INVESTORS’ BEHAVIOUR ANALYSIS IN PEER-TO-PEER (P2P) LENDING PLATFORMS

A Thesis
Presented to the Faculty
of ISM University of Management and Economics
in Partial Fulfillment of the Requirements for the Degree of
Master of Management

by
Simonas Mulevičius
January 3, 2023

Abstract

This thesis analyses the investors' behaviour in the selected Peer-to-Peer lending platform with the aim of measuring clustering behaviour among individual investors. The literature review section provides multidisciplinary tools and methods to measure the connectedness and similarity of different investor groups. For example, the same UPGMA algorithm that is used to construct an evolutionary tree of different living organisms in Bioinformatics can also be used to construct a similarity tree of different investors in Peer-to-Peer lending platforms. Additionally, prior work in Social Sciences used the Triadic Closure property to measure the tendency of different networks to converge. Hence, this metric has been selected for this thesis as the primary tool to test the selected hypotheses. To measure investors' tendency to converge to similar investment strategies, Secondary Peer-to-Peer lending market data was collected from one Lithuanian Peer-to-Peer lending platform (i.e., *Gosavy.com*) using a custom data collection program with custom hardware. Collected information about 1778 unique loans and 8605 investors was used to prove that the Triadic Closure was present among the investors in the selected Peer-to-Peer lending platform. Additional experiments generated with Monte-Carlo simulation methods show that a similar effect could have been observed if investors were to invest randomly. Hence, a correlation effect has been observed between the number of loans the investors tend to have in common, given their past similarities. This observed phenomenon could be explained by the fact that investors who invest in many loans have higher chances of co-investing with more investors at random into the same set of loans. The thesis presents an application of a new unsupervised clustering technique that could be used to efficiently deanonymize humans in online transactional data, for example, in Blockchain. Unfortunately, strategies of main groups of investors could not be constructed reliably because the choice of the clustering technique greatly affects the size of clusters and, hence, group-internal strategies. By reading this thesis, individual investors would benefit by learning how they can reconstruct individual competitors' portfolios by looking at their past investment decisions. In the meantime, Peer-to-Peer lending platforms could construct a loan recommendation system based on the existence of the observed Triadic Closure property. Finally, this thesis demonstrates that market regulators should ensure the anonymity of individual investors by requiring lending platforms to preserve k-anonymity.

Keywords: Investors' behaviour analysis, Peer-to-Peer lending, Unsupervised clustering, Triadic Closure

Table of contents

List of figures

1. Introduction

1.1 Relevance of the Topic

The current high inflation rate forces investors to seek high-yield investment opportunities. For instance, the Bank of Lithuania announced that in February of 2022, inflation in Lithuania was 14% (Bank of Lithuania, 2022), surpassing the returns of traditional investment tools. For example, S&P 500 index grew 6.6% on average over the last twenty years (Statista, 2022). However, nowadays, there are numerous different investment instruments that appeal to different investors of different risk tolerance levels. In addition to traditional equity or loan investments, today there are some more unconventional, innovative, and sometimes experimental ways to invest money. For instance, now both institutional and individual investors can invest in recent innovations in Finance Engineering: Initial Coin Offerings (ICOs), Initial Public Offerings (IPOs), crowdfunding platforms and Peer-to-Peer (P2P) lending platforms (Chen, Hu, & Ben, 2021). This thesis analyses investors' behaviour in online Peer-to-Peer lending platforms as it is a relatively new personal finance innovation. Moreover, as presented in the subsequent Literature review section, there is evidence that some investors investing in Peer-to-Peer lending platforms are losing money. As a result, a better understanding of different investors' behaviour patterns is required. Hence, the focal point of this thesis is to answer whether investors tend to converge to similar investment strategies. Such analysis has been enabled by the available online transactional data of the selected Peer-to-Peer lending platform. Finally, this thesis provides recommendations for required additional legislation based on the large collected quantitative data set. Implementation of some of the recommendations would facilitate sustained growth of the Peer-to-Peer lending community.

1.2 Research Question, Aim and Objectives

The research aim. Determine past changes in investment strategies among different groups of investors that invest in Peer-to-Peer lending platforms.

The research question: What are the past changes in investment strategies among different groups of investors that invest in Peer-to-Peer lending platforms?

Finally, there are five **research objectives** for this thesis:

- Examine the academic literature on the theory of investors’ biases, herding behaviour and investors’ preferences in Peer-to-Peer lending platforms.
- Provide research methodology explanation allowing readers to reproduce the steps.
- Empirically evaluate if Triadic Closure property exists among different groups of investors in Peer-to-Peer lending platforms in Lithuania.
- Empirically evaluate if investors tend to converge to similar investment strategies and thus invest in the same set of loans throughout time.
- Provide theoretical discussion about the Triadic Closure property measurements and compare the behaviour of different groups of investors throughout time. In addition, offer recommendations on how individual investors and Peer-to-Peer lending platforms can benefit by analysing the investors’ behaviour while investing in Peer-to-Peer loans.

1.3 Research Design

1.3.1 Data source and collection method

In order to determine temporal behaviour similarities between different groups of investors, a history of transactions will be used. To be precise, for this study secondary market data from Lithuanian peer-to-peer lending platform will be used. These transactional records contain information about the actually issued loans (borrowed amount, date and time of borrowing, annual interest rates of the loan, payback period), borrowers’ characteristics (age, gender, education, household size, borrower’s income at the time of borrowing, number of dependant people that the borrower has to take care, debt-to-income ratio, city of the borrower, purpose for borrowing) and a list of co-investors that invested in given loans (data and time when every individual investor invested, the exact amount of money the investor invested and investors anonymised ID). Previous studies analysed investors’ behaviour by using a static snapshot of the market (e.g., Golovkina, 2022). But this research project has a unique access to secondary market records of the major Lithuanian Peer-to-Peer lending platform “Paskolų klubas”. The aforementioned records specify the amount of investment of each individual investor as well as the time when that investment decision was made. The initial investigation of the data sample reveals that about 12,000 loans are available and every loan usually has around one hundred co-investors that jointly lend money to the borrower. As a result, the unique and rich dataset enables the historical analysis of investors’ behaviour.

1.3.2 Analysis techniques

In order to make any inference from this data, the initial data pre-processing would have to take place. During this step, the raw secondary market logs would have to be transformed into more structured records about the investors and their behaviour patterns. Then, an analytical investigative analysis would have to take place. This analytical step constitutes of two parts. At first, based on similarity investors would have to be grouped into certain categories based on loans these investors invest into. Then, for every such group a time analysis of past investment decisions would have to take place. The main research objective would be to answer if investors tend to converge investing into loans with similar characteristics. A random sample of several investors reveals that certain investors tend to invest large sums of money (e.g., in the form of five-hundred-euro denominations) on an ad-hoc basis. Hence, the secondary research objective would be to answer whether these bursts of investments actually correlated between similarly investing investors.

1.4 Sequence of the Thesis

Here is a list of topics that are presented in this Master thesis. At first, the extended literature review on the topic of Peer-to-Peer lending behaviour is provided in the Literature review section. The Literature review section is split into two main areas – potential dangers for investors in Peer-to-Peer lending platforms and then solutions that aim to help investors. Based on the identified literature gap and recommended techniques, the Methodology section aims to explain steps that have to be taken to identify convergent behaviour among investors. Research findings section provides a summary of the results. Here, calculations of the Triadic closure property would have to be provided. The discussion chapter aims to explain why certain hypotheses were (or were not) rejected. Finally, the conclusions chapter provides a concise list of caveats and recommendations for both platforms, regulators and individuals based on the research findings.

2. Literature review

Peer-to-Peer lending serves an important role in the economy but there are numerous ways how P2P lending ecosystem can be derailed. As noted in the literature, Peer-to-Peer lending has far-reaching economic benefits for the general public. For instance, peer-to-peer lending helps the economy by providing both small and medium enterprises to get access to the additional funding (Cummins, Lynn, Mac an Bhaird, & Rosati, 2019). This is an important achievement because previously these businesses were from conventional banks as too risky clients (Cummins, Lynn, Mac an Bhaird, & Rosati, 2019).

2.1 Introduction to Peer-to-Peer lending

Peer-to-Peer lending platforms have proliferated only in the recent years. For example, in Western countries, the first peer-to-peer lending platforms (e.g., Zopa.com in the UK or Prosper.com in the US) were established around 2005 (Lee & Lee, 2012). While in Lithuania the first Peer-to-Peer lending platform SAVY (Kreditai.info, 2019) was established only in 2014 (Verslo žinios, 2022). Hence, there are only a handful of studies that analysed this relatively young P2P lending market in Lithuania as previous studies mostly focused on analysing investors' behaviour in peer-to-peer lending platforms from the US and China (Cummins, Lynn, Mac an Bhaird, & Rosati, 2019). In the recent years number of lending platforms in Lithuania grew substantially – at the time of writing there are 24 different platforms that have a license from the Bank of Lithuania to be Peer-to-Peer lending operators (Bank of Lithuania, 2022). With the growing number of platforms, there is a responsibility to educate novice investors about the advantages and risks of Peer-to-Peer lending markets.

