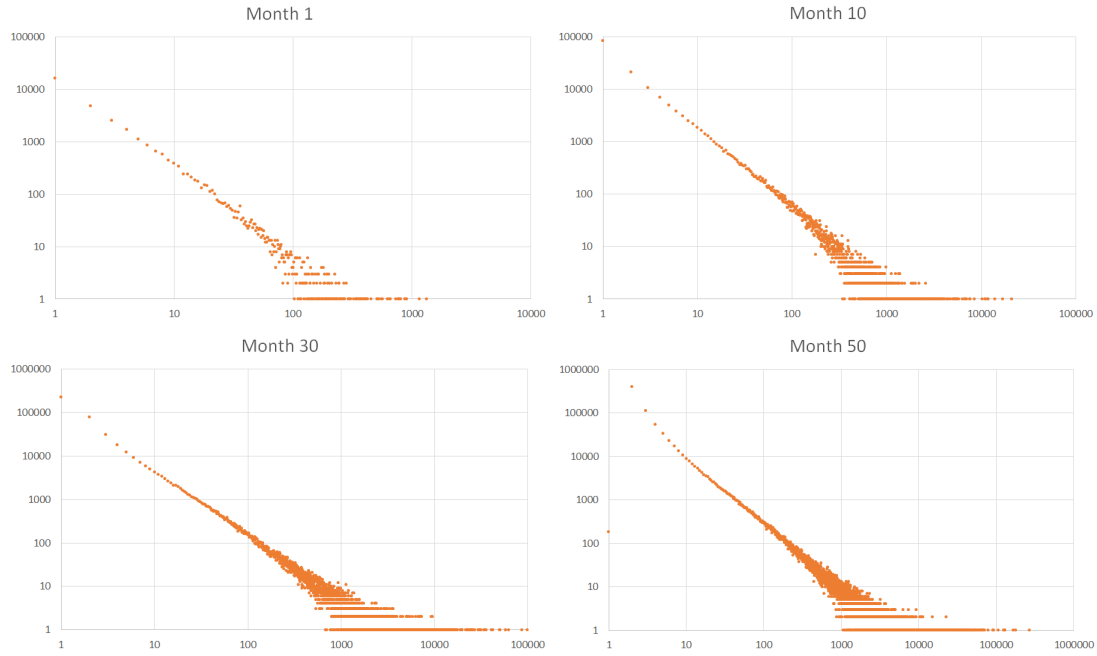

Figure 6.2: Distribution of degrees of stores in total cumulative networks for 4 example months (X-axis shows the degree, Y-axis the number of stores, which have it)

the log-log plot, which indicates possible presence of the power-law distribution[1], a necessary condition for a network to be scale-free. Indeed, we observe the presence of *long-tail* – large hubs which have a high degree, but there are only a few of them – and the presence of many low-degree nodes (stores with only a few customers in the whole time interval), which create majority.

This distribution is often present in networks, which evolved using the *preferential attachment* [Barabási and Albert, 1999]. According to this theory we could say that clients picked stores according to their popularity throughout time. On the other hand, this does not seem as the definitive evolution rule for ATM and store networks, because clients tend to be limited by their location and their preferential attachment is limited to stores and ATMs in their city.

The power-law nature of the data must be confirmed by finding a parametrized power-law distribution, of which data might have been created. Such parameters can be found by the maximum likelihood estimation, which is able to determine parameters for various distributions. The powerlaw library can perform this estimation and reveal the models, which are the most probable.

Stores	truncated power-law	power-law	lognormal
truncated power-law	–	0.69	0.34
power-law	−0.69	–	−0.34
lognormal	−0.34	0.34	–

Table 6.1: Comparison of the likelihood ratio between most probable distributions of degrees in the Store subnetwork

[1]https://necsi.edu/power-law

E-shops	truncated power-law	power-law	lognormal
truncated power-law	–	17.42	3.82
power-law	−17.42	–	−13.60
lognormal	−3.82	13.60	–

Table 6.2: Comparison of the likelihood ratio between most probable distributions of degrees in the E-shop subnetwork

ATMs	truncated power-law	power-law	lognormal
truncated power-law	–	6.82	0.53
power-law	−6.82	–	−6.28
lognormal	−0.53	6.28	–

Table 6.3: Comparison of the likelihood ratio between most probable distributions of degrees in the ATM subnetwork

In Tables 6.1, 6.2 and 6.3 we can see the comparison of three most probable distributions for all examined networks – power-law, truncated power-law (truncated in the tail part) and log-normal (distribution, whose logarithm follows normal distribution). While power-law can be characterized as "rich get richer", for log-normal it holds that "rich get richer, but with upper bounds" and it appears often in natural processes.

The numbers in the cells of these tables are the loglikelihood ratios between two compared distributions – this ratio is positive if the distribution in the row has higher probability and negative for the higher probability of column distribution.

As we can see, for ATM and Store networks the truncated power-law and the lognormal distributions are both good fits and we cannot claim any of them as significantly better. For the e-shop network the truncated power-law appears to be the best fit.

There are intuitive differences between the degree distributions of these networks. Relationships of clients with ATMs and stores are influenced also by geographical positions on both sides, hence there may be some localised preferential attachment. Smaller hubs, which are closer to the residence of a client has bigger probability to be picked than bigger hubs further away. This geographical consideration can lead to the upper bound, which are characteristic for the lognormal distribution. On the other hand, in interactions with e-shops these geographical considerations have smaller influence and e-shops are more likely picked by the preferential attachment.

6.3 Single-mode Networks

For a determination of many properties in the networks it is easier to use their projected single-mode variations. They have lower complexity, as the number of nodes and also their types is decreased, and algorithms, which would be too slow on two-mode networks can be more efficient on their projections.

For each type we will create a similarity network with the weights of relationships equal to the Ružička index of relationships of the cumulative two-mode networks. The meaning of such relationships is that when two clients conducted

transactions on the same terminal, the weight of the relationship determines how many terminals do they have in common and how much money they spent on them.

The advantage of this projection is that if both clients visited the same terminal, then the similarity increases, and if only one of them visited it, the similarity is decreased. Also, the fact that there is some terminal, which was not visited by any of the clients, does not impact the similarity at all. The main disadvantage is, however, ignorance of time, because two clients may be extremely similar even if the difference between transactions of the first client and transactions of the second one is several years.

However, an improved customized projections can be created and the structure of the single-mode network will stay the same. For instance, we can decide to consider also the time of transaction and mark two clients similar only if they visited the same terminals in the same months.

In our projections we ignore insignificant relationships with similarity lower than 0.1 (10%). Such pairs of clients have at most 1 out of 10 terminals in common and their spending habits are probably entirely different. If we include also clients with no connection to an ATM, store or e-shop, our single-mode network contains a lot of lonely nodes. These nodes can be in some analyses ignored, e.g. in community analysis.

6.4 Community Structure

We can separate clients into communities by the similarity obtained during the single-mode projection. An example of such separation is shown in Figure 6.3 – more similar clients A and B are placed in one common grey community, node C is placed in its own community. We expect that the community structure will be more observable in ATM and store networks, where clients are biased by location and tend to pick closer objects, while in the e-shop network the community structure should more consider the type of an e-shop.

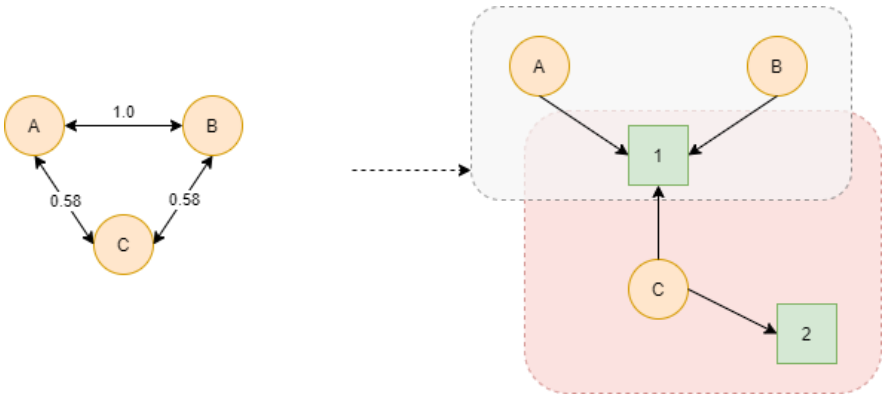

Figure 6.3: Example division of clients into communities and results after backtracking to two-mode network

We choose the Louvain modularity optimizing algorithm, as it not only yields a possible community structure, but also modularity, so the goodness of the structure can be measured. As all networks contain tens of thousands of nodes

and relationships, Girvan-Newman algorithm would be inefficient, as it looks for an optimal community structure, however, we expect Louvain to yield near optimal communities.

The basic comparison of community structure among all three networks is presented in Table 6.4. The modularity values imply good community structure in all three networks. Details are discussed in the following sections with the subnetwork analyses.

	Clients in network	**Modularity**	**Total communities**
e-shop	155,704	0.71	143
store	80,682	0.91	137
ATM	257,326	0.89	199

Table 6.4: Basic properties of network community structures

In the following sections we will focus more on the e-shop and store subnetwork. The ATM network is very similar to the store network because of the regional patterns and its analysis is therefore analogous. On the other hand, it lacks the further information about the purpose of the money being in transaction, therefore we will not analyse it in greater level of detail.

6.5 E-shop Network

E-shops differ from the ATMs and stores in a way that a client does not have to travel anywhere to spend money, therefore greater distance of an e-shop's address (which is usually the address of the headquarters) should not have any impact. On the other hand, there may be e-shops with poorer shipment support, whose majority of customers live or work nearby.

Even though we discovered over a million of e-shops from our transactional data, the vast majority of them appeared only in a few transactions throughout the whole observed period. As we can see in the comparison in Table 6.5, e-shops marked as active traded more than five times bigger amount of money in more than four times bigger number of transactions, even though they make up only 3.4% of all e-shops. Therefore we will focus on the active part.

	Number of e-shops	**Total money spent (CZK, in mil.)**	**Number of transactions**
Active	43,443	6,651.8	5,700,535
Inactive	1,221,209	1,300.8	1,322,046
Total	1,264,652	7,952.6	7,022,581

Table 6.5: E-shop spending - comparison of active and inactive e-shops

Similarly to the store network, e-shops changed their online terminals during the observed period. The same e-shop then appears multiple times with a similar name. In the overall analyses we merge them together.