According to the taxonomy, peer-to-peer lending is a type of crowdfunding (Cummins, Lynn, Mac an Bhaird, & Rosati, 2019). Peer-to-Peer lending platforms act as intermediates that connect both lenders and borrowers. To be precise, they allow individual borrowers to borrow money from a group of willing lenders directly without going through a traditional financial institution such as a bank (Chen, Lai, & Lin, 2014). These lenders, more frequently referred to as investors, can choose their investment preferences and then give money to a selected group of individuals that meet the criteria (Chen, Lai, & Lin, 2014). In the meantime, lending platforms serve several key roles: 1) platforms do limited checks on the potential borrowers to ensure that only valid candidates can apply for funding; 2) lending platforms ensure match-making process between investors and borrowers; 3) platforms take care of debt collection process (Chen, Lai, & Lin, 2014). However, the main difference between banks and Peer-to-Peer lending platforms is that lending platforms do not take a responsibility to guarantee that the borrowers will pay back the loans and this risk is then shared with individual investors. This in turn allows Peer-to-Peer lending platforms to cut operational costs because of the reduced overheads of not having to take the risk of issued loans (Maudos & Fernandez de Guevara, 2004). This is a root cause why Peer-to-Peer lending platforms offer higher returns than bank deposits. However, there are some other Internet-based process improvements such as automatic reminders, an ability for investors to invest automatically or aggregated debt collection which in turn allow platforms to further reduce operational costs (Chen, Lai, & Lin, 2014). These efficiency improvements result in greater returns for investors (Mild, Waitz, & Wöckl, 2015). Other minor advantages about Peer-to-Peer lending platforms include: 1) increased visibility and control that allows investors to offer money to borrowers that lenders prefer (Slavin, 2007) and 2) reduced barrier of entry for private investors by requiring lower initial investments compared to other means of investment such as stocks (Chen, Lai, & Lin, 2014). However, a study of one Danish-African peer-to-peer lending platform showed that individual investors usually do not have the required analytical skills to correctly estimate financial gains and risks associated with loans (Chen, Lai, & Lin, 2014). Authors of the study about the aforementioned Danish-African P2P lending platform demonstrated that an average investor of the platform was actually losing money (Mild, Waitz, & Wöckl, 2015). The main cause of this issue was that borrowers were charging insufficiently high interest rates for their high-risk loans issued to borrowers in Africa (Mild, Waitz, & Wöckl, 2015). For example, statistical analysis revealed that defaulted borrowers on average took loans with less than 2% higher interest rates compared with the borrowers who paid back loans in full (Mild, Waitz, & Wöckl, 2015). The risk of defaults, however, is an important aspect in P2P lending platforms because there is no amortisation mechanism (e.g., an intermediate bank) and lenders take the full risk in case borrowers become insolvent. Moreover, borrowers with worse financial situation are the main users of P2P lending platforms (Haewon, Byungtae, & Myungsin, 2012). As a result, to improve financial portfolio management it is the uttermost priority for investors on peer-to-peer lending sites to better evaluate the risks associated with an investment in risky loans.

But the question remains whether investors in Peer-to-Peer lending platforms are capable of discovering winning strategies that balance the risk with potential financial gains over the long term. To check this claim, one has to conduct a historical analysis of investors' networks. To be precise, one can consider two investors connected by a "co-investor" relationship if they borrowed money to the same borrower at least once. As a result, a network of co-investors (see Figure 3) might emerge in the analysis of Peer-to-Peer lending market. And it worth mentioning that a new school of thought, coined by D. Easley and J. Kleinberg, analyses different social phenomena through the lenses of networks. For instance, the network analysis can identify group boundaries with a relatively high precision – in one scenario an investigation of interactions between gym members and two rival couches helped researchers predict how this collapsing community would fall apart (Easley & Kleinberg, 2012). A simple network parameter correctly predicted for all but one gym member which couch gym members would select after the split of the gym (Easley & Kleinberg, 2012). Network analysis provides a set of tools that could be helpful for investors in Peer-to-Peer lending markets. However, relevant studies mostly focused on the borrowers’ side of the P2P lending networks (e.g., one study investigated impact of borrowers’ relationships to the success of them getting a loan (Mingfeng , Nagpurnanand , & Siva , 2013)). Hence, it would be interesting to switch the argument and focus on the investors’ networks instead. There are two major theories that support such claim. According to Efficient Market Hypothesis (which states that all market participants are rational agents) investors should invest, at least in theory, to those loans that give the highest long-term returns. Another supportive theory is called Triadic Closure property. It states that if A and B are interacting as well as A and C are interacting, then there is an increased chance that both B and C will start to interact too (Easley & Kleinberg, 2012). Triadic Closure property was first discovered in social networks and in summary it can be summarised as “my friend’s friend is also my friend”. In the context of Peer-to-Peer lending, existing Triadic Closure property would suggest that over time investors should converge and start investing in similar set of loans. In practical terms such investors’ network analysis would help both legislators and Peer-to-Peer lending platforms to identify whether individual investors tend to converge on the winning and thus sustainable strategy. Alternatively, irrational, and non-converging investors’ behaviour might reduce investors’ returns and hence it can be detrimental to both lending platforms as well as the underserved borrowers.

2.2 Potential dangers for investors in Peer-to-Peer lending platforms

In order to keep the Peer-to-Peer ecosystem alive, the investors have to get reasonable returns that would motivate them to continue investing in such risky loans. One of the problems is that investors in Peer-to-Peer lending platforms face a serious uncertainty when investing in loans. For example, it has been mentioned in the literature that borrowers are in the better power position as they know their current individual financial status better (Cummins, Lynn, Mac an Bhaird, & Rosati, 2019). Moreover, borrowers may choose not to disclose certain unfavourable aspects of their lives that could in turn lead to a lower financial score. Hence, as individual investors usually are not professionals, they can make poor financial decisions because of the asymmetric information about the quality of a loan (CarloSerrano-Cincas et al., 2015, p. 1).

However, investors also need to be aware of any possible biases they or other market participants might have. For example, one study found that personal characteristics such as beauty play a role in a person's likelihood of acquiring a loan even though these personal attributes do not affect the probability of loan default (Ravina, 2019). Similarly, another investigation revealed the presence of racial biases in peer-to-peer lending investors' decisions (Pope & Sydnor, 2011). Another similar example of bounded investors' rationality includes the study of Chinese Peer-to-Peer lending platforms that concluded investors tend to borrow money to people of the same race or borrowers who live in nearby locations (Chen, Jin, Zhang, & Yang, 2016).

Furthermore, investors' decisions may be influenced by other investors. Banerjee (1992) coined the phrase 'herding behaviour' which means that individuals follow the actions of the crowd even if there are signs that such behaviour might be detrimental to the follower. Within the context of P2P lending, this crowd-following is frequently observed. Interestingly enough it has been identified that in certain circumstances investors' herding behaviour acts as a positive mechanism to curb information asymmetry (Cummins, Lynn, Mac an Bhaird, & Rosati, 2019). For example, Haewon et al. (2012) analysed herding behaviour in a peer-to-peer lending platform in South Korea. In that case, the creditworthiness of the loan was estimated by individual investors as they could vote on the attractiveness of the loan in the presence of information asymmetry (Haewon, Byungtae, & Myungsin, 2012). On the one hand, the authors argued that this allows investors to better estimate the probability of loan default. But on the other hand, it seems that such a voting mechanism might introduce certain biases that could be apparent in a homogenous society. Another study concluded that herding in P2P lending auctions had a negative effect on investors' financial returns (Wang & Greiner, 2010). However, not all herding behaviour was deemed harmful. For instance, one study analysed a US-based peer-to-peer lending platform and identified the presence of so-called rational herding which is an investment strategy adoption throughout the time based on other investors' decisions (Zhang & Liu, 2012). On the one hand, these results seem contradictory at first. But on the other hand, such results occurred because different studies investigated different platforms that have distinct mechanisms of allowing investors to invest in P2P loans. Even though herding might have different consequences for the performance, investors should still be aware of the existence of such effects.

Another cause of potentially irrational decisions stem from the timeliness of events related to Peer-to-Peer lending markets. An analysis of peer-to-peer lending platforms in China revealed that individual investors react swiftly to the negative news about the platforms they invest into (Chen, Hu, & Ben, 2021). To be precise, it was revealed that investors usually tend to substantially reduce their investments into the peer-to-peer lending platforms that have recent signs that the owners of these platforms were planning on running away with investors' money (Chen, Hu, & Ben, 2021). In addition, the authors of the previous study emphasise that investors' behaviour change for a limited amount of time without any more noticeable long-term consequences (Chen, Hu, & Ben, 2021). This can be explained by the 'limited horizon' theory that suggests that more recent events have higher impact on the future events while impact of past events dwindles with time (Chen, Hu, & Ben, 2021). Furthermore, this study proposes a model that uses public attention as the mediator between the negative news about the platforms and investors' reaction (Chen, Hu, & Ben, 2021). However, the aforementioned study focused on the impact that these negative effects have impacted Peer-to-Peer lending platforms. But no recommendations for individual investors were provided. Hence, there is a literature gap about the investors' historical behaviour that this thesis will try to address.

Another future problem for P2P investors could arise from the fact that lending platforms have a greater power than individual investors. For example, a critical study of the major US-based Peer-to-Peer lending platform revealed that almost all the loans shortlisted in these platforms receive funding from the investors (Balyuk & Davydenko, 2019). Moreover, lending platforms use the historical transactional data to train machine learning models to better predict the risk of loans (Balyuk & Davydenko, 2019). Consequently, this historical knowledge is then used to give credit ratings to different loans. Then, using the automatic investment robots even novice investors can automatically participate in the Peer-to-Peer lending markets (Balyuk & Davydenko, 2019). As noted in the literature, on the one hand such financial innovations reduce the barrier of entry for the less educated investors, but on the other hand, such automatic investment mechanisms encourage investors to be more passive (Balyuk & Davydenko, 2019). Consequently, Peer-to-Peer lending platforms could gain substantial power that could be used to the detriment of the individual investors (e.g., platforms could decide to hike the transaction fees). This centralisation of power in the hands of platforms has not only the economic implications, but, as the authors of the aforementioned study emphasise, the ethical consequences (Balyuk & Davydenko, 2019). Nevertheless, competition and independent market regulation should keep these power struggles between investors and platforms at bay.

2.3 Solutions aimed to help investors in Peer-to-Peer lending platforms

In order to increase the returns from peer-to-peer lending, different scholars proposed several tools and insights that can help investors. For instance, one study created a simplified and improved decision support tool that allowed investors to quickly assess the loan quality (CarloSerrano-Cincas et al., 2015). The aforementioned analytical tool used only a handful of key parameters about the borrower (CarloSerrano-Cincas et al., 2015). After the analysis of the discriminatory power of several loan parameters (e.g., indebtedness or loan purpose) the most descriptive indicators were selected (CarloSerrano-Cincas et al., 2015). This analysis of peer-to-peer loan defaults in the USA revealed that a risk grade assigned by the P2P platform was the most significant parameter that determined the likelihood of loan default (CarloSerrano-Cincas et al., 2015). Such a decision support tool allowed investors to make better investments faster.

Another explanatory mathematical model was based on the data from Peer-to-Peer lending platforms in Lithuania. One of the most recent analysis investigated the factors that determine interest rates set by the loans that Peer-to-Peer lending platforms (Golovkina, 2022). As noted by the author of that study, lending platforms in Lithuania set interest rates themselves prior to posting the loans to the primary market where individual investors can invest (Golovkina, 2022). Hence, investors in Lithuanian Peer-to-Peer lending market are to some extent "price takers" as they do not have the power to negotiate different interest rates other than the ones offered by platforms. Other researchers identify this price setting practise as a problem and call this phenomena "reintermediation" because previous middlemen in the loan issuing industry (e.g., banks) were replaced with a new set of middlemen (i.e., Peer-to-Peer lending platforms) (Balyuk & Davydenko, 2019). As a result, for individual investors to thrive in Peer-to-Peer lending markets, investors have to be aware of the dynamics of the platforms. As previously alluded, one such key element of the Peer-to-Peer is the interest rate setting mechanism. Usually, the precise mathematical algorithm that sets interest rate for a particular loan is a proprietary black-box mechanism that only the Peer-to-Peer lending platforms know (Golovkina, 2022). In her analysis, Golovkina uses publicly available information about loans in Lithuanian Peer-to-Peer lending platforms to infer the impact of different parameters to the interest rates of the loan. The study reveals that duration of the latest employment and credit score were the most influential factors that determined the cost of borrowing in Lithuanian Peer-to-Peer lending platforms (Golovkina, 2022). However, credit rating is issued by the platforms based on the borrowers' financial (e.g., debt-to-income ratio) and personal information (e.g., age). Hence, as credit score depends on other variables, it should have been excluded from the analysis. The main reason being that the main focus of Golovkina's research was to provide an explanatory tool that could help borrowers assess their potential interest rates if they were to borrow from Peer-to-Peer lending platforms. Nevertheless, the aforementioned study finds that factors such as borrowing purpose or gender do not affect the cost of borrowing in Lithuanian Peer-to-Peer lending platforms (Golovkina, 2022). This previous analysis focused on the borrowers' side of the lending platforms (e.g., cost of borrowing). However, questions whether some investors in Lithuanian Peer-to-Peer lending platforms discriminate borrowers according to gender, age or location is still left unanswered.

2.4 Clustering Techniques to Group Investors

In order to analyse group behaviour, first of all different clusters of investors have to be established. As Peer-to-Peer lending field is relatively new, it would be wise to look for clustering techniques from different and more mature fields. For instance, the problem of grouping individuals based on their features has been already solved in Bioinformatics. To be precise, there are numerous algorithms and clustering techniques that help to group animals based on their similarity in genes (Compeau & Pevzner, 2014). Hence, similarly to DNR sequence, one could encode behaviour of an individual investor in a binary sequence of zeroes and ones. Each number at a particular location would correspond to a specific investor's decision – whether she or he invested into a specific loan. Assuming that there are L loans in the analysis, investor's signature would constitute L number of binary elements. A sequence of such bits would represent investor's *signature*. For the subsequent analysis, let's call this behaviour signature as a feature vector of length L.

Given this numeric representation of investors’ behaviour, the question remains how to compare different investors and group them. One such mechanism is called *k-means* algorithm (Compeau & Pevzner, 2014). This is an iterative process that helps to identify *k* hierarchical clusters in the given data (Compeau & Pevzner, 2014). K-means clustering technique combines data in bottom-up manner (Compeau & Pevzner, 2014). To be precise, each node (in this case, an investor) is assigned to an individual cluster. Then, at every step two closest clusters are grouped together and then replaced with a representative cluster that aims to combine features of both clusters (Compeau & Pevzner, 2014). The process is repeated until there are *k* clusters left. Provided that each investor is encoded as a feature vector of length *L*, the application of k-means clustering algorithm would combine different L-dimensional vectors until there would be only the predefined number of investor groups left.

To better understand constructed clusters of investors, a data visualisation tool is required. One way to represent the hierarchically clustered data would be to construct a dendrogram of all the investors based on their past behaviour (see Figure 29 and Figure 30). A dendrogram is a topological data structure where analysed entities are stored at the bottom, and entities are connected by a tree-like structure where a branch indicates that entities have a common ancestor or similarly behaving entity. In the context of investors’ behaviour analysis, each entity (or the so-called *leaf*) would correspond to an individual investor. Then, connecting arcs or branches would represent a local split in behaviour. In order to construct such behaviour similarity dendrogram, one could use UPGMA (“Unweighted Pair Group Method With Arithmetic Mean”) algorithm (Brunton & Kutz, 2019). Similarly to the k-means clustering technique, the UPGMA would build a similarity tree in a bottom-up manner. Finally, after constructing a similarity dendrogram one could further analyse the investors’ behaviour by first splitting them into the groups and then focusing on the group behaviour.

2.5 Conclusion of Literature Review

There is a rich literature on the topic of Peer-to-Peer lending. Some researchers focused on the problems that investors face (e.g., intrinsic issues related to the bounded investors’ rationality or power struggles between investors, borrowers, and platforms). Other academics investigated ways to help either borrowers or investors. However, thorough investors’ historical behaviour analysis of Lithuanian Peer-to-Peer lending platforms has not been conducted before. Hence, this Master thesis will aim to bridge the gap in literature by providing an extensive analysis on the possible strategy convergence of different investor groups in Peer-to-Peer lending Lithuanian platforms. To achieve that, Efficient Market Hypothesis with be tested using Triadic Closure property.

3. Research Methodology

3.1 Research Aim and Objectives

The research aim of the thesis. Determine historical changes in investor behaviours among different groups of investors that invest in Lithuanian Peer-to-Peer lending platforms.
The research aim of this chapter. Provide research methodology explanation allowing readers to reproduce the steps.

Finally, there are the following **research objectives for this chapter**:

- Provide a theoretical model and hypotheses of the analysis.
- Describe and explain the selected research design methodology, instrument and expected data sample structure. In particular, specific attention has to be paid to discuss potential weaknesses of the selected approach.
- Describe chosen research process steps.
- Provide ethical justification for the appropriateness of this analysis.

3.2 Research Theoretical Model and Hypotheses

3.2.1 Theoretical Model

Key ideas of this thesis are built on the following theoretical model (see Figure 1). In social sciences, difficult real-life phenomena can be simulated by a simplified network of entities and relationships between them. In this case nodes represent investors in Peer-to-Peer lending platforms while connecting arrows represent "co-investor" relationship. In other words, two investors are related by the "co-investor" relationship if they invested money into the same loan. As alluded previously in the Literature Review part, Triadic Closure property would be a tendency for co-investors over time to invest in the same loans as other investors who already have similar loans. Taking example provided below (see Figure 1), if Triadic Closure property was present in the selected Peer-to-Peer lending platform, then, provided that the first investor has some common investments with both second and third investors, then there would be a higher probability that these two investors (No. 2 and No. 3) would have a common investment at some point in the future than with some other random investor. As a result, if Triadic Closure property is present, then the investors' network would "tend to close" (Easley & Kleinberg, 2012) meaning that missing "co-investor" links would tend to form over time.

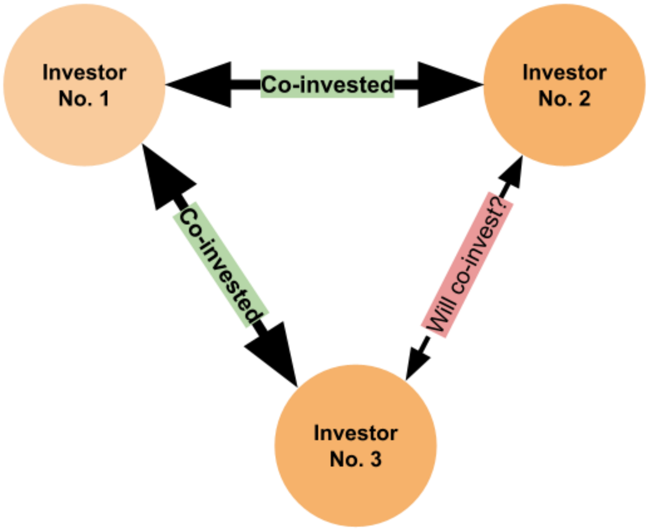


Figure 1. Triadic closure property in co-investors' network

3.2.2 Hypotheses

H1 – Triadic Closure property is not present in the selected P2P lending platform between groups of similarly investing investors.

H2 – Throughout the time different groups of investors observe a statistically significant shift in their average group investment parameters suggesting that investors tend to converge to the same group-internal strategies.

3.3 Research Design

The primary information source for this Master Thesis is secondary data – transaction logs collected from one Lithuanian Peer-to-Peer lending platform (i.e., "Savy"). Analytical data analysis will be conducted using a secondary market data of the selected P2P lending platform. A secondary market is a place where current loan owners re-sell their loans to other investors with or without a premium. In addition to the traditional loan parameters (e.g., borrower's income or indebtedness level) there will be more information such as loan notification history associated with the loan repayment. Furthermore, next to every loan advertisement, the selected platform provides a list of co-investors who also invested in the same loan. This additional information about the co-investors could reveal potential herding behaviour in investors' strategies. Finally, the selected information source is accessible to the general public[1] meaning that other researchers have a chance to test the results.

3.4 Research Instrument

Construct validity of the data sample is ensured by the selected Peer-to-Peer lending platform. To be precise, the aforementioned lending platform collects borrowers' information such as gender or age. In addition, certain parameters such as debt-to-income ratio are provided or checked by the external parties (such as credit scoring institutions). However, information about borrower's assets or the quality of these assets is not checked. Furthermore, there is a slight risk that the chosen Peer-to-Peer lending platform may show incorrect data but there is no way to evaluate this risk.

Content validity. Triadic closure property is measured by the level of interaction between several different entities. There are several proxies for this type of observation. For instance, some studies investigated electronic communication between two parties, while others observed day-to-day interaction between people (Easley & Kleinberg, 2012). Hence, to measure triadic closure, there are several aspects that have to be covered: there have to be some entities (in this case – lenders) that are connected by some relationship ("co-investor" relationship). This information is provided in the secondary market logs. To analyse the strength of these "co-investor" relationships deeper one can investigate further the number of overlapping investments that two lenders have made. Furthermore, in order to capture change of the Triadic closure one needs to have historical information about the change in the relationships over time. This can be achieved by using transaction logs from the secondary Peer-to-Peer lending market[2]. There one can find a list of co-investors that lent money to the same set of borrowers. Moreover, these transaction logs contain both the invested amount as well as the time of the investment. Hence, historical comparison could be performed using the secondary market data.