As we can see in the comparison of key e-shops in Figure 6.4, most of the top e-shops are domestic with the few exceptions, e.g., foreign e-shop **AliExpress.com**

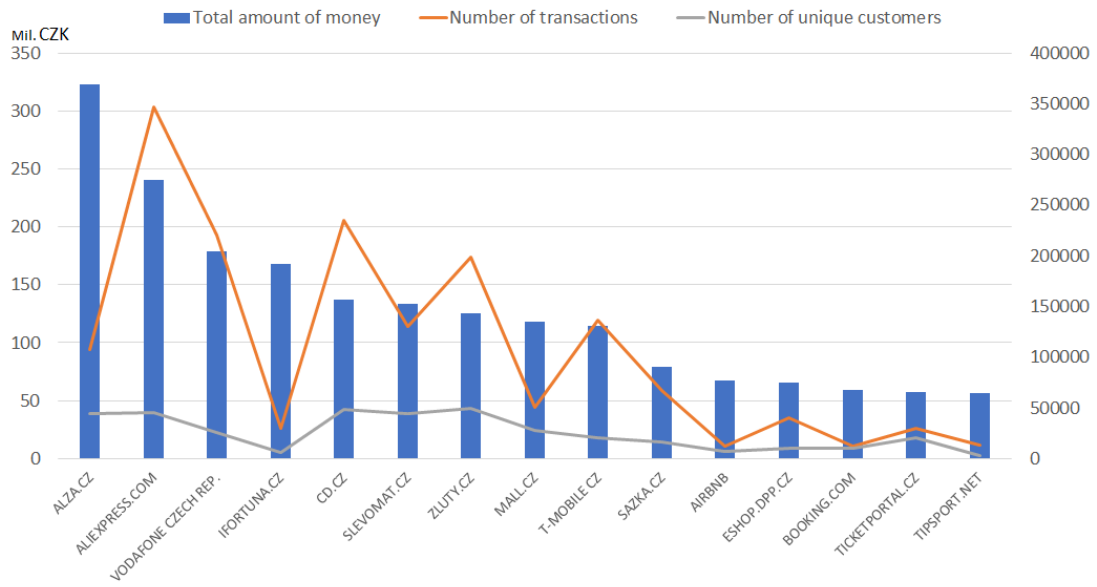

Figure 6.4: Top 15 e-shops according to their total amount of traded money (left axis) along with numbers of transactions and customers (right axis)

or popular accommodation platforms **AirBnb** and **Booking.com**. The orange line adds information about the total number of transactions, through which we can determine also the average amount of money spent.

We can see that clients of our bank spent the most of money in the domestic e-shop **Alza.cz**, also, large domestic e-shops as **Alza.cz** or **Mall.cz** have higher income from one transaction than foreign **AliExpress.com**, which focuses on more items sold by a low price.

Another interesting observation is that 3 domestic betting companies (**iFortuna.cz**, **Sazka.cz**, **Tipsport.net**) appear in this comparison higher than e-shops offering widely used services, like mobile phone operators or transportation services.

The regional view can reveal new patterns or anomalies. We selected 4 e-shops and displayed them on the map of the Czech Republic in Figure 6.5 – if the district is dark, the amount of money spent normalized by the number of clients in the district is high, if the district is white, there was no transaction in the e-shop from a client from the particular district conducted. For instance, in the upper left map we can see spending in the e-shop **Alza.cz**. Because of its good transportation services, there are transactions in every district, even in the rural regions, although the most transactions are conducted in Prague and its neighbourhood districts, possibly because of the spread of stores for delivery pick-up throughout the city.

On the other hand, in the upper right map we can see a locally bounded e-shop, **DameJidlo.cz**. The vast majority of transactions were conducted in two largest cities, Prague and Brno, because of the e-shop activity – it did not offer services in the smaller cities or in the countryside. As of today, this e-shop expands also to other cities. However, the network contains many locally bounded e-shops, e.g., other e-shops limited in the largest cities or local infrastructure services. Prague public transportation, shown in the lower left map, is the example of such

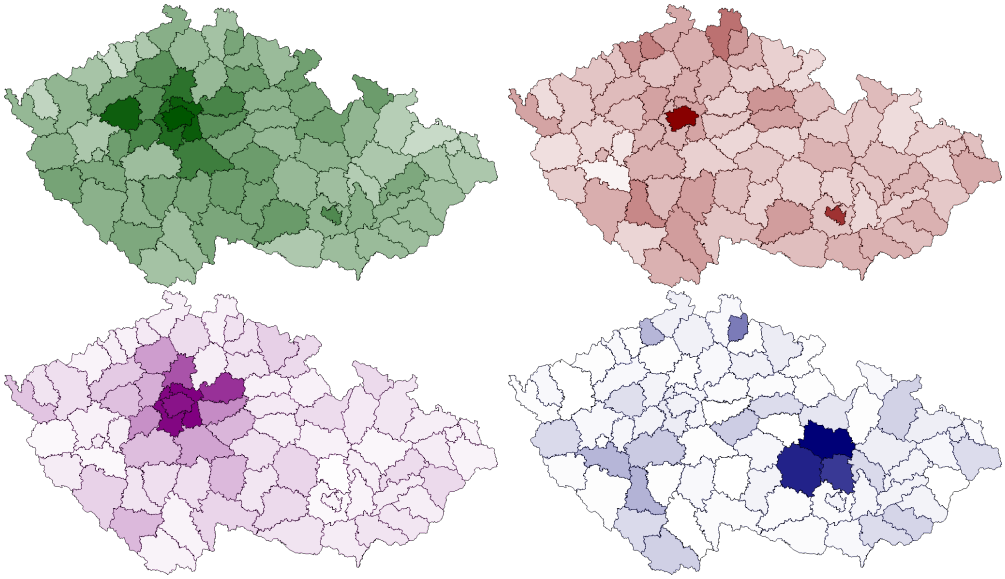

Figure 6.5: Total amounts of money spent by clients with residence in particular districts normalized by the number of clients in those districts in 4 e-shops – **Alza.cz** (green), **DameJidlo.cz** (brown), **DPP.CZ** (purple), **B365** (blue)

regional e-shop. Spending in these e-shops can contribute to determining the client's movement even without the knowledge of store transactions.

Other rather interesting e-shops are betting companies. They differ from another e-shops by receiving a large amount of money in a few transactions from a small group of clients.

Even though from the lower right map for the foreign betting company **B365** it appears that the three districts in the northern Moravia suffer from a betting addiction and that this e-shop is an example of regional e-shop, a detailed analysis proved otherwise. These districts are a residence districts for a small group of clients, who conducted hundreds of transactions for over 1.5 million CZK each. Due to a small size of these districts, their overall value appears to be high. Identifying clients, who spend large amounts of money at the betting sites, however, may be important from the point of offering a loan or a mortgage.

Community Structures

The community structure is quite different in comparison to the other networks. There is one dominant community containing one third of all clients of the e-shop network and it is centered around largest e-shops in Czechia. The majority of these clients are born after 1985 and have the residence in Prague or its neighbouring districts. They can be characterized as clients, who like to use online shopping for almost every aspect of life – over 44% of them used at least 5 e-shops from top 15 e-shops introduced in Figure 6.4. These clients also form over 75% of customers of the biggest e-shop **Alza.cz**.

Other communities are significantly smaller. The second largest community is entirely different from the dominant one – **Alza.cz** is the 19th most popular e-shop here, the most popular e-shops are those with various kinds of deals, like **Slevomat.cz** (selling vouchers for food, wellness, etc.), **AliExpress.com**

and **Wish.com**. The price of an e-shop item is certainly important for these clients. Smaller communities are all centered around the most popular e-shops, which can be used to describe the community simply, although such description should be confirmed or refused by a further analysis. For instance, the next two communities are centered around two largest mobile phone operators **Vodafone** and **T-Mobile** – they are receivers of the most of the money spent by these clients. These clients spent significantly less money in other e-shops as other goods can be bought in stores, however, mobile phone payments are nowadays almost required to be paid by credit card online. Another such community spent a large amount of money on the two biggest train travelling companies **ČD.cz** and **Zluty.cz**, less money were then spent in other e-shops from the the most popular 15, we could say that these clients travel by train more than others, e.g., for work or school.

The biggest communities are not often also most interesting ones. We expect that the behaviour of majority of clients is "normal" – they receive salary, shop in the mainstream e-shops and stores and do not cause any problems for the bank. These clients are not interesting for the bank's analysis. However, there are also client, whose behaviour differs, e.g., by higher probability to take o loan, or worse, to not be able to repay it.

We can directly search for communities, clients of which behave in some way. For instance, if we search for communities with the highest ratio of loans per one client, all communities on the top of this list are centered around betting companies like **iFortuna.cz**, **Tipsport.net** or **Sazka.cz**.

Comm. ID	**Loans and members**	**Top e-shops**
60213	1,330 loans per 1,092 clients	iFortuna.cz, Sazka.cz, AliExpress.com
24616	87 loans per 77 clients	SynotTip.cz, BWin.com, Wish.com
39229	1,492 loans per 1,607 clients	TipSport.net, iFortuna.cz, Chance.cz
19189	3,616 loans per 4,164 clients	Sazka.cz, AliExpress.com, GooglePlay
25334	3,454 loans per 4,129 clients	TMobile.cz, AliExpress.cz, Alza.cz

Table 6.6: Overview of communities with the highest ratios between the number of loans and the number of members

Even though the fact, that a clients applied for a loan does not necessarily mean that he is more risky, there is apparently a connection between taking a loan and spending on betting sites. However, inability to repay leading to the termination of a loan is a serious issue and banks try to avoid it. In Table 6.7 we present the top communities from the point of the loan termination ratio. As we can see, the majority of them includes clients spending in similar shops like the clients from communities detected above (community 39229 appears in both tables).

Communities with higher overdraft ratio are similar to the "loan communities", their top stores are mostly related to betting or gambling. However,

Comm. ID	Loan termination rate	Top e-shops
122025	23 / 196 (11.7%)	iFortuna.cz, B365
4037	30 / 367 (8.1%)	iTunes.com, LeoVegas, PlayStationNetwork
80645	44 / 566 (7.7%)	T-Mobile.cz, AliExpress.com, Aukro.cz
39229	110 / 1,493 (7.3%)	TipSport.net, iFortuna.cz, Chance.cz

Table 6.7: Overview of communities with the highest loan termination ratios

communities with the higher mortgage ratio do not spend significant amounts on bettings sites, spending of these clients is very similar to the biggest communities and include e-shops like **Alza.cz**, **IKEA**, etc.

6.6 Store Network

Similarly to the analysis of the e-shop network we can explore the store network. Spending in this network is, as mentioned, influenced by addresses of a store and the movement of a client.

The network contains over 800 thousands stores, of which the most are not interesting for our analysis, as there are only a few transactions with a small amount conducted in them. Therefore we selected, like in the e-shop network analysis, stores with at least 200 transactions as active. The basic comparison of active and inactive stores is presented in Table 6.8. In addition, we can see that this network, in comparison with the e-shop network, contains 10 times more transactions with the total value 4 times bigger.

	Number of stores	Total money spent (CZK, in mil.)	Number of transactions
Active	33,308	25,430,3	57,799,620
Inactive	806,929	6,030,3	7,487,835
Total	840,237	31,460.6	65,287,455

Table 6.8: Store spending - comparison of active and inactive e-shops

As we can see in the comparison of best stores in Figure 6.6, the clients conducted transactions with the largest total amounts of money in the large **IKEA** furniture stores, followed by the large showroom of the most popular e-shop **Alza.cz**. On the other hand, the next most profitable stores are almost all stores from large supermarket chains all over the Czech Republic, especially **Globus**, **Tesco** and **Albert**. The number of unique customers is low in comparison to the e-shops, mainly because of the locality – a store is mostly visited by customers, who live or work nearby.

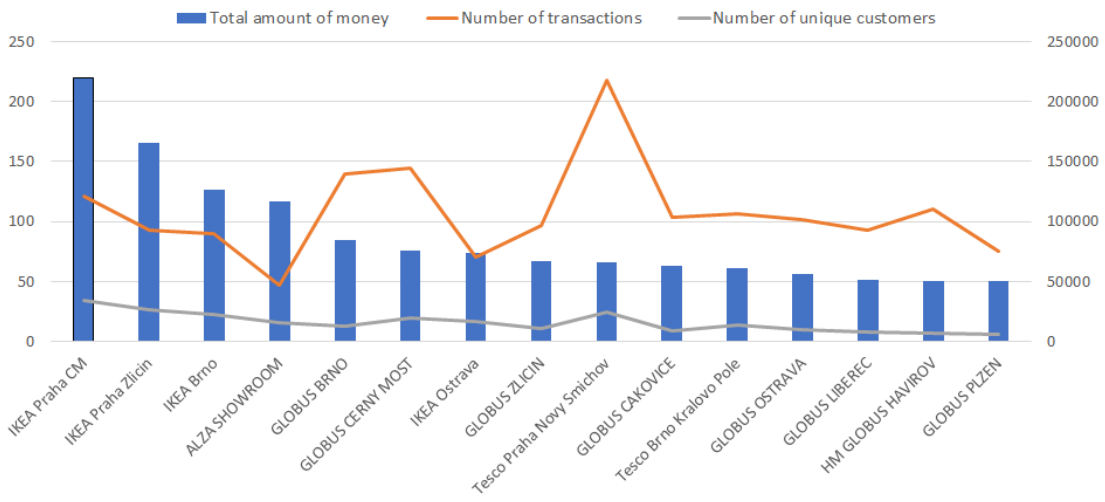

Figure 6.6: Top 15 stores according to their total amount of traded money (left axis) along with numbers of transactions and customers (right axis)

Community Structures

Community analysis implies a straightforward separation - clients are separated geographically according to locations of terminals, where transactions were made. Largest communities from the point of the number of clients are more balanced than in the previous analysis, there is no dominant community. The biggest community of clients interacts mainly with large stores in the second most populated city, Brno, especially furniture stores **IKEA** or supermarket chains. Other communities have similar interaction structure, but are located in other Czech cities, in addition, Prague is a common city to several major communities. Prague customers were also divided into regions inside the city - clients from the largest Prague community mainly visit malls and supermarkets in the west, while the second largest in the north-east area.

We can again, like in the previous section, search for communities with higher loan or overdraft ratio. However, due to the regionality of this subnetwork, we do not get "suspicious stores" (vast majority of them are supermarkets), but "suspicious regions". The examples of such communities are presented in Table 6.9.

Comm. ID	Loans and members	Regions
28801	302 loans per 298 members	Praha, Rumburk, Varnsdorf
9295	669 loans per 680 clients	Most, Litvinov
5802	594 loans per 606 clients	Karviná
27581	124 loans per 127 clients	Tachov, Marianske Lázně
14323	920 loans per 972 clients	Chomutov, Jirkov

Table 6.9: Overview of the store network communities with the highest ratios between the number of loans and the number of members

6.7 Spending After Certain Life Situations

Thanks to the complexity of our network we can analyse spending patterns as the consequences of another situation in life of a client. We can then propose and verify some theories, which might be useful for a bank to sell its products to those clients, who will be probably more interested.

For instance, are there some e-shops or stores, which are typical for spending after applying for a mortgage, e.g., furniture stores or hobby markets? If yes, can we estimate that a client got his/her application approved in some other bank? We can then, as a bank, offer refinancing after a short time.

We will try to answer these questions in Chapter 8 focused on evolution mining.

6.8 Association Rules

Knowing the overall structure of the network, but also the local microstructure, we can propose rules present in the networks and use them to discover causalities between pairs or even larger groups of shops.

Frequent Pattern Mining and Apriori Algorithm

A well-known approach *frequent pattern mining* [2], often used in machine learning (e.g., market basket analysis, recommendation systems), can be also used in our case.

The Apriori algorithm, proposed by Agrawal and Srikant [1994], is certainly the most commonly used one to mine the rules. The input of the algorithm is a set of transactions, each transaction is a set of items bought together. The task is to identify groups of items, which are most likely to be bought together and therefore the purchase of some of them raises the probability of purchasing others. Whenever we talk about transactions in this chapter, we mean transactions as defined in the frequent pattern itemset analysis.

For our use-case we could modify the idea of a transaction from a customer picking groceries in a store to a client choosing in which stores will he shop over some time period. An *item* is then one of the stores or e-shops and a *transaction* is a set of them.

As in other analyses, we can take one month as a period for transaction. Such transactions are showed on the left side of Figure 6.7 – client 1 shops in A, B and X in the first month, and in A, B, X and Y in the second month. The algorithm would suggest that if a client shops in B and X, he would shop in A with a high probability too, as this combination have a high support (2/3) and confidence (100%). Similarly, rule with Y as precondition would be mined with the same results.

However, as the presence of two shops in one transaction may be an evidence to a causality between them, a month is a long time to consider them as dependent. Intuitively, if the two purchases occur at the beginning and the end of month, respectively, the client may conduct them without a causality. On the other hand,

[2] Aggarwal, C. C. and Han, J., Frequent Pattern Mining, 2014, Springer International Publishing

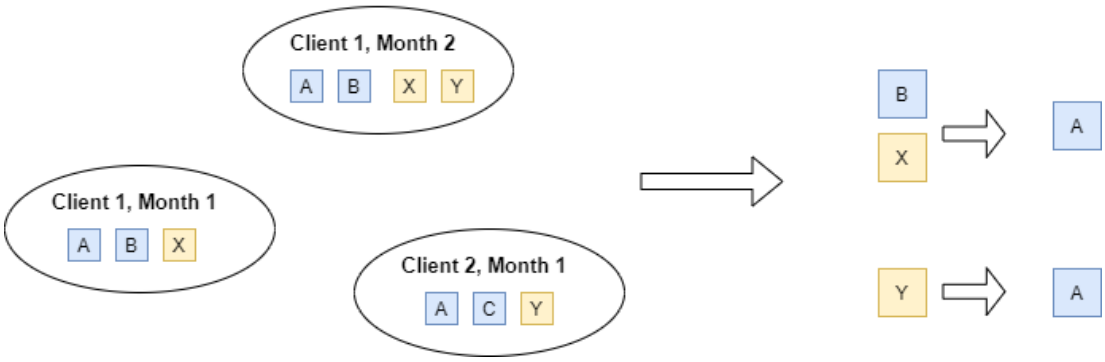

Figure 6.7: Shops used by a client in a particular month grouped into *transactions* (left) and rules mined by the Apriori algorithm (right)

if a small amount of time, as a day or a week is used as a period for transaction, some spendings may in fact imply others. For instance, if there is a grocery store purchase followed with a fast-food purchase in enough daily transactions, we can suggest a causality: if a client goes shopping, he usually also dines in this particular fast-food.

Results

Daily transactions may be the most interesting ones, although their analysis is the most complicated. The number of combinations *client + date*, even if we omit those, which contain only one item, reaches up to 15 million. The Apriori algorithm, however, efficiently eliminates stores and e-shops (and then combinations of them) with the support lower than our threshold. As shown before in Figure 6.2, the majority of stores have a small degree for our dataset with millions of transactions. Even though such percentage is low compared to the normal usage of the algorithm, we consider a pattern occurring in around 1000 transactions to be significant.

Although we do not expect any causality between spending in a store and spending in an e-shop, we still analyse them together. The causality between two e-shops should be also less probable, as a client is not motivated to shop in two e-shops at one moment like when he shops in stores, but some connections may appear.

The results of the Apriori algorithm on the stores network are presented in Table 6.10. As we can see, the most supported rules are those which imply a connection with a large supermarket and a small specialized store placed near the exit of this supermarket. Such stores are mostly pharmacies, drugstores or newsagents.

Another interesting pattern involves the furniture stores **IKEA** and their restaurant services. We can see that over 50% of clients spend money in furniture stores after spending in restaurant.

The most of inferred rules have a high confidence only for direction $A \rightarrow B$, where store B is a popular store with large individual support and A is a less popular store in the neighbourhood. However, there are a few, where the confidence is equally high for both directions. As an example there is a rule between two highway toll booths in Poland near the Czech borders, which are

used together almost every time.

Some of the rules emerged also among the e-shops, they are shown in Table 6.11. However, most of them are irrelevant for practical usage. The majority of them are correlations between two terminals of one e-shop. Thanks to the decentralized business model of **AliExpress.com**, where goods are not bought from one supplier, but from a lot of subcontractors operating in different countries, payments on these terminals form the most of e-shop rules. The only pair with the support slightly below the threshold, but occuring between two independent e-shops, are vignette e-shops for highways in Slovakia and Hungary, as these countries are both on the way to a popular summer vacation destination, Croatia. Indeed, the vast majority of these vignettes were purchased in the summer months.

In contrast to the analysis of daily transactions, we further explored weekly transactions, but also transactions composed from purchases in a shorter period. If we consider this period to be 3 days, the input transactions of the Apriori algorithm are all pairs of shops, for which there exist transactions within the span of 3 days. This transaction would cover the situation, when two shops are in causality, but the first approach did not discover it as the transactions are not always conducted in the same day. However, this approach did not bring any new results, only supported those already presented in the previous case.

Conclusion

In this chapter we have explored the parts of our network where the relationships are related to spending of money in various ways. First we introduced all considered subnetworks of both two-mode and single-mode nature. We also described the process of creating a single-mode network as a projection of the original two-mode one using structural similarity. We explored the degree distributions in all networks according to the type of the terminal and provided an explanation of the reason why they differ – locality. The e-shop and store networks were examined in the higher level of detail, we discovered and described the key players and the community structure. Finally, we used the association rule mining as a tool for the search of causalities between e-shops and stores.

Store A	Store B	Description	Support	Confidence	Reverse confidence
IKEA Restaurant, Praha 9	IKEA, Praha 9	restaurant → IKEA	13,164	61.2%	18.6 %
IKEA Restaurant, Praha 5	IKEA, Praha 5	restaurant → IKEA	9,706	56.3%	17.8%
IKEA Restaurant, Brno	IKEA, Brno	restaurant → IKEA	9,089	53.7%	17.1%
Globus PHM, Brno	Globus, Brno	gas station → supermarket	5,565	53.3%	8.4%
DM, Eden, Praha 10	Tesco, Eden, Praha 10	drug store → supermarket	4,246	55.2%	7.1%
GECO, Havířov	Globus, Havířov	newsagent → supermarket	3,600	80.4%	6.8%
Don Pealo, Mlada Boleslav	Tesco, Mlada Boleslav	newsagents → supermarket	3,296	73.2%	9.8%
Dr.Max, Brno	Globus, Brno	pharmacy → supermarket	3,039	77.7%	4.6%
BENU, Praha Zličín	Globus, Praha Zličín	pharmacy → supermarket	2,266	78.6%	4.7%
Stalexport Autostrada, Balice, POL	Stalexport Autostrada, Myslowice, POL	highway toll booths	2,182	90.9%	88.9%

Table 6.10: Frequent pattern discovered in daily store transaction data

Store A	Store B	Description	Support	Confidence	Reverse confidence
AliExpress.com, GBR	AliExpress.com, GBR	two terminals of one e-shop	24,144	53.2%	7.9%
AliExpress.com, LUX	AliExpress.com, GBR	two terminals of one e-shop	7,109	53.6%	2.3%
Google Play App, GBR	Google Play App, GBR	two terminals of one e-shop	3,292	70.7%	5.3%
Matrica, HUN	eznamka.sk, SVK	e-vignette e-shops	947	79.5%	10.7%

Table 6.11: Frequent patter discovered in daily e-shop transaction data

7. Income Analysis

This chapter is focused on the income of clients. With the knowledge of client's income, we can explore all other aspects of financial habits and raise new hypotheses.

The original data do not contain any reliable information about the type of an incoming transfer, certainly not every incoming transfer is a salary. Therefore, a supervised detection procedure must be performed. There are some intuitive properties, which a salary should follow:

- payment's additional data – the variable, constant and specific symbols of a payment, along with an optional message, can indicate its type
- amount and date – a salary for each client should follow patterns in terms of the amount and date of transaction, e.g., transactions on 15th of every month with the amount around 20,000 CZK
- counterparty – complementary to previous properties; if multiple transfers, which we consider probable to be a salary, came from one account, this account may be an account of some employer and all transactions can be confirmed as salaries

While this procedure marks transactions as being a salary, it also estimates some of the external counterparties to be an **employer**, which can be used as a new artificial entity. A company, which employs some of our clients and also has an account in our bank, is not represented by the employer entity, but as a client with type "legal entity". Naturally, one client can receive the salary from more than one employer, it is up to our analysis whether we sum them up or consider every salary separately.