However, there is one caveat that not the entire market will be visible for this analysis. As the primary data source is secondary market, there might be some loans that have already been paid back in full or have been cancelled. Moreover, information about

[1] https://mano.gosavy.com/lt/antrine-rinka

[2] https://mano.gosavy.com/lt/antrine-rinka

some borrowers may be not accessible. For instance, in order to see information about the borrower and the list of lenders that gave money to the borrower, it is sufficient to find just a single investment that is being re-sold in the secondary market. In other words, if a single co-investor tries to resell the loan on the secondary market, then investment information about all the remaining co-investors would be also visible. However, there is a small chance that there might be some loans that are not present on the secondary market and thus are invisible. As a result, there might another group of investors whose behaviour could not be analysed if they do not tend to resell their current investments on the secondary market. Finally, another part of the market could be hidden by the fact that some investments are being bought automatically. To be precise, the selected Peer-to-Peer lending platform allows investors to buy and sell loans on the secondary market automatically. Hence, if there is a high demand for particular type of loans, then they could be quickly bought by the automatic investment function. On the one hand, for an external study such as this one it is difficult to estimate the actual extend of the hidden market information. On the other hand, the initial data analysis revealed that each loan was financed by around one hundred investors. This in turn means that a single investment sold on the secondary market can shed a lot of light on the market behaviour on the remaining co-investors. This last property of long co-investor lists suggests that secondary market data should be sufficient to draw generalisable conclusions about the overall behaviour of the market. However, one has to admin that the behaviour analysis of some niche investment patterns might be limited for a small number of investors.

3.5 Research Sample

For this analysis a limited target population sample will be used. To be precise, population size would be limited to just one lending platform from Lithuania. In addition, as alluded before, some part of the market behaviour could be missing because of the aforementioned factors related to hidden secondary market information. For this research project a population sampling of the whole Lithuanian Peer-to-Peer lending market was attempted but out of three major consumer lending platforms ("Savy", "Paskolų Klubas", "Finbee") only one agreed to participate in the study. Nevertheless, information about the 1778 loans from the selected Peer-to-Peer lending platform would be sufficient to conduct quantitative research.

One might want to sanitize the data and normalise it by selecting equal number of loans with different credit scores. However, in such scenario a lot of valuable information about behaviour peculiarities could be lost as number of loans with different credit score is unbalanced.

Other sampling techniques are not applicable because this study performs an observatory behaviour analysis and other techniques might be inferior. For example, during a qualitative interview a selected investor might find it hard to accurately describe his or her investment strategy that the investor used two years ago. However, a qualitative interview could provide more explanation about the way how the investors select the preferred investment strategy. For instance, the investor could pick the strategy based on the investment recommendations from other investors. Or alternatively, the investor could choose to avoid loss by not investing into loans that resemble previously defaulted loans from his or her portfolio.

3.6 Data Collection Method

Unconventional IT skills and tools will be leveraged for this research project. For example, market data about loans will be collected in an automatic way using an already constructed system (see Figure 2 and Figure 3). The loan extraction hardware system consisting of Raspberry Pi micro-computer (see Figure 2) and software (see Figure 3) can extract data from the selected P2P lending platform. This robotic data extraction technique is called *web scraping*. The automatic data collection tool is written in Python using Selenium library.

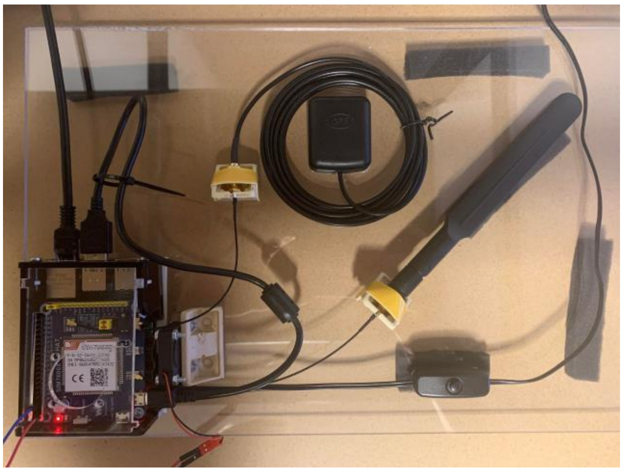


Figure 2. Required hardware for automatic market data extraction

The entire data analytics pipeline is presented in the following illustration (see Figure 3). At first, the loan extraction robot collects market information from the Peer-to-Peer lending platform. Raw information is then cleaned, transformed, and stored for subsequent analysis. Finally, as demonstrated in the last steps, investors’ clustering analysis has to be performed and some statistical parameters (such as Triadic Closure measurements) have to be calculated.

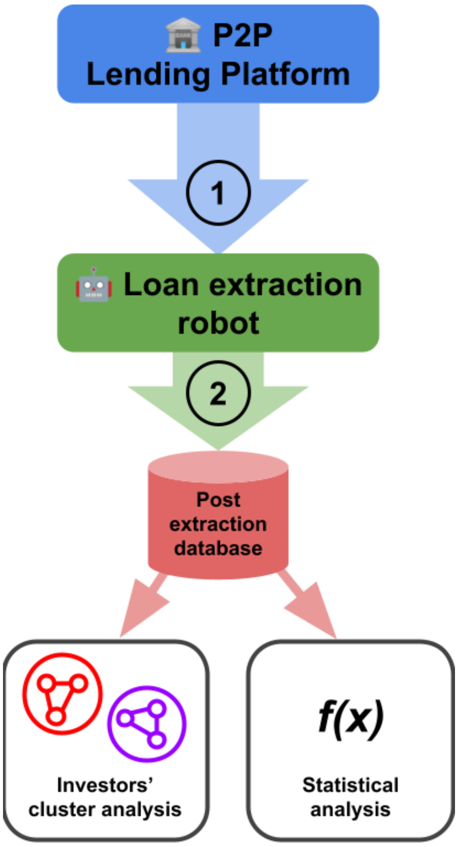


Figure 3. The architectural structure of the analytical pipeline

Finally, it is worth reiterating that the selected lending platform granted legal permission to use secondary market information for this research project. In order to mitigate potential disruptions this automated data collection tool can cause, the web-scrapping robot waits at least one second before making subsequent requests. This ensures that the data collection program will not cause a Denial-of-Service attack for the lending platform.

3.7 Research Process

After taking a static snapshot of the secondary market of the selected Peer-to-Peer lending platform, the data collection robot will store information in a specific database for further investors’ cluster analysis. Hence, one limitation of this study is that more historical data is not available. However, the initial investigation revealed that automatic data collection tool is capable of collecting all the market data within one day. Consequently, there should be no major inconsistencies related to the time it takes to collect the data.

The graph-oriented database (i.e., Neo4j[3]) will be employed as it allows users to construct and analyse different networks. As shown in the diagram below (see Figure 3), information from the database will be used to visualise clusters and analyse the behaviour of investors that invest using similar strategies. It is expected that there should be several main investment strategies which could correspond to noticeable clusters. For instance, in the example provided below bubbles (or nodes) represent investors while arcs (or relationships) represent that certain bubbles are related. For example, in the provided example, investor No.13 acts as a gatekeeper – it has co-invested in the same loans as investors from No. 11 to No. 15. However, investor No.13 also co-invested with investor No. 8 into another loan. Such example network suggests that there are primarily two groups of investors.

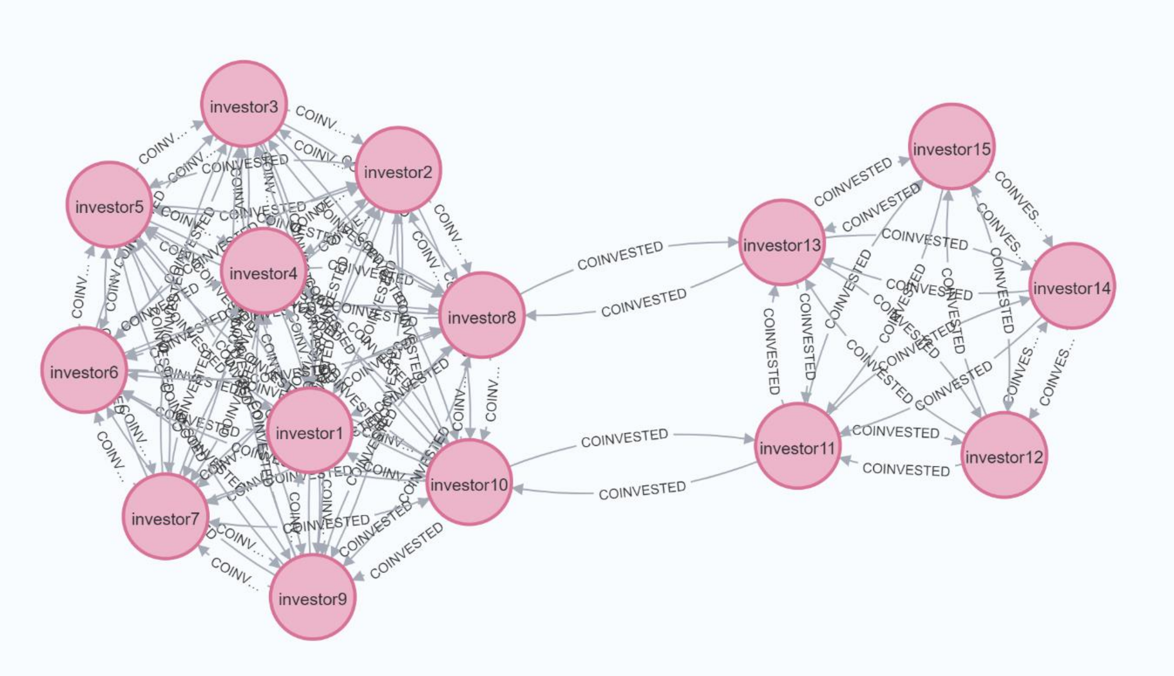


Figure 4. Anticipated structure of investors' clusters

This research project will build heavily on the ideas of the earlier analysis of peer-to-peer lending platforms. For example, the previously mentioned investigation of the predictive capabilities of different loan attributes (i.e., CarloSerrano-Cincas et al., 2015) did not include information about the household size. Furthermore, for the current research project it is possible to add several additional parameters such as the geographical location of the borrower. Location of residence may or may not affect the borrower's likelihood of solvency. As a result, investors could ignore this parameter when searching for the most optimal

[3] https://neo4j.com/

investment strategy. In addition, it is worth mentioning that not all results that were discovered by previous scholars will be applicable in this case because different peer-to-peer lending platforms operate differently. For example, on some sites borrowers are asked to provide their profile pictures. However, in the selected P2P lending platform personal information (such as race or profile picture) is hidden. Hence, results from Ravina’s study about investors’ discrimination based on race and appearance might not be relevant to this research project. Nevertheless, there are plenty of other parameters that investors can take into account while deriving an investment strategy. Different loan parameters could act as differentiators of different investor group boundaries.