However, the analysis of the variable symbol is not trivial. This property of a transfer, which is usually used for the identification of a payment for its receiver, is optional in real world. It often contains identification number of an individual or a company, who sent the payment, which are sensitive data, not available to our research. Hence, our data only contain a flag determining whether the transfer had an identification number as the variable symbol. This procedure showed that the vast majority of salary transfers have the value of variable symbol set to 9.

Not all detected salaries match the common definition of salary, some of them do not seem reliable, e.g., the amount outliers. For instance, the data contain approximately 4% of salaries with the amount lower than 500 CZK, although there can be an explanation to this, e.g., daily wage or wage split into multiple payments. Therefore, we will work with "joint" salary from one employer for a particular month. Those salaries which would be lower than 500 CZK, even when joint, will be ignored. As the concept of split payments can be a characteristic of an employer, we will note also the number of transaction for one joint salary. On the other side of the amount spectrum there are payments with large amounts, approximately 0.6% of all salaries have the amount higher than 100,000 CZK. These higher amounts can indicate either that a client has a job with a high salary and should be further observed, or that these transfers represent a salary for a longer term, e.g. three months or a year.

7.1 Employers

We will start with the analysis of the artificial employer node type. Basic properties are summed up in Table 7.1. Employers can be divided into 2 groups according to their total number of salary transactions. Approximately one third of all employers conducted more than 10 transfers, other two thirds conducted less. An employer with less than 10 salaries over the observed period is either an insignificant employer, which would not be relevant for further hypotheses, or a non-employer account, which was falsely marked as an employer – we will ignore these employers for simplification. If the bank however possessed the name of an account or would suspect it to be an employer for another reason, it could include it manually.

	# employers	# transactions	Transferred amount
Active	27,644	2,055,014	43,642,154,467 CZK
Inactive	50,817	103,610	2,695,803,934 CZK
All	78,461	2,158,624	46,337,958,402 CZK

Table 7.1: Basic overview of **employer** node type

Naturally, there are differences in the amount and the periodicity of salaries among clients and employers. Employers also differ in the number of unique clients, to which they sent a salary in the observed period. The relationship between the average salary and the number of employees is presented in Figure 7.1.

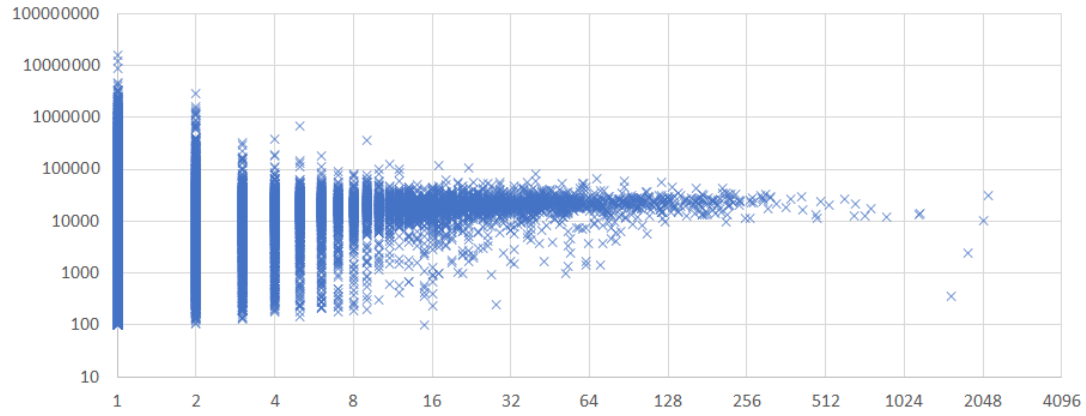

Figure 7.1: Overview of employers by the number of clients - employees (on X-axis, logarithmic scale) and the average salary paid in one month regardless of the number of transfers (on Y-axis, logarithmic scale)

We have to realize that the numbers of employees are not the actual numbers of all employees, as we can include only clients of our bank. The real size of an employer can be similar, if the majority of employees use our bank, or entirely different. We expect equal representation of all banks among employees and a linear dependence between our detected number of employees and the actual number of employees.

We can see that for the majority of employers the average salary lies between the minimum wage (ranging from 9,200 CZK to 13,350 CZK in the observed

period [1]) and the average wage in the Czech Republic.

7.2 Clients as Employees

On the other side of the salary relationship there are clients of our bank, however, not all of them. For the majority of our clients (295,389 – 72.6%) we have not been able to identify any income as salary. There may be various reasons:

- a client with income other than salary – an income of these clients can be unemployment allowance, maternity allowance or an income from a member of his/her family,
- a client with the main account in a different bank – these clients receive income to other accounts and use our bank only for specific purposes – savings, loan, etc.

We will for the purpose of this analysis ignore clients, who have no salary income. This leaves us with 111,132 clients, who received at least one salary.

7.2.1 Amount of Salary

The first and the most straightforward characteristics of a client's salary is certainly its amount. The overview displayed in Figure 7.2 shows the distribution of the average yearly salaries. We divided clients into "buckets" according to their average yearly salary sum because of the great variability in the salary periodicity (discussed more in Section 7.2.2). Although the largest bucket contains clients with the yearly salary between 128,000 and 256,000 CZK, which is slightly above 10,000 CZK per month, the majority of clients are below this amount. Some clients may indeed have a low salary, but we think that for most of them either the inference of their salary missed some of their income, or they have most of their salary sent to an account in a different bank and we see only a small part of it.

Demographics Behind Salary Amount

There are many assumptions about the amount of money that an employee can get for a job, e.g., inequality of men and women or regional inequalities (cities vs. countryside). We will try to analyse these assumptions in our data and see if our clients can be viewed as a good sample of a population. We are limited by the scope of data, which are available for our inferred network, as well as by the reliability of the data, which we already have – although gender is indisputable in most cases, some data, which change over time, may be obsolete (e.g., addresses).

To make this analysis more accurate and to be able to compare it with the general population, we will pick only salaries from 2018, because the number of clients with salary is highest in this year. We will again work with the yearly salary to avoid problems with irregular periodicity of payments and then take the median value – the average value may be biased by outliers.

[1] https://www.mpsv.cz/prehled-o-vyvoji-castek-minimalni-mzdy

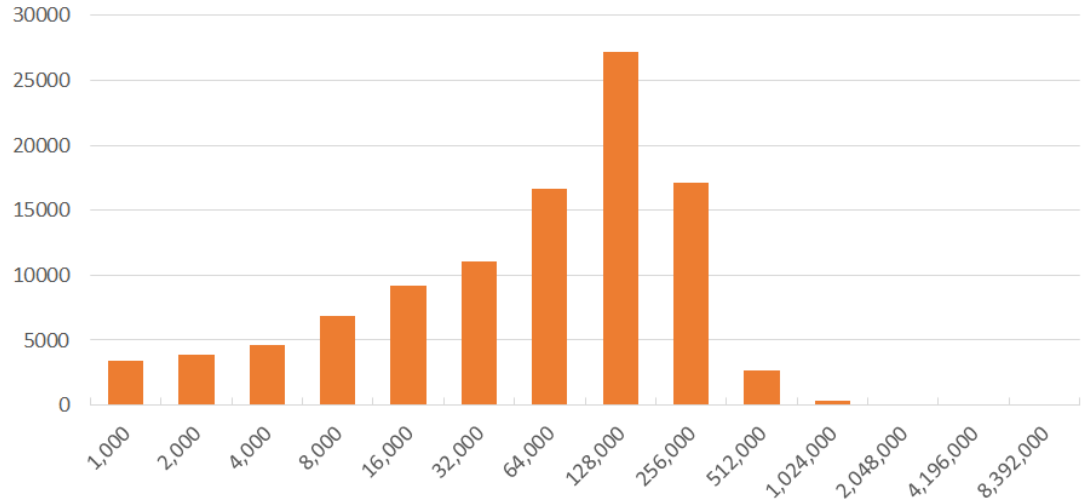

Figure 7.2: Numbers of clients by the average of their yearly salary

Assumption 1 Men receive more than women.

The analysis of dependence of salary and gender is the easiest, as the gender data are available for every client. For the year 2018 we have 39,984 women and 39,223 men, who received at least one payment inferred to be a salary. For every month and gender we took the median value. The population data were published by CZSO[2].

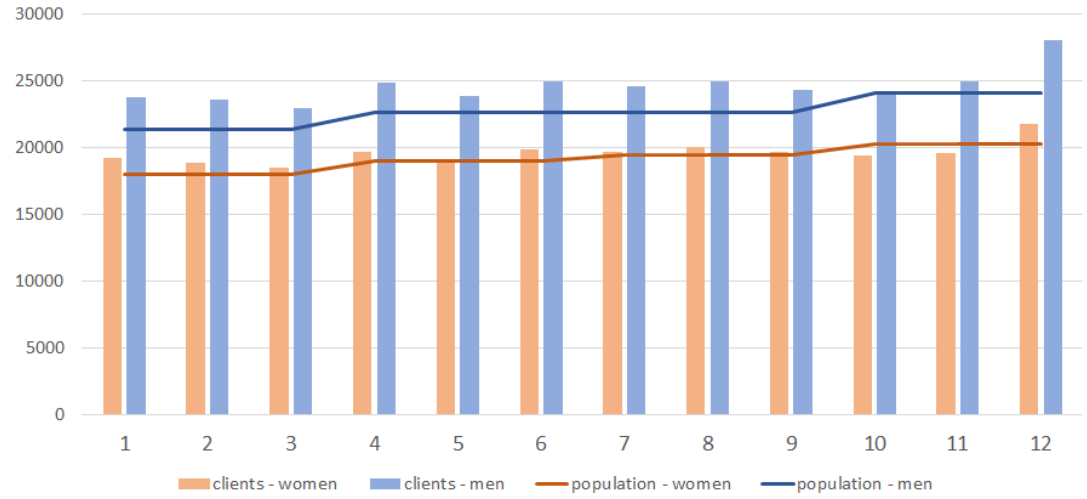

Figure 7.3: Comparison of median salaries (in CZK) of men and women, which are our clients (bars), with median salaries of men and women in population through the months of 2018

The results are presented in Figure 7.3. We can see that there are large differences between men and women throughout all months, as expected. We can also see that clients of our bank receive more than the general population, which can be caused by the orientation on municipal clients.

[2]https://www.czso.cz/documents/11350/60622098/gpmz030819.xlsx/a737d23e-4f79-448a-bfb3-d4401e9c5fa1?version=1.0

Assumption 2 Clients in Prague and other main cities have higher income.

The second analysis of salary amounts is focused on regional differences. Although most analyses comparing salaries on the geographical basis focus on the addresses of employers, we do not possess any data about them. We can work only with the addresses of employees. We suppose that the most of clients also work in the district of their residence, although this probably does not hold for the districts surrounding large cities, as their residents usually commute to this city.