3.8 Data Analysis Methods

Finally, after all the data pre-processing, a time series analysis would have to take place. To be precise, Triadic Closure parameter would have to be computed at different time intervals and then compared. Existence of Triadic Closure property in the investors’ network can be identified by calculating the tendency over time for investors to invest into a similar set of loans. For this property to be measured, another metric called neighbourhood overlap has to be introduced (Easley & Kleinberg, 2012). In the context of Peer-to-Peer lending investors’ networks the neighbourhood overlap would be equal to the number of loans that two independent investors have both invested into. But the initial data analysis revealed that there are certain investors whose investment portfolios exceed one hundred thousand euros whiles others invested only up to a thousand euros. Hence, in order to avoid potential skews because of the larger investment funds, neighbourhood overlap metric for two investors would have to be normalised by diving neighbourhood overlap by the total number of loans that these two investors have combined. As a result, provided that investors do actually converge on similar optimal strategies, throughout the time one would expect to see an increasing neighbourhood overlap between those investors who have identified the optimal sweat spot and the ones that are still searching for the optimal strategy.

It is worth mentioning, that Triadic Closure measurements would have to be compared over some time period. One way would be to slit data into two halves of equal duration. However, to better capture Triadic Closure dynamics (e.g., related to possibly altered investors’ behaviour after the Covid-19 pandemic), data will be aggregated on the monthly basis.

3.9 Ethical Considerations

As mentioned in the previous sections, data for this master thesis will be collected from the selected Lithuanian Peer-to-Peer lending platform rather than individual investors. Moreover, the provided data does not contain personally identifiable information about the investors – identities of all investors are anonymised, and only system-wide ID is used to identify individual lenders. As a result, an official permission to use publicly available secondary market data was obtained from the selected investment platform ("Savy") prior to the research. In addition, the aforementioned logs contain extensive information about the borrowers. However, these records do not contain personally identifiable information. Hence, GDPR consent forms from the borrowers are not required in this particular case.

In order to limit the footprint of this research, information used for the analysis will be deleted one year after publishing the results.

4. Empirical Research Results

This section starts with a high-level overview of the collected data. Then, co-investor network analysis is provided with the aim to identify the presence of the Triadic closure property among investors. After that, an in-depth hypothesis testing is provided to investigate whether an observed Triadic closure property is statistically significant. Finally, another hypothesis is analysed and several clusters of differently behaving investors are provided with the results of their behaviour change over the span of several years.

4.1 High-level Overview of the Collected Data

For the empirical study a data sample of the selected Peer-to-Peer lending platform is used. This analysis uses a web scrapped information from the secondary Peer-to-Peer lending market. To be more precise, an automated script written in Python programming language collected a secondary market snapshot of all the listed loans on the 19th of October 2022. The dataset has 1778 unique loans that received funding from 8605 unique Peer-to-Peer lending investors. See Appendix A for all 40 available features that were collected about each loan and borrower.

In order to independently verify the calculations, one would have to collect the same secondary market data sample from the selected Peer-to-Peer lending platform. However, the secondary market data is constantly changing so the independent validator would have to use Web Archive[4] to retrieve the same historical snapshot of the secondary market or ask the lending platform to provide the data.

4.2 Statistical Description of the Collected data

4.2.1 Information about the Investors

Using the collected loan information from the secondary Peer-to-Peer lending market, one can infer several parameters about the investors themselves. As each loan contains a list of co-investors who issued money to the specific loan, it is possible to reverse-engineer investors' portfolio sizes, number of loans and a list of similarly investing investors.

As alluded in the previous section, the study identified 8605 unique investors. Investor No. 114907 had the smallest total investment fund of one euro and a single

[4] https://archive.org/web/

investment. In contrast, the wealthiest investor on the platform was the platform itself (with a unique investor code "SAVY"). Collected data reveals the total visible portfolio of SAVY which stands at 2.9 million euros. Moreover, SAVY invested into 5498 loans which accounts to about 63.9% of all loans available on the secondary market. A plausible explanation of this is that SAVY re-invests money of the security fund. This SAVY security fund aims to spread the risk for individual investors by guaranteeing payments on behalf of late borrowers. However, it is worth mentioning that such risk-sharing service comes at a cost – investors who want to use SAVY security fund have to pay a premium when investing as part of their investment goes to this fund which is then reinvested to other loans. It is worth pointing out that SAVY's dominance in its own platform will be visible in the subsequent sections where similarity of different investors will be compared and it SAVY would play an important role there by sharing similar investments with almost all investors.

The analysis of secondary market revealed that a total of 13.7 million euros were invested in SAVY Peer-to-Peer lending platform. Comparing the total investment amount with SAVY's portfolio it seems that SAVY provided more than one fifth (21.2%) of all funds. Putting the portfolio sizes into perspective (see Figure 5), one can see that the green part of the pie chart which represents total share of SAVY's portfolio inside SAVY's platform is remarkable. Other wealthier individual investors (investor No. '1' represented by an orange slice and investor No. '42920' that is represented by a light blue slice) are dwarfed by the sheer size of SAVY's portfolio size. Moreover, it is worth mentioning the analysis revealed that SAVY was investing multiples of 500-euro investments into the same loans while individual investors are prohibited from doing so[5].

Other individual investors do not own a substantial part of the capital that is traded on the platform. Hence, wealth of individual investors in Figure 5 is consolidated into the groups based on the total investors' portfolio size. Medium investors who have made investments from 1,000 to 10,000 euros each constitute the largest group of investors – they provided approximately 6 million euros or 36% of all the platform capital. In contrast, investors who invested less than one hundred euros in the platform each, provide less than 1% of all the capital. Their combined contribution is less than the portfolio size of the second wealthiest investor on the platform (i.e., investor No. '1' with a portfolio size with more than 200,000 euros).

[5] https://gosavy.com/investavimo-rizika/

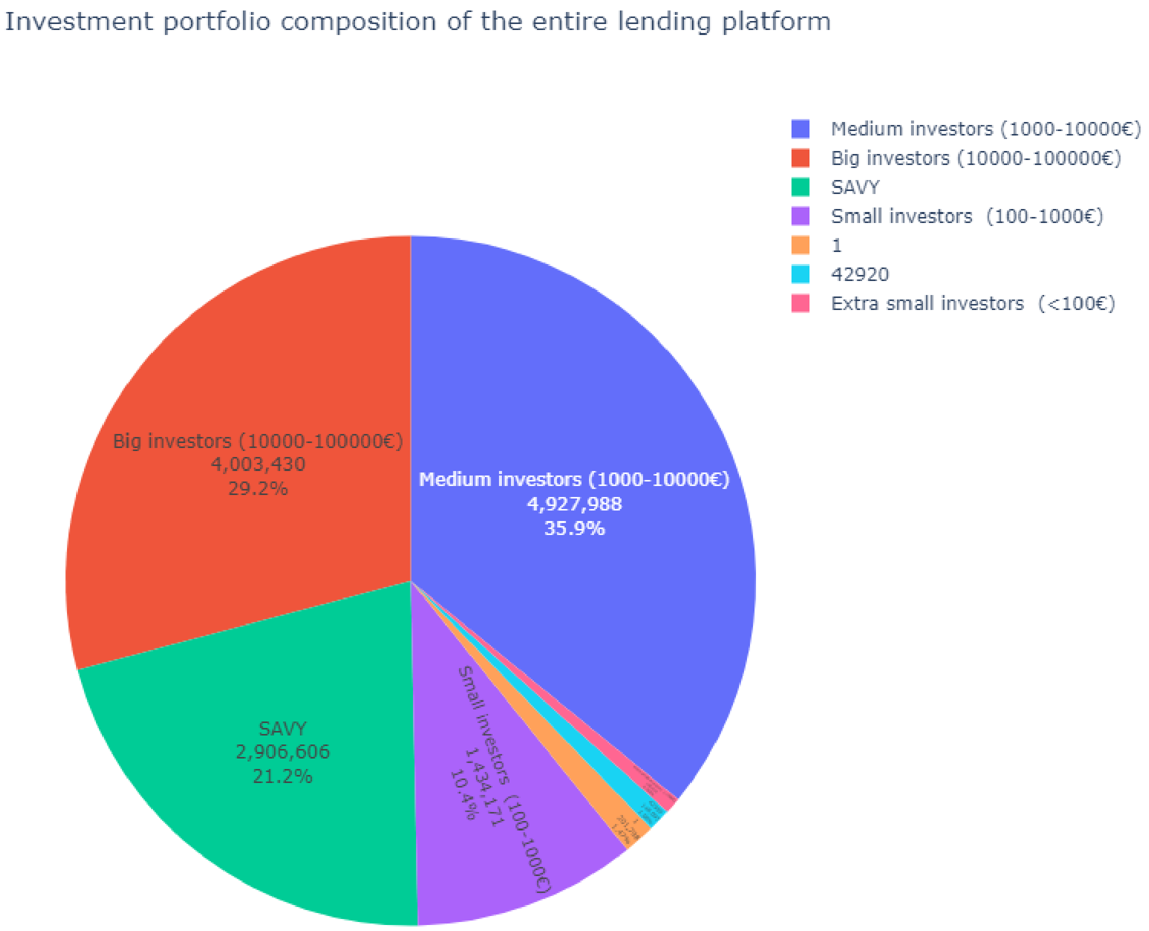


Figure 5. Entire platform investment portfolio composition

Wealth distribution of different investors is presented in Figure 6. Here investors are sorted by their portfolio size. SAVY's portfolio as well as the other nine remaining wealthiest investors have been excluded in order to make the contribution of smaller investors comparable. This diagram demonstrates disproportional wealth distribution among the investors. Investors with larger portfolios have more opportunities to invest in more loans compared with less wealthy investors. As a result, one could expect that a small proportion of investors would have made significantly more investments compared with the rest. As shown by the red Ordinary-Least-Squares trendline in Figure 7, this correlation does hold among the selected group of investors.

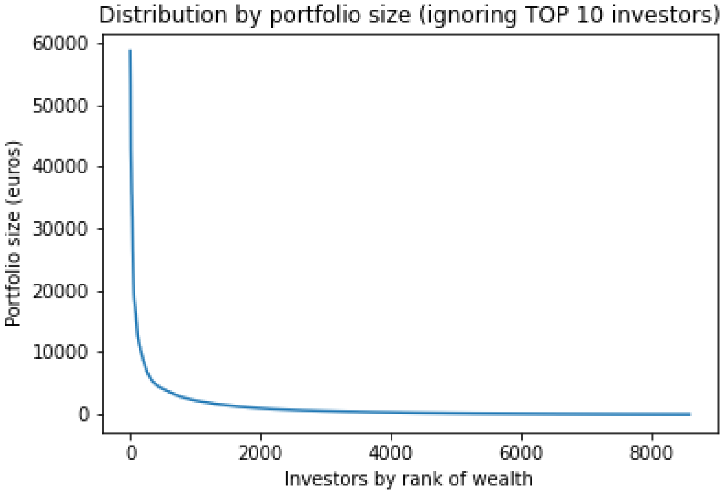


Figure 6. Distribution of portfolio size from the richest to the poorest investor

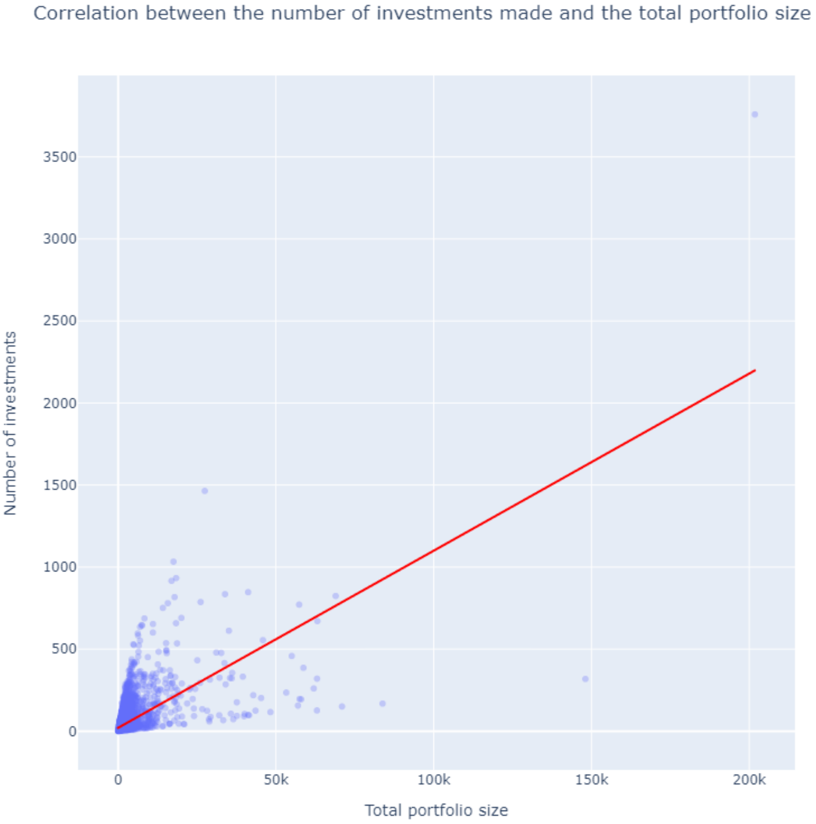


Figure 7. Correlation between the number of investments made and the total portfolio size

Finally, cumulative composition of the total portfolio (see Figure 7) reveals that the top 16.0% wealthiest investors contribute to 84.0% of the platform wealth (the red dot in the figure). Y axis in this cumulative distribution tells the reader the proportion of total wealth owned by the top X% of investors. By knowing this information, the lending platform could offer preferential treatment to the top investors (e.g., in the form of lower fees or the earlier ability to invest in lucrative loans than the general public).

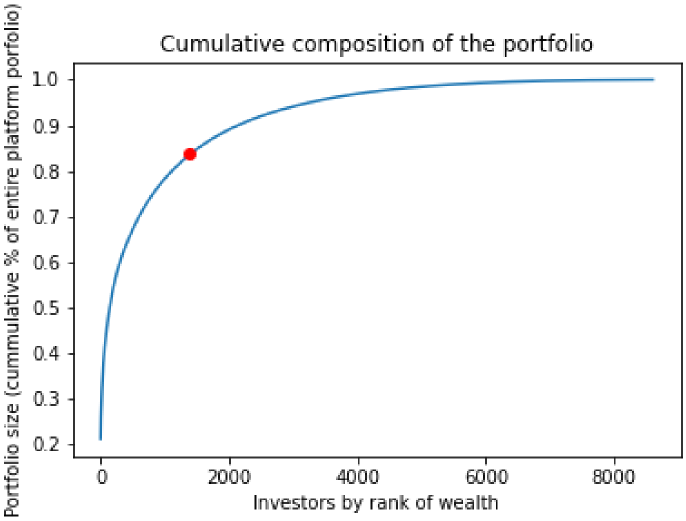


Figure 8. Cumulative portfolio composition

4.2.2 Information about the Loans and Borrowers

It is essential to analyse the distribution of different loan parameters that are listed in Appendix A. Different groups of investors could be compared by investigating distribution of specific loan parameters. For instance, one might expect that one group of investors might be more willing to issue money to women than men. Hence, a list of such distribution diagrams of the most relevant parameters is provided below.

4.2.2.1 Gender Distribution Among Borrowers

Gender distribution among borrowers is presented in Figure 9. From all loans, 713 of them were issued to women while 968 were issued to men. However, it is worth pointing out that there is no precise mechanism to disambiguate borrowers that took loans from this platform several times using the current data as personally identifiable borrower's information (such as name or personal ID number) is not presented in the secondary market loan listings. Hence, a simplifying assumption was made to assume that every loan was taken by a different person.

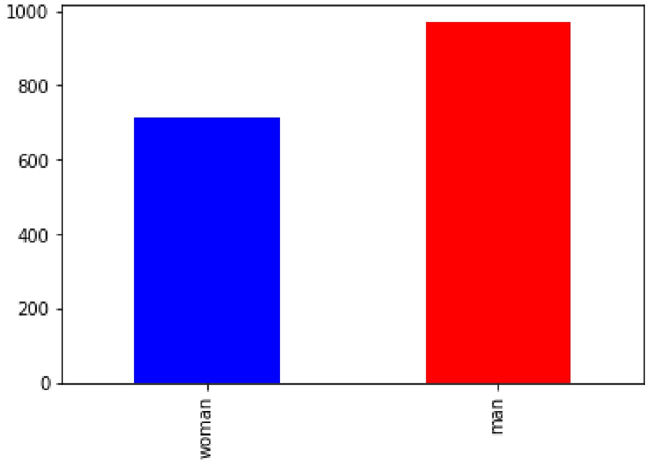


Figure 9. Gender distribution among borrowers

4.2.2.2 Age Distribution Among Borrowers

Age distribution among borrowers is presented in Figure 10. The median age of the borrower is 36 years while the average age is around 38 years. The range of ages varies from 19 to 72 years suggesting that even senior citizens are able to acquire loans in the selected Peer-to-Peer lending platform[6].

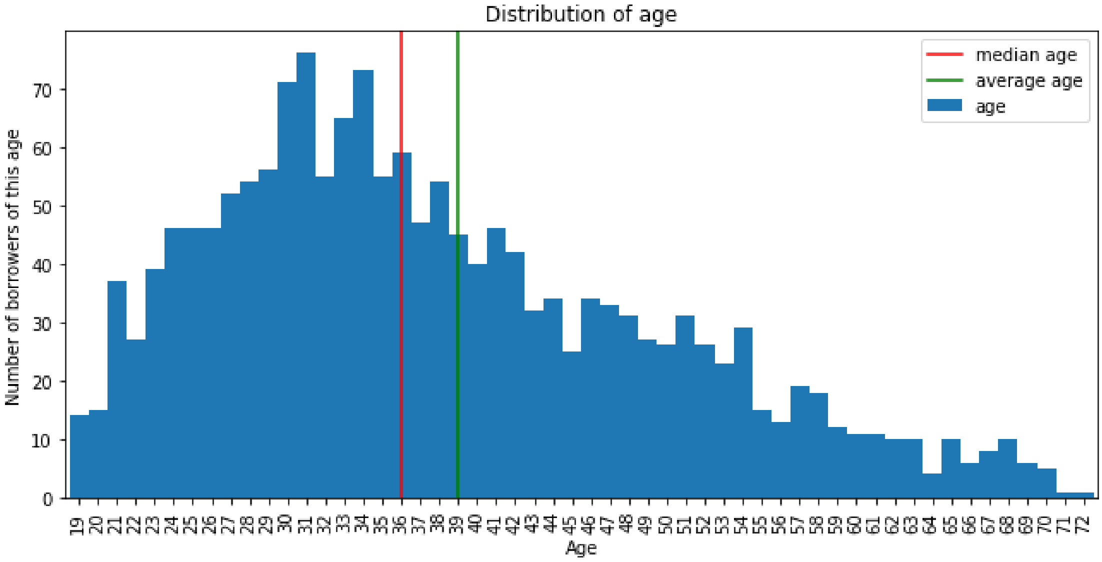


Figure 10. Distribution of borrowers' age

4.2.2.3 Geographical Distribution Among Borrowers

As demonstrated in Figure 11, geographical distribution of the borrowers is rather diverse. The most borrowers live in the major cities of Lithuania such as Vilnius (299 borrowers) or Kaunas (198 borrowers). Nevertheless, there is a so-called 'long tail' of

[6] https://www.sodra.lt/lt/senatves-pensijos-amziaus-lentele

borrowers from smaller cities, villages, and rural areas. Moreover, a significant proportion of all borrowers do not disclose city of their residence.