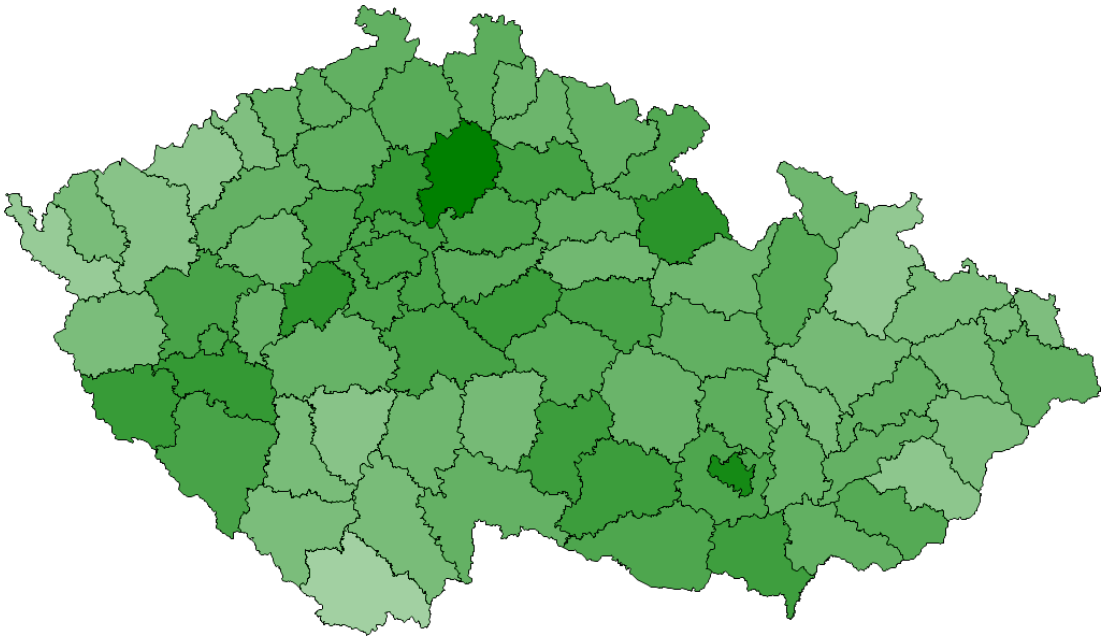

Figure 7.4: Comparison of the median yearly salary in districts of the Czech Republic in 2018

As we can see on the map in Figure 7.4, the districts are highly unbalanced. On the other hand, the districts with the highest salary are a bit surprising. In Section 5.1 we discovered that our bank is especially popular in the district of Mladá Boleslav. This region is the home region for the second largest employer in the Czech Republic, the car manufacturer **Škoda Auto a.s.**. Its high salaries combined with the popularity of our bank among its employees raise the median value (257,390 CZK) above the value of all other districts. On the other hand, there are many districts with particularly low salaries, with yearly summary being below 120,000 CZK. We can see such districts in West Bohemia (Cheb, Karlovy Vary), Southern Bohemia (Český Krumlov, Písek) and Silesia (Opava, Karviná).

Despite the expectations about the largest cities, there are certain countryside districts with even higher value than Prague or Brno. These two large cities have similar salaries, both around 190,000 CZK. Some large cities, as Olomouc or Ostrava are even below the majority of countryside districts. While it appears that this assumption can be rejected, we have to realize that our addresses of the residence of clients may not be reliable, as they may be the residents of their city of birth, but in fact they live elsewhere.

Assumption 3 Highly educated clients have higher income.

In the last amount analysis we compare clients according to the education, that they claimed to achieve. The education status is one of the personality traits which are not available for every client. Therefore, this analysis works with the least data. We will again work with the sum of all income of year 2018.

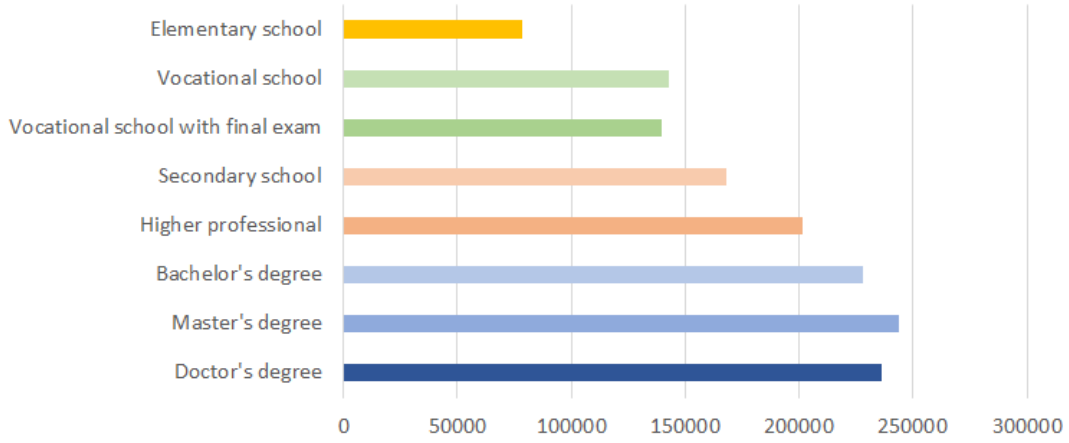

Figure 7.5: Comparison of different education groups by the yearly salary in 2018

Even though the number of considered clients is smaller (27,101), the results, presented in Figure 7.5, match the expectations the most. The highest salaries can be observed among clients with the university education, although the differences in the degree are insignificant. With the lower education the salary also decreases, clients with the elementary school education have median salary at one third of clients with the university education. There is one interesting discovery – the final exam apparently does not help to increase salary for the students of vocational schools.

7.2.2 Periodicity of Salary

The amount of salary is not the only important aspect of the salary analysis. A client, which receives a certain amount of money every three months probably behaves differently than a client with a regular income. We will consider only the range of months in which clients got their first and last salary. This constraint can help us to make the division more accurate, e.g., if there is a student, who did not receive any salary in his/her first months of the membership in the bank, but then started to receive a monthly salary, we will consider him as a client with monthly, not irregular, income. We will create 6 groups:

1. regular monthly income – a salary received every month,
2. almost regular monthly income – a salary received in 90% of included months,
3. every 3 months – a salary received at least every 3 months,
4. every 6 months – a salary received at least every 6 months,
5. annual – a salary received at least once a year,
6. irregular – a salary, which does not match any of the patterns above.

We created the second group because there is a significant number of clients, which have a nearly perfect series of salaries, but there are 1 or 2 months with

no income, so they cannot be included in the first group. We believe that these clients may have a special behaviour distinguishing them from other groups.

This division into groups is, however, simplified – if there are clients, who receive their salary in a different regular pattern, e.g., 4 months, we may not notice them and include them into the "annual" or "irregular group". A supervised analysis with a deeper knowledge of the salary domain can probably produce more accurate division.

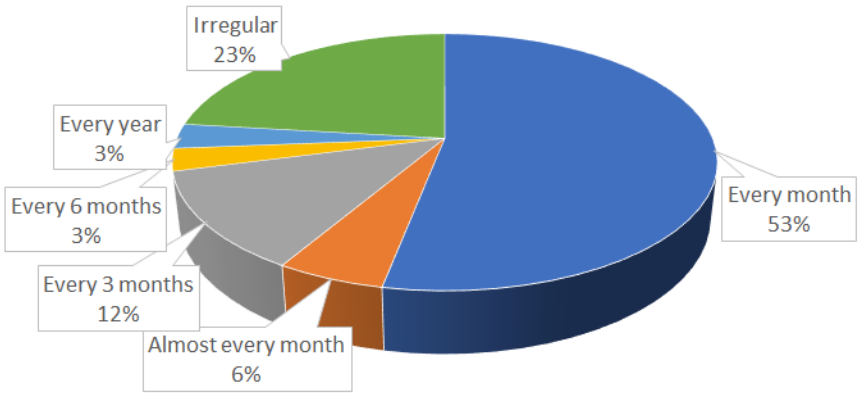

Figure 7.6: Overview of the periodicity of salary

Figure 7.6 shows the groups and their sizes. The majority of clients received a salary every month of their observed period. Together with the group of clients, who received their salary "almost every month" we have nearly 60% of clients. Approximately 18% of clients have a job with the salary periodicity different than one month and finally, approximately 23% do not have regular income from the viewpoint of our bank. This group is composed of clients, for which we recognized only a few incoming transfers as salaries, or those, who have large gaps in their salary series. This division allows us to propose and verify some ideas, which may be intuitive, but must be confirmed in the real data.

Assumption 4 Regular salary leads to a better risk class.

While the risk class is usually based on multiple factors, the regularity of the salary should be certainly the important one. We expect that clients with regular salaries have in general better rating of risk than those with irregular salary payments.

	More often than 3 months	**Less often than 3 months**	**Irregular**
A+	37.9%	37.2%	37.2%
A	16.0%	14.9%	16.3%
B	17.3%	16.9%	17.1%
C	9.7%	10.2%	10.0%
D	11.0%	12.2%	11.4%
E	8.1%	8.6%	8.0%

Table 7.2: Comparison of risk class distribution among clients from different salary groups

However, Table 7.2 shows that there is no difference in risk class membership based on the identified periodicity groups. The distribution of risk classes is in all groups almost identical to the distribution among all clients.

Assumption 5 The interest in certain bank products is influenced by the regularity of income.

Having clients with a salary divided into periodicity groups, we expect this division to confirm or deny straightforward hypotheses. We will focus specifically on particular bank products. For instance, do clients with regular income apply more for a mortgage? Do clients with a lower periodicity or those with irregular income use more overdraft service? We will examine dependence of a client having/not having one of products on client's income periodicity.

We will use Pearson's chi square test of independence. Through this test we will try to reject null hypothesis H_0: variables *membership in a periodicity group* and *possession of a certain product* are independent. There is no problem with the test reliability – all values are safely above the recommended threshold 5. As there are 6 groups, we have 5 degrees of freedom for every test. χ^2 value for this level of freedom and confidence level 99% is 15.09.

Results

	Observed	**Expected**	χ^2
Monthly	1,545	1,398	15.45
Almost Monthly	162	143	2.52
Every 3 months	263	327	12.52
Every 6 months	52	74	6.54
Every year	56	82	8.24
Irregular	557	609	4.44
Sum	**2,635**		49.71

Table 7.3: Chi-square test for mortgage data

In Table 7.3 we present results for dependence of mortgages. As we can see, total χ^2 value is 49.71, which is significantly above 15.09. We can thus reject H_0 and claim variables dependent. The biggest differences from the expected values are for clients with a monthly income, which applied for more mortgages, while clients, who receive their salary less often, applied less.

	Observed	**Expected**	χ^2
Monthly	13,318	13,873	23.12
Almost Monthly	1,646	1,402	42,46
Every 3 months	3,441	3,195	18.94
Every 6 months	733	725	0.08
Every year	745	805	4.47
Irregular	5,870	5,959	1.32
Sum	**25,753**		90.39

Table 7.4: Chi-square test for loan data

Table 7.4 shows results for loan owners and their distribution into groups. The total value of χ^2 is again high and we reject the null hypothesis. The most significant difference between reality and expectation is observed for clients with a nearly regular income – higher interest in loan can be caused by a sudden decrease of the money on account due to the loss of job. Clients with 3 months between salary payment also seem to apply for loans more, while clients with a longer gap do not. On the other hand, clients with regular income apply for loans less, probably because they did not experienced sudden financial problems.

	Observed	**Expected**	χ^2
Monthly	6,955	6,804	3.35
Almost Monthly	769	698	7.22
Every 3 months	1,651	1,591	2.26
Every 6 months	339	361	1.34
Every year	383	401	0.80
Irregular	2,726	2,967	19.57
Sum	**12,823**		34.54

Table 7.5: Chi-square test for overdraft data

For overdraft data the results are similar to the previous cases, the total χ^2 value is over threshold and H_0 can be rejected. However, as we can see in Table 7.5, the group with the biggest difference from the expected value is the one with irregular income. This group, however, is a group with the most unknown background. Some of them may indeed get an irregular income based on their seasonal job, but some of them may receive their income to an account in some other bank, as they considered our bank as their main bank for a while, but changed it over time. If we ignored this group, χ^2 value would be only a slightly over the expected value.