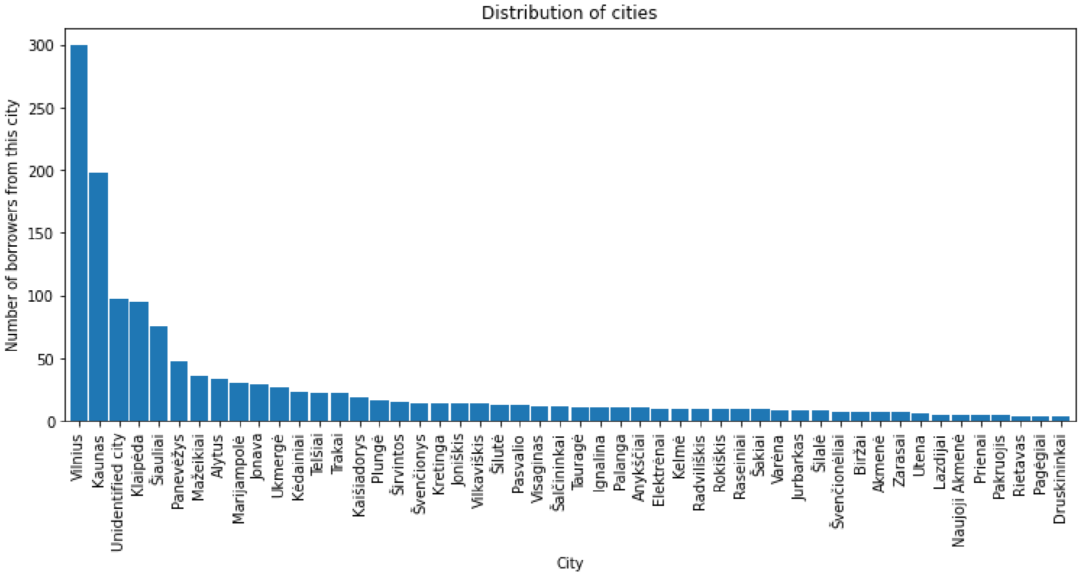


Figure 11. Top 50 most frequently mentioned cities

4.2.2.4 Credit Rating Distribution Among Borrowers

It is worth analysing the distribution of lending platform issued credit ratings[7]. The proprietary credit score rating aims to capture the risk of the loan. Furthermore, loans with the lower credit score are issued with the higher interest rates[8]. As shown in Figure 12, in the secondary market one could mostly find B and C category loans.

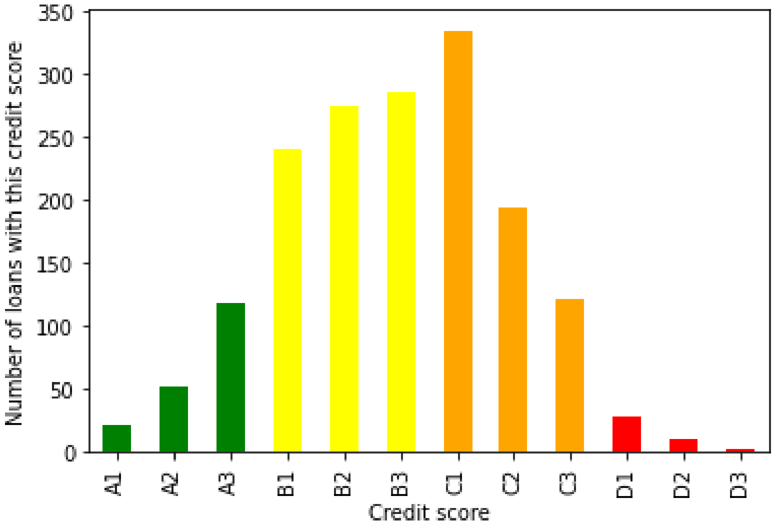


Figure 12. Distribution of credit scores by the number of loans

[7] https://gosavy.com/investavimo-rizika/
[8] https://gosavy.com/investavimo-rizika/

4.2.2.5 Total Loan Duration Distribution Among Borrowers

Another interesting insight comes from Figure 36 which depicts the most popular loan durations in terms of months. It is apparent that people usually borrow money for number of months that is equal to a multiple of twelve. For example, as emphasised by the red bars in Figure 36, the most frequent loan durations were 24, 36, 48, 60, 72 and 84 months which correspond to periods from 2 to 7 years. Hence, an interesting observation is that people do not tend to borrow for different time increments other than a year (e.g., not for 4.5 years)

4.2.2.6 Distribution of Repayment Delays

Repayment delays are presented in the pie chart below (see Figure 13). Most of the loans are paid back regularly. However, legal debt recovery process has been initiated for almost 10% of all loans. Proportion of defaulted loans could serve the purpose in the later steps of the analysis by providing a proxy metric that would allow us to evaluate the actual quality of loans. However, even better loan quality evaluation metric would have been an explicit actual repayment schedule which, for instance, would allow us to different slightly due loans (i.e., due for 30 days) from the more problematic loans (e.g., loans that are due for more than 90 days). More granular repayment schedule breakdown is present in another Peer-to-Peer lending platform “Paskolų klubas” but not in the selected SAVY Peer-to-Peer lending platform.

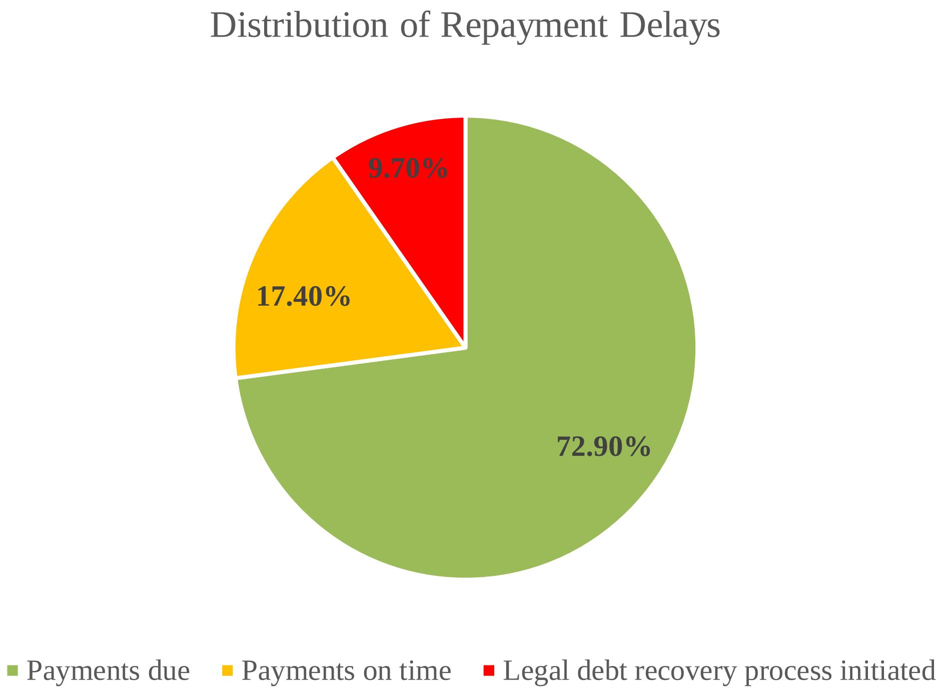


Figure 13. Distribution of repayment delays

4.2.2.7 Distribution of Loan Interest Rates

Distribution of the annual interest rates is presented in Figure 14. Most of the loans have interest rates from 8% to 24%.

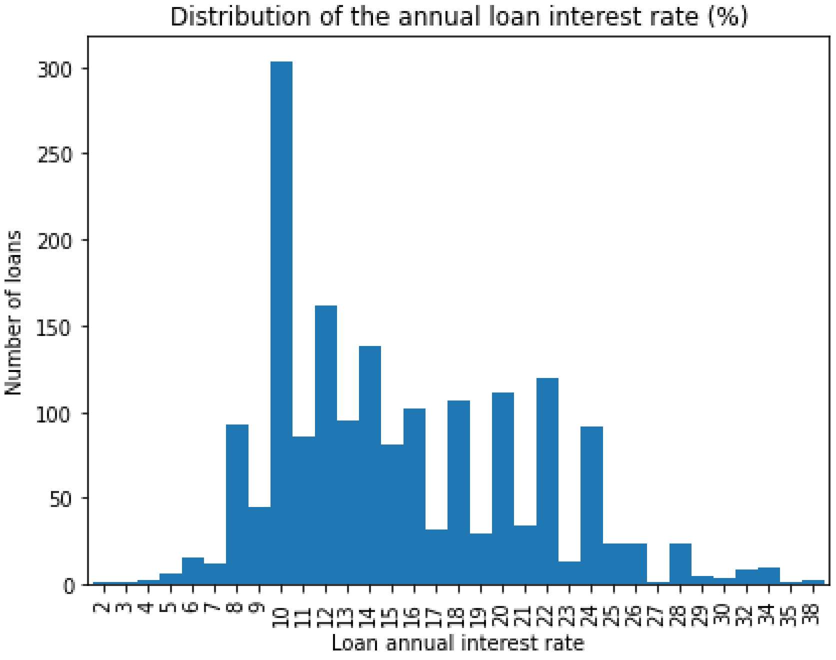


Figure 14. Distribution of the annual interest rate

4.2.2.8 Distribution of Loan Purpose

Distribution of loan purpose is shown in the pie chart below (see Figure 15). The vast majority of loans are taken to refinance the previously taken loans. Remaining loans are taken for consumption, house restoration projects or other niche uses cases such as purchase of a vehicle, buying real estate or financing a business.