Another interesting revelation is that clients with regular income use overdrafts in a similar amount as clients with an almost regular income and more often than clients, who receive their salary only once or twice a year. On the other hand, the regularity of their income may assure them that they will be able to settle this overdraft and therefore they feel safer when spending money.

We explored correlations between the membership in one of our salary periodicity groups and possession of a loan-like product. Although independence was rejected in all cases, the rejection of independence of an overdraft product rely mostly on the group of clients with irregular income. This group would certainly require more examination and possibly be further divided by a supervised analysis. The dependence of mortgages and loans was, however, confirmed, we also discussed the most biased behaviour in relation with these products.

7.3 Structural Communities

We will try to discover possible community structures. As this part of the network offers great variability in the level of detail, we will go through multiple community divisions based on the different types of relationships. We will then for every community division search for a community, which is different from some point of view, e.g., demographics or higher interest in certain products.

7.3.1 Similarities of Salary Histograms

First, we will ignore employers. We will create a single-mode subnetwork with similarities of clients based on the difference of their monthly salary histograms. The histogram for a client contains 50 values for each observed month, the values are sums of all joint salaries from all employers. Through this network we should be able to group together those clients, whose value of the salary was similar at the same time.

Histograms and their similarities were inferred only for the preselected group of clients, which live in Prague. The advantage of such group is that the regional differences in the salary amounts are ignored. The average salary in Prague is significantly higher than the average salary in other regions of the republic. If then all clients are included, we consider similar an average client from Prague with a well-paid client from the countryside. On the other hand, a preselected client from one region can be biased in other ways compared to all clients or the general population. For instance, the number of clients with the university education is two times higher than among all clients (7.5% vs 4.5% for the bachelor's degree, 15.1% vs. 7.2% for the master's degree) and, at the same time, the number of clients with the elementary education is at 60% ratio compared to all clients (3.8% vs. 6%). A community with the high number of clients with the highest education would not be extraordinary, it would just copy the regional characteristics. The preselected group of clients were used for multiple descriptor and histogram analyses, the number of members of this group is 54,904. However, not all of these clients have their salary sent to the account in our bank, the number of clients with salary histograms is 18,892.

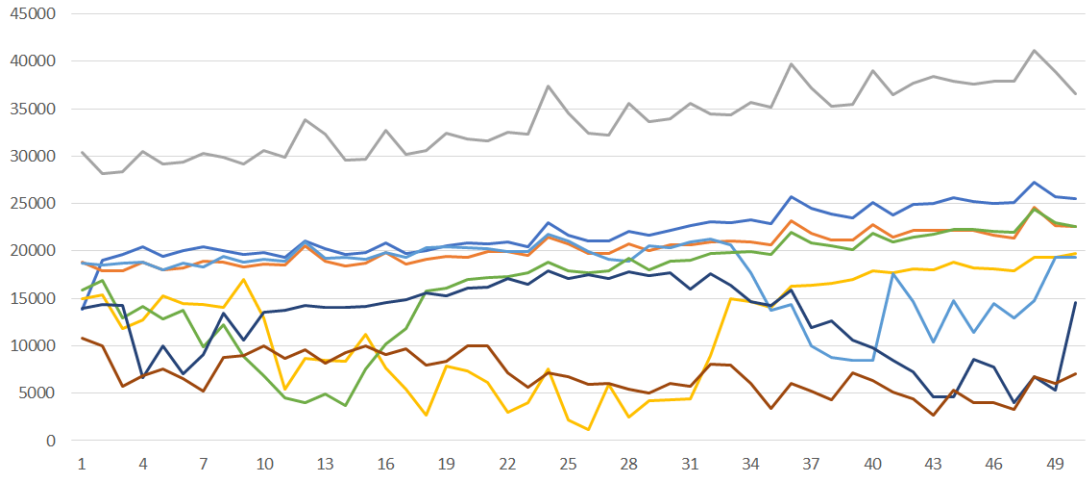

Figure 7.7: Comparison of 8 clients communities created from similarities of salary histograms

Figure 7.7 offers comparison of medians of salaries by community through the observed period. As we can see, 8 communities were created.

One of the communities, marked with the grey colour, is above all others in terms of the salary amount. The value for every month is above the average salary in Prague at that time. This community differs from the others by even higher ratio of clients with university education, also 70% of these clients are men. This group also contains a significantly higher number of clients between

the age 35 and 50, while at the same time only a few clients under 30. Clients of this community are less likely to take a loan, but more likely to take a mortgage.

On the other side, there is the brown community. Salaries of these clients are low, they do not exceed 10,000 CZK after the first month. This community is composed mainly from clients between 20 and 35 years old.

We can also see, that only 3 communities have perfectly stable income during all 50 months, there are only positive peaks in the end of every year (possibly Christmas bonuses). Others either started low and grew over time (green, yellow), started high and then decreased (light blue), or first gradually increased and then decreased (purple).

The disadvantage of the histograms and their comparison is that it does not have to provide the reliable comparison of two clients and their salary series. For instance if there are two clients, who receive the same amount of money every 3 months, but each of them in different months, the distances between histogram values would be large, even though the two clients have nearly identical salary.

7.3.2 Involvement of Employers

The second attempt to divide clients according to their salary will include also employers. The similarity relationship used for the community detection algorithm will be based on aggregated relationship of a client and an employer. Various aggregation metrics can be used – the total amount received, the total number of salaries, the average salary, etc., we will use the total amount of money.

Results

We used Louvain algorithm for the community detection, the structure is near-perfect (modularity over 0.9). However, communities are not very interesting, as they are mainly created by the employees of particular employers. Due to the insufficient information about employers, such as an area of expertise or an address, the employer entity itself is not very important in this salary analysis. If we, for instance, knew the economic segment of the employer, we could use it as a basis of new analyses (average salary in an economic segment, a client changing his/her segment, etc.).

With our limitations, as long as employers are comparable, they can be interchanged. Therefore, we will replace the employer entity as the target of a salary relationship with its corresponding artificial entity of categorical nature.

7.3.3 Replacement of Employers With Their Categories

New artificial entities will be created from the combination of an average salary and the size of an employer. If then two clients will be in a salary relationship with two different employers, however, both these employers will be members of the same clusters, the two clients will now be in a relationship with the same artificial entity. Similarity of clients will be based on the number of transactions between a client and an artificial employer as the amount of money is encoded in the employer itself.

Results

The Louvain algorithms yielded good community structure, the value of modularity is above 0.93. We have obtained 2,130 communities, majority of them (approximately 75%) have less than 30 members, only 11 have more than 1000 members. Large communities are mainly composed from employees, who for the most of their time in the bank worked at the same (or similar) employers. For instance, the largest community contains clients, who worked for the smallest employers with the lowest average salary in the whole observed period. On the other hand, there are also communities, for which there is no dominant category of employer. In these communities we can observe clients, who changed their employment in similar way from the point of view of our chosen categories.

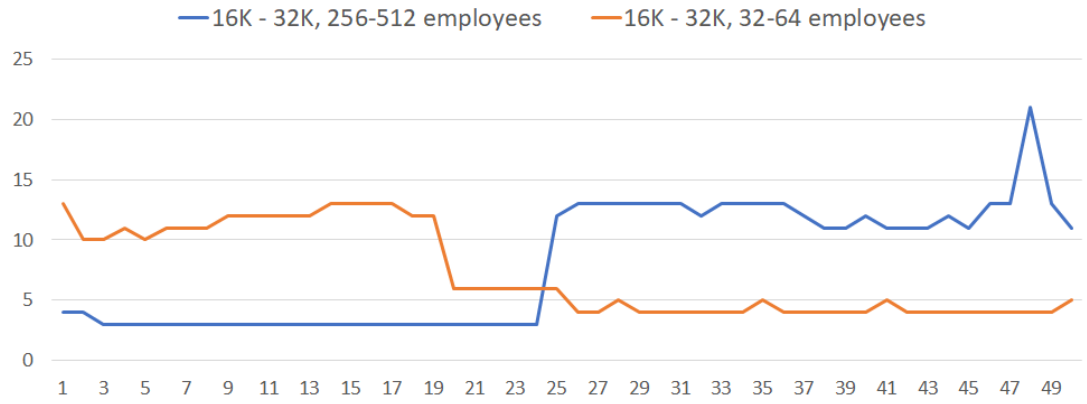

Figure 7.8: Example of community with similar change of employment over time (Y-axis shows the number of clients working for an employer of certain characteristics)

An example of such community is presented in Figure 7.8. Most of its members worked almost half of the observed period at a medium-sized employer, average salary of which is between 16,000 and 32,000 CZK. As we can see, the majority of them then changed their employer to an employer with the similar average salary, but the number of employees in range of 256 and 512. Even though these clients probably do not know each other (they have different year of birth, address, etc.), we have been able to capture their common behaviour.

7.4 Friends or Colleagues

The **friendship** relationship was created entirely on common transaction history between two clients. Some of them are possibly even more bound, one of the main binding relationships can be sharing of the common employer. This idea is motivated by the principle of the social influence – the change of behaviour in order to resemble the behaviour of one's friends. Even though two clients do not share the same employer, they can influence each other to get the similar jobs.

We will divide all friendships into categories from the salary point of view:

- colleagues – friends, who also receive their salary from the same employer for a few consecutive months;

- friends with similar jobs – friends, who have never worked for the same employer, but who have worked for two employers, which are somehow similar (e.g., average salary, size, periodicity);
- no work relationship – friends, which do not have similar jobs and never did.

We then expect, that clients, which are friends, but also are members of the same group, are more likely to have the same characteristics, than those with no work relationship. We will compare their risk class and social and education status. If this division into clusters has an impact on the risk class, we get an interesting tool for its prediction based only on relationship with another already known client.

Results

The disadvantage of the friendship relationship is its sparsity – there are only 37,743 relationships between 47,404 unique clients. If the bank would want to use this relationship for such analyses, it could extend the strict threshold on the relationship probability.

Colleagues	**Similar jobs**	**No confirmed relationship**
4,290 (11.4%)	5,066 (13.4%)	28,387 (75.2%)

Table 7.6: Overview of created friendship groups

As we can see in Table 7.6, only for 25% of friend pairs we were able to confirm the professional closeness, while for the rest there is no evidence. Even for those pairs, where in fact there is a relation, we do not have to be able to detect it because the clients use an account in some other bank as their salary account.

	Colleagues	**Similar jobs**	**No rel.**	**All friends**
Social status	89.4%	78.7%	69.2%	73.0%
Education status	42.8%	37.1%	37.0%	37.7%
Risk class	25.0%	24.5%	23.9%	24.1%

Table 7.7: Comparison of selected properties through all friendship groups – the percentage value indicates how many pairs from the group have the same value of a certain property

Table 7.7 shows the results of property comparison. The fact that two friends were colleagues have the biggest impact on the social status. On the other hand, two clients tend to have the same social status more if they are colleagues, the friendship relationship does not have an impact. As for the education status, we can see that there is a small difference between friends, who are also colleagues, and others. However, the group of friends, who worked together and for which we also know their education status, is too small, a further analysis with the extended friendship relationship would be needed. The most interesting property, risk class, is apparently independent of the work relationship of two clients.

The detection of colleagues among friends or friends with similar jobs and their comparison to those pairs of friends, who do not seem to have any work

relation, does not bring significant conclusions. However, this idea cannot be entirely rejected, as our inferred data do not provide entire picture of the real world. In the ideal network, where we possess salary, friendship and personality data about all clients, such analysis would be more reliable.

8. Graph Evolution

In Chapter 6 we used frequent pattern mining to mine rules and causalities from the network. Transactions were composed of shops, in which clients spent their money. However, our network contains a lot of different types of relationships between different entities. Encoding of a more complex part of the network would be difficult and the number of transactions and unique items would be too big to be efficiently handled by the algorithm. Some information, e.g., that a client has a friend, who also shops in the same e-shop, would not be coverable. However, there are also methods, which were developed directly for network-like data.

8.1 Graph Evolution Rule Mining

An example of such method is Graph Evolution Rule Miner (GERM) presented by Bringmann et al. [2010]. It is a method for extracting rules of evolution from a network, which can then be used also to predict future development. It requires a network with a temporal dimension, where an evolution can be observed. Our network, where the majority of relationships have their timestamp, is suitable. Although the basic algorithm does not assume a lot of different relationship types, it can be easily extended.

In GERM the network G is represented as series of snapshots $G_1, ...G_t$, where G_i stands for network in time i. The main idea of the algorithm is matching of *patterns*, i.e., small graphs with unbound variable nodes. Pattern P then occurs in a graph, if there exists assignment of the variable nodes in P to the real nodes in G.

Every relationship type can be used in a pattern – those relationships, which do not have a temporal dimension, would be simply a part of every snapshot. We can use also properties of nodes by transforming them into new nodes, e.g., gender or birth year.

A pattern then can be used like a search query in the network. In Chapter 6 we proposed some questions regarding spending of a client after a certain event in his life. We will use these questions as network queries. These queries and their matches can be used in various ways:

1. pattern occurrence – basic and the simplest usage; for a given pattern we find assignments of all its variables and count these assignments; there are 2 occurrences of pattern on the left in Figure 8.1 in the small network in the middle (one node can be a part of multiple occurrences);
2. pattern probability – we can estimate the probability of a certain edge of a pattern based on the number of occurrences of a full pattern and the number of occurrences of a pattern, which is missing the evaluated edge; there is 1 match of the full pattern in Figure 8.2 and 4 matches of the pattern without the evaluated edge, therefore the probability of this edge is estimated to be 25%;
3. prediction – patterns and their probability can be used for the future prediction – probability of a "future" edge in the current network should approximate the probability of an event represented by this edge.

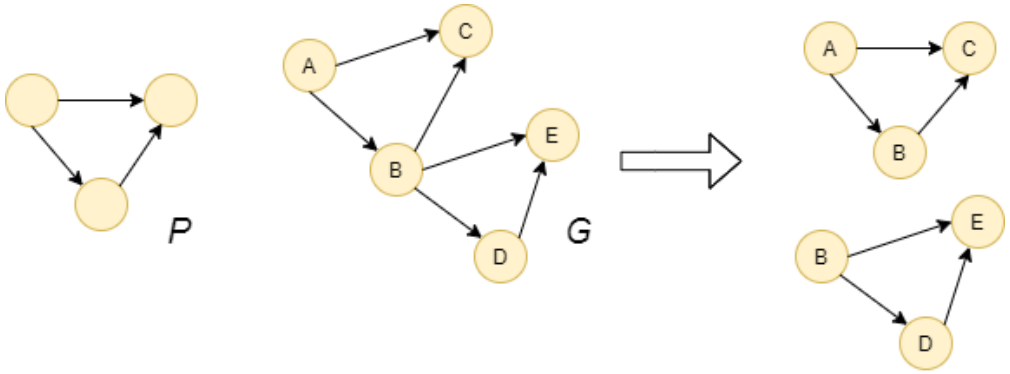

Figure 8.1: An example of a pattern P and its mapping in a graph G - nodes A, B and C form one occurrence and nodes B, D and E form another one

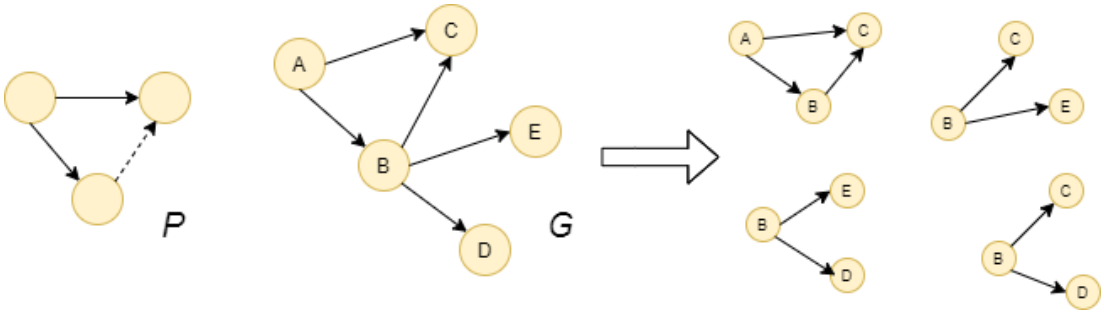

Figure 8.2: An example of the mapping of a pattern P and its subpattern with a missing edge – only 1 triplet matches the full pattern, 4 triplets match the subpattern

Our goal will be as follows: we will take a hypothesis, based on real-world questions, and encode it into the pattern. Then we will search for matches in our network a observe the occurrence of the patterns as well as the specific nodes, which were mapped.

8.2 Friendship Relationship

Relationship of two clients "being friends" was inferred after a complex analysis of their transactional behaviour. Can it bring new information about other common interests? Does this relationship mean that these clients influence each other? We believe that possible communication between two clients may cause visible patterns in their behaviour.

In SNA there is an idea of *social influence*, according to which an actor may alter his behaviour according to the behaviour of other actors to which he has a social connection, e.g., a relative or a friend. This influence may have an impact on direct properties of an actors, like the residence address of a client, or even on connections with a third actor or an object.

If the influence has an impact on a third actor, we observe a *triadic closure* and if there is an object, to which one of the actors is connected and the second one is not, we talk about *membership closure*. The examples for these closures are presented in Figure 8.3 – the membership closure between two friends and the loan product and the triadic closure between three friends.

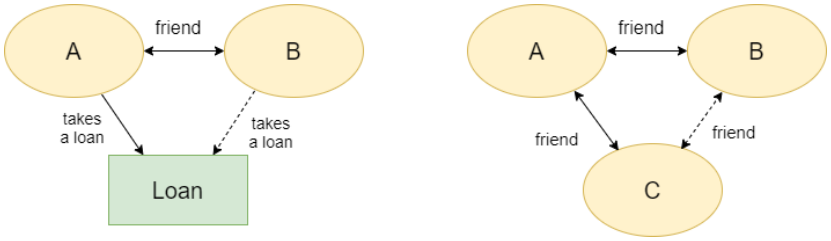

Figure 8.3: An example of the membership closure (left) and the triadic closure (right)

As an opposite to the social influence there is a *selection* – two actors create a relationship based on the mutual relationship with an object (e.g., people becoming friends after attending the same lecture). This results in the *focal closure*. It seems to be similar to the process of inferring our **friendship** relationship, where two clients are friends if they share the common stores and ATMs at the same time, but it is not. We do not know the exact time, where the two clients become friends. As it is more likely that two clients, who have a common transactional history, were friends before that history, than that they became friends after some common transactions, we assume that the real-world friendship was created before they even entered the bank. Therefore, we will not work with the focal and triadic closures and focus more on the membership closure, which is more useful in the banking world.

8.2.1 Membership Closures with E-shops

This analysis is focused on the influence of the friendship relationship on spending in shops and withdrawing from ATMs. However, since the **friendship** relationship was created from the common transactions in stores and ATMs, we will examine only e-shop transactions, which were not used for this inference.

Therefore, we will search for the matches of a pattern consisting of following rules:

1. a client has a friends (or multiple friends),
2. client's friend(s) shopped in some e-shop in some month,
3. the client did not shop in this e-shop in this month,
4. consequence relationship – the client shopped in the e-shop in the following month.

Table 8.1 shows the results of the pattern matching. In the **Support** column there is the number of incomplete patterns – those matching all but the consequence relationship. In the **Confidence** column there is the number of matches of the full pattern containing also the consequence relationship.

The simplest pattern – having a friend, which shopped in an e-shop in previous month – does not have a high confidence. The number of consequence relationships reaches only to 2%. However, with the higher number of friends the confidence of pattern also increases. For instance, if there are 3 clients, who shopped in an e-shop, their friend visited this e-shop next month with the 12.6% probability. However, the probability is still too low to be able to tell whether a client will shop in the particular e-shop in the next month.

# friends in the previous month	# friends in the current month	Support	Confidence
1	0	5,599,358	97,972 (1.7%)
2	0	512,353	48,552 (9.4%)
3	0	1,247,393	157,503 (12.6%)
1	1	968,621	109,395 (11.3%)
2	1	3,478,052	535,693 (15.4%)

Table 8.1: Results of the matching of various patterns on the e-shop subnetwork

We looked also on patterns, in which some of client's friends shop in the previous month and some of them in the month of consequence relationship. As we can see, the confidence of such patterns is even higher, but still low. The higher probability of such pattern may be caused by the bigger influence – the two friends talked about this e-shop only recently.

If we look at the specific e-shops, which were most often matched for the pattern in the third row of Table 8.1, we see mostly e-shops related to the transportation, e.g., **Taxify**, **Uber** and **Zluty.cz**. The third has high support also among all clients, as we saw in Section 6.5, and its support in these matches is not significantly higher. On the other hand, **Taxify** (today known as **Bolt**), which is not very high in the list of the top e-shops, appears to be significantly more popular among pairs of friends. These e-shops are then followed by popular e-shops like **Alza.cz** or **DameJidlo.cz**, which are most probably matched only by coincidence caused by high overall number of transactions.

8.2.2 Membership Closures with Loans

Similarly we can examine the friendship influence on the bank's products, e.g., loans. We will ask whether the fact, that one (or more) friends taking a loan in past months, increases or decreases the probability of a client taking a loan. The pattern, which will be matched, is shown as an example of the membership closure in Figure 8.3.

As the friendship relationship is based on the transactional history of a client, we will consider only clients with a transactional history. Therefore, we will ignore clients, who only took a loan in our bank and their transactional history is in possession in some other bank.

# friends in the previous month	Support	Confidence
loans		
1	39,237	1,422 (3.6%)
2	1,688	93 (5.5%)
3	435	24 (5.5%)
overdrafts		
1	9,823	157 (1.5%)
2	165	10 (6.0%)

Table 8.2: Results of matching of patterns containing loans and overdrafts

Mortgages have low support in general, and their support among the clients with friends is below a reasonable threshold. Therefore we focused more on loans and overdrafts. Results are presented in Table 8.2. As we can see, approximately 3.6% of clients applied for a loan only a month after a friend of theirs applied in the month before. The confidence increases with the number of friends with a loan. However, as the number of these cases is under 100, we cannot confirm correlation between the number of friends with a loan and applying for a loan.

8.3 Spending After Taking a Mortgage

We proposed questions regarding spending and bank products in Section 6.7 – are there certain stores or e-shops, which are visited by clients who have just recently applied for a mortgage? We will be matching the pattern composed of the following rules:

1. a client applied for a mortgage in some month;
2. the same client shopped in an e-shop/in a store in the next month (or more months);
3. the client did not shop in this e-shop/store in previous months.

In contrast to the previous cases, no rule is marked as the consequence rule, as we are not interested into the number of matches. We are looking for stores and e-shops, which were matched the most often.

8.3.1 Results

As for the e-shops, the matched ones correspond completely to the most popular e-shops introduced in Section 6.5, such as **AliExpress.com**, **iTunes.com** or public transportation e-shops. For the store patterns, the stores occurring in most matches are the **IKEA** stores, which, however, also belong to the most popular stores in general (see Section 6.6). The number of matches is not significantly higher in comparison to the other shops, therefore the connection of applying for a mortgage and shopping in the **IKEA** stores cannot be confirmed, even though it seems intuitive. However, the hardware store **Hornbach**, which does not belong to the most popular stores, occurs more often than in all transactions. Due to the low support (only 43 matches for the pattern with transaction in one month after mortgage) we cannot claim connection to mortgages.

Discussion

We introduced a way of mining evolution rules using graph patterns. We also used it for measuring the social influence between two friends, which are expected to shape each other's behaviour. Specifically, we measured the number of loans and overdrafts which were taken after possible influence. The confidence of these patterns was not high enough for us to be able to claim them as rules. In the analysis of patterns in e-shop spending we observed a higher support and confidence, but again not high enough to create useful evolution rules.

In the end we also discussed hypothesis drawn in Chapter 6. However, our procedure did not find any e-shop or store, which is used with a significantly higher probability when a client applies for a mortgage.

Conclusion and Future Work

The main goal of this thesis was to analyse network, which was created from the real-world operational data of a small bank, as if it were a social network. We wanted not only to confirm or reject interesting hypotheses from the financial world, but also to provide a walkthrough for analysing any similar dataset.

In the beginning we reviewed the input data and discussed their origin and quality. Although such network can be composed of a various nodes and relationships from different data sources, not all of them are equally useful or reliable. For instance, we have seen products and their ownership by a client, whose data are directly stored and maintained by the bank, then employers, which were inferred from the counterparties of transfers estimated to be salaries, and the friendship relationship, which is created entirely on the basic of transaction history of two clients and may not be real at all.

We then reviewed the concept of a social network, what properties social networks usually have and how to examine them efficiently. The most important techniques included analysis of degree distribution, key player analysis and community detection. However, not all properties can be examined in all social networks. We also introduced software tools, which allow to use the mentioned algorithms on the real data. Although custom adjusted solutions can be used for a more detailed analysis, these open-source tools should be at the beginning of examination of every social network.

Our next step lead to the analysis of basic entities of the domain. The knowledge of the domain, together with its verification in the data and possible comparison to the general population, is the essential phase of every data mining process. For instance, we were able to understand the demographics behind the clients of our bank, which was helpful during drawing conclusions of various SNA methods.

Then we focused more on the certain parts of the network at a greater level of detail. We examined the overall structure of the spending subnetwork and the possible distribution of degrees in it, which can suggest the evolution model of the network. For each type of spending we analysed the key players and the structural communities, which enable us to divide the clients into groups consisting of clients with similar spending. We revealed that there are shops, which, if often visited by a client, can suggest his/her behaviour.

Another subnetwork with more detailed analysis is the network of clients, employers and salaries. We described the process of salary inference, as the original data did not contain any information about the motivation behind the money transfer. Using the complexity and the various different relationships in the network we discussed hypotheses often connected to a salary, e.g., gaps between men and women, regional differences, etc. We observed, that some nodes and relationships are more reliable and analyses referring to them are particularly relevant. On the other hand, the inference process brought also relationships, which seems to be interesting at first, however, they must be examined with caution, e.g. friendship relationship or personality traits of a client.

The last analysis was aimed at the evolution of the network and possible social influence. We explained the concept of social influence and tried to measure it for

the friendship relationship. We expected that two clients, who are friends, may influence each other by recommending various bank products or shops. Although we did not mine any useful rule with a good support and confidence, we discussed possible influence observable by pattern confidence increased with more influence factors, in our case higher number of influencing friends.

In general, we believe that inference of the "hidden" network from the data, which do not seem to correspond to a network at first glance, and its analysis using the knowledge of the dynamic and growing SNA is a useful approach, which could be complementary to the traditional statistical approaches.

Data Issues and Ideal Network

Our network suffers from several data-related issues. The first issue is its small size. Although we discovered over 400,000 clients, the majority of them do not use our bank as their primary bank (with a regular salary and most of their payments). If we look at the list of the biggest banks in the Czech Republic (as of the number of clients in 2021[1]), presented in Table 8.3, we can see the existence of significantly bigger banks with a longer tradition. Most clients in late productive age or retirement are probably more conservative when changing their bank. Therefore, they are not very interested in new banks with modern products.

Bank	**# clients (in mil.)**
Česká spořitelna	4.5
ČSOB	4.2
Komerční banka	2.3
Moneta Money Bank	1.4

Table 8.3: Overview of biggest banks in the Czech Republic

This leads to the second issue of our data – the composition of clients do not reliably copy the composition in general population. Young banks mostly gain clients among the young adults, who pick their first bank, and among clients of other banks, who only use one product because of its convenience in comparison to their main bank, and the current account in our bank is only a "side effect". The third issue is related to the temporal dimension of our data – young age of our bank also causes shorter transactional history. We therefore cannot track changes in spending or income in the life of a client from his/her student's years to the retirement, like a bank with the lifespan exceeding 30 years could.

If our network had been created from the data of one of the dominant banks or, even better, from the data of all banks, these issues would have been solved. If we did not have to handle the anonymization of the data and also had an access to more data sources, e.g., the administration registers, we would get the ideal network with all data (personality, demographics, etc.) for every client. For instance:

[1] https://www.penize.cz/bezne-ucty/425357-nejvetsi-banky-v-cesku-zebricek-podle-poctu-klientu-i-penez

- our analysis of communities with risky behaviour would be more accurate, as we would have more detailed data about transactions, but also about loans and mortgages – there would not be a case, when we had client's full loan history, but no spending history;
- inference of the friendship relationship would be more reliable, because we could compare the full financial behaviour of two clients (although it would be more computationally demanding) – the analysis based on influence between two clients could be improved;
- having more data about employers (with the knowledge of all employees and possibly data from the government's registries, e.g., the Business Register[2]), we would be able to connect an employment in a certain economic segment and specific behaviour;
- thanks to the significantly higher total number of clients, groups of clients with the distinguishing behaviour would be also bigger, we would not have to deal with the support too low for a hypothesis confirmation.

We realize that such "ideal" network is not possible to create outside the bank itself due to the privacy issues. On the other hand, we suppose that with the larger and more complex network social structures would be even more visible and such analysis would bring conclusions applicable to the real life.

Future Work

Although we analysed and discussed the core parts of the networks, thanks to its complexity there are areas, which could be examined by similar methods and algorithms. For instance, we did not cover savings accounts, because in our network they are not as dynamic as current accounts. However, in general they may also contain interesting patterns.

Another simplification of the most analyses was the narrowing of timestamps to months. Even though it may be sufficient for some analyses, there are others, which might have better outcomes when considering weeks or even days. For instance, in evolution analysis the influence between two clients could be measured on a day-to-day basis. Although such analysis may be more demanding in time or space complexity of algorithms, it might bring interesting conclusions.

As we discussed in the previous section, the application of same or similar approaches on a richer network would most definitely bring more reliable statements about inferred social networks. Finally, the idea of social network inference is not limited to the banking world. For instance, any organization with its members, communication between them and arbitrary processes inside the organization would certainly be able to create a network with observable social structures.

[2]http://www.obchodnirejstrik.cz/

Bibliography

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994. URL `http://www.vldb.org/conf/1994/P487.PDF`.

Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, Jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0085777. URL `http://dx.doi.org/10.1371/journal.pone.0085777`.

A.-L. Barabasi. Network science book, 2015. URL `http://networksciencebook.com/chapter/9`.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science, vol. 286*, pages 509–512, 1999. URL `https://barabasi.com/f/67.pdf`.

V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/p10008. URL `http://dx.doi.org/10.1088/1742-5468/2008/10/P10008`.

Björn Bringmann, Michele Berlingerio, Francesco Bonchi, and Aristides Gionis. Learning and predicting the evolution of social networks. *Intelligent Systems, IEEE*, 25:26 – 35, 09 2010. doi: 10.1109/MIS.2010.91.

Anna D. Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1), Mar 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08746-5. URL `http://dx.doi.org/10.1038/s41467-019-08746-5`.

Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, Dec 2000. ISSN 1079-7114. doi: 10.1103/physrevlett.85.5468. URL `http://dx.doi.org/10.1103/PhysRevLett.85.5468`.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), Dec 2004. ISSN 1550-2376. doi: 10.1103/physreve.70.066111. URL `http://dx.doi.org/10.1103/PhysRevE.70.066111`.

P. Erdos and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

Diogo Fernandes and Jorge Bernardino. Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb. In *Proceedings of the 7th International Conference on Data Science, Technology and Applications* -

Volume 1: DATA,, pages 373–380. INSTICC, SciTePress, 2018. ISBN 978-989-758-318-6. doi: 10.5220/0006910203730380.

D Vasques Filho and Dion R J O'Neale. The role of bipartite structure in R&D collaboration networks. *Journal of Complex Networks*, 8(4), Aug 2020. ISSN 2051-1329. doi: 10.1093/comnet/cnaa016. URL http://dx.doi.org/10.1093/comnet/cnaa016.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. ISSN 0027-8424. doi: 10.1073/pnas.122653799. URL https://www.pnas.org/content/99/12/7821.

J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. 3rd Edition. Morgan Kaufmann Publishers, Burlington, 2011.

R. A. Hanneman and M. Riddle. *Introduction to social network methods*. 2005. URL http://faculty.ucr.edu/~hanneman/nettext.

I. Holubová, M. Svoboda, D. Bernhauer, T. Skopal, and P. Paščenko. Inferred social networks: A case study. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 65–68, 2019. doi: 10.1109/ICDMW.2019.00019.

P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 1912. URL https://nph.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-8137.1912.tb05611.x.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983. doi: 10.1126/science.220.4598.671.

J. Liu and T. Liu. Detecting community structure in complex networks using simulated annealing with k-means algorithms. *Physica A: Statistical Mechanics and its Applications*, 389(11):2300 – 2309, 2010. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2010.01.042. URL http://www.sciencedirect.com/science/article/pii/S037843711000097X.

R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, pages 95–116, 1949. URL https://doi.org/10.1007/BF02289146.

S. Milgram. The small-world problem. *Psychology today*, pages 60–67, 1967. URL http://snap.stanford.edu/class/cs224w-readings/milgram67smallworld.pdf.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), Feb 2004. ISSN 1550-2376. doi: 10.1103/physreve.69.026113. URL http://dx.doi.org/10.1103/PhysRevE.69.026113.

T. Opsahl. Triadic closure in two-mode networks. redefining the global and local clustering coefficients. *Social Networks*, 35:159–167, 2013.

José J. Ramasco, S. N. Dorogovtsev, and Romualdo Pastor-Satorras. Self-organization of collaboration networks. *Physical Review E*, 70(3), Sep 2004. ISSN 1550-2376. doi: 10.1103/physreve.70.036106. URL http://dx.doi.org/10.1103/PhysRevE.70.036106.

M. Ružička. Anwendung mathematisch-statistischer methoden in der geobotanik (synthetische bearbeitung von aufnahmen. 1958.

P. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Introduction to Data Mining*. Second Edition. Pearson, 2019.

K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks. *CoRR*, abs/cs/0702048, 2007. URL http://arxiv.org/abs/cs/0702048.

Stanley Wasserman and Katherine Faust. *Social Network Analysis in the Social and Behavioral Sciences*, page 3–27. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. doi: 10.1017/CBO9780511815478.002.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

Chu-Xu Zhang, Zi-Ke Zhang, and Chuang Liu. An evolving model of online bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 392 (23):6100–6106, Dec 2013. ISSN 0378-4371. doi: 10.1016/j.physa.2013.07.027. URL http://dx.doi.org/10.1016/j.physa.2013.07.027.

List of Figures

List of Tables