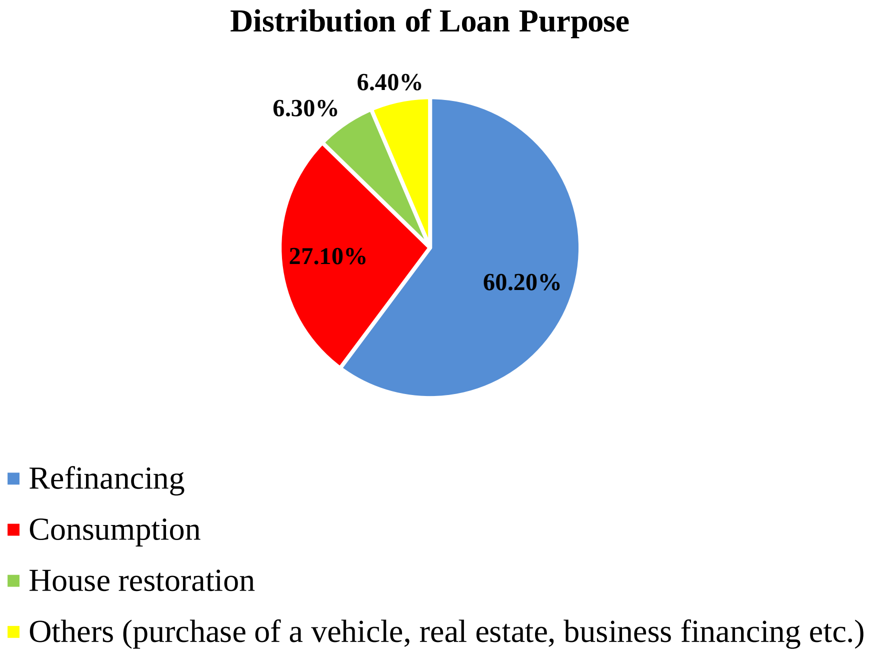


Figure 15. Distribution of Loan Purpose

Finally, remaining feature distribution charts were left out as they do not add any significant information that would help in the clustering analysis. For instance, borrowers' work industry turned out to be especially uninformative as from all the loans almost half of the borrowers (i.e., 888 to be precise) did not specify the industry they work in.

4.3 Measurements of the Triadic Closure Property

In order to measure a possible Triadic Closure property among different investors in the selected Peer-to-Peer lending platform, it is worthwhile to establish some common ground for further exploration. To begin with, in order to determine if this property exists, neighbourhood overlap has to be calculated first. The term neighbourhood overlap was mentioned in the prior analysis (Easley & Kleinberg, 2012) and in this context the neighbourhood overlap is equivalent to:

$$neighbourhood\ overlap\ (X,Y) = \frac{number\ of\ loans\ that\ \boldsymbol{both}\ investors\ X\ and\ Y\ invested\ into}{number\ of\ loans\ that\ \boldsymbol{at\ least\ one\ of}\ investors\ X\ and\ Y\ invested\ into}$$

Calculating a neighbourhood overlap ratio for all pairs of investors yields the following diagram (see Figure 16). On each axis there are different names of investors sorted by their portfolio size. Each cell at the coordinates *(x, y)* stores neighbourhood overlap score and the brighter the colour, the bigger the overlap. For example, there is an apparent diagonal line (where *x* is equal to *y*) in the middle of the diagram which states a trivial fact that an investor portfolio is identical 100% to itself (hence the maximum overlap). Furthermore, there are several apparent dark lines which indicate that certain investors have extremely low overlap with other investors. This peculiarity requires further attention.

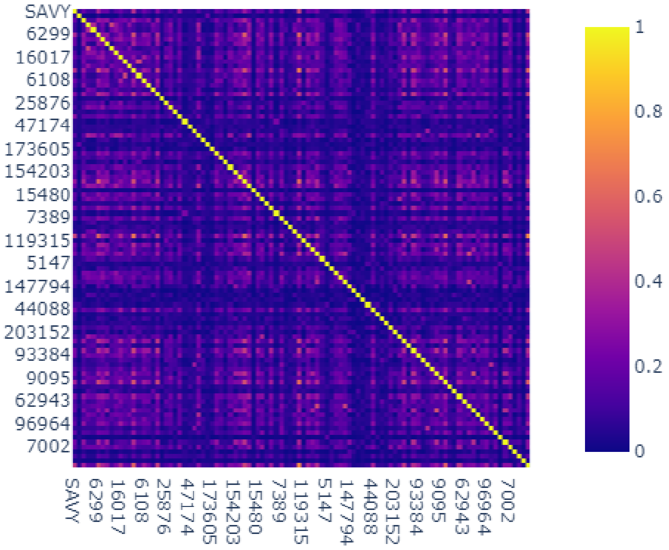


Figure 16. Relative similarity score matrix

Another similarly looking distance matrix uses absolute similarity score (see Figure 17). To be precise, a single cell at the position *(x, y)* in this matrix captures only the total number of overlapping loans as opposed to the relative proportion of overlapping loans. Absolute similarity score is one of the key building blocks for the subsequent Triadic closure calculation logic. It is worth emphasising the fact that in the absolute similarity matrix almost all the cells are dwarfed by the sheer size of loans that the SAVY investor (i.e., the platform itself) has issued (let's recall the fact that SAVY manages almost one fifth of all platform assets). Hence, one of the drawbacks of using the absolute similarity score is that information about the similarity of smaller investors is diluted by the bigger investors.

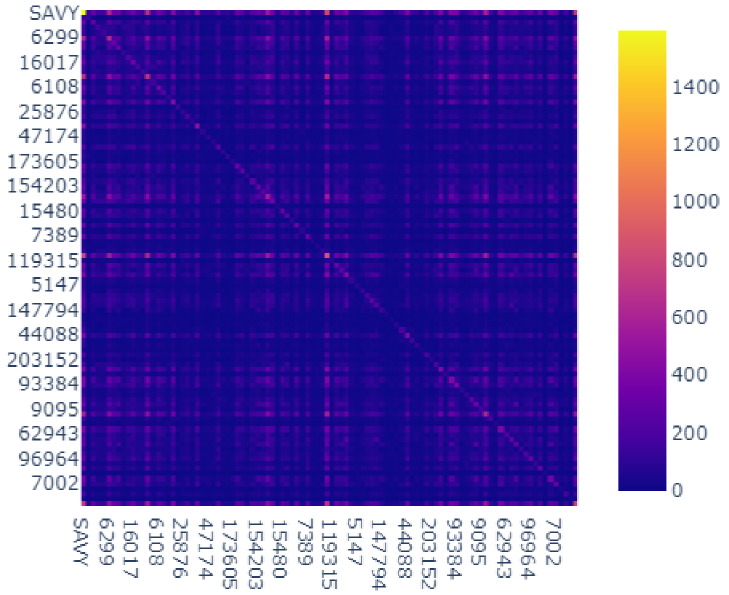
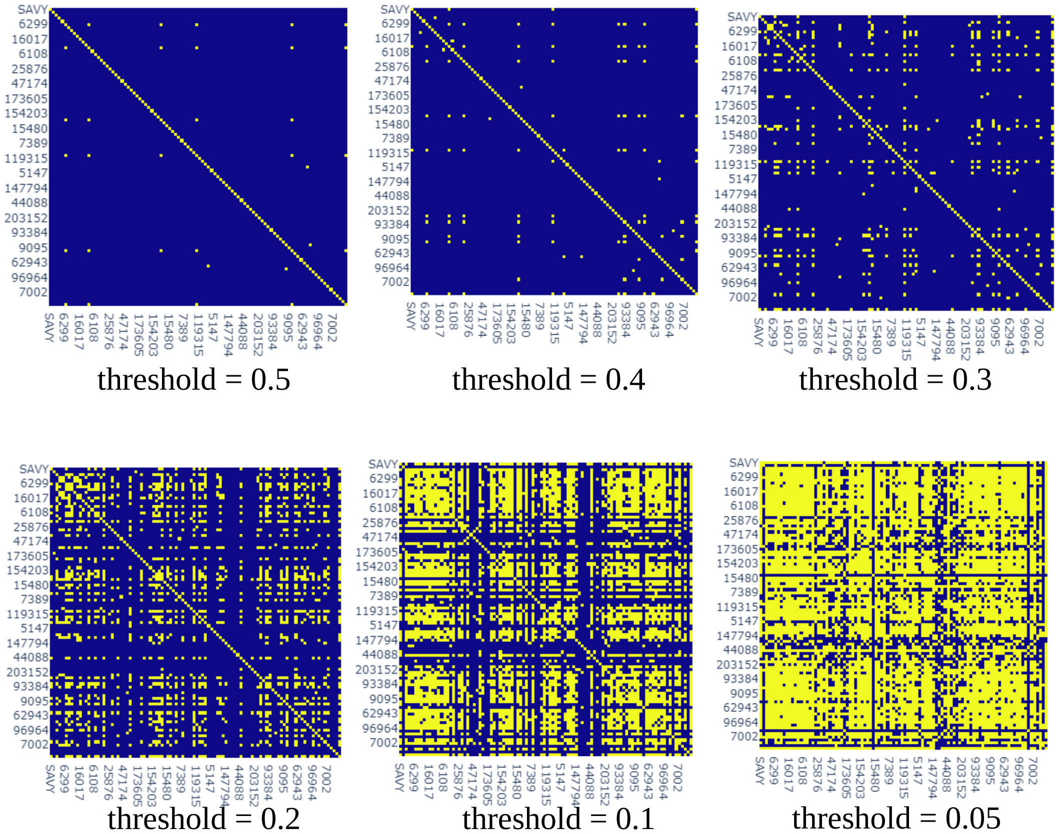


Figure 17. Absolute similarity score matrix

Before delving into Triadic Closure calculations, it is worthwhile analysing whether certain clusters emerge in the co-investors' network structure. For this purpose, a binary similarity score is used. To be precise, every value of the relative similarity score matrix (see Figure 16) is cast to either 0 or 1. Value of 1 is given only if the relative similarity between two investors is greater than certain threshold. Several examples of applying different thresholds to the relative similarity matrix is presented below (see Figure 18). Decreasing thresholds (0.5, 0.4, 0.3, 0.2, 0.1, 0.05) reveal an increasing number of connections between different investors that are denoted by yellow dots. One could call these new vested co-investor relationships as *strong co-investor relationships* because they exceed some predetermined threshold of chance. Separation between strong and weak relationships helps to establish a more precise definition of Triadic closure (Easley & Kleinberg, 2012). To be precise, Triadic closure states that if two pairs of entities are connected by a strong relationship (e.g., *A* with *B* and *A* with *C*), then if the Triadic closure is present, then the two unconnected entities (i.e., *B* and *C*) should at least establish a weak connection (Easley & Kleinberg, 2012).



threshold = 0.5 threshold = 0.4 threshold = 0.3

threshold = 0.2 threshold = 0.1 threshold = 0.05

Figure 18. Examples of binary relative similarity matrices

Visualisation of these strong co-investor relationships (using a threshold of 30%) yields the following result (see Figure 19). As previously mentioned in the literature, one could expect to see the naturally occurring networks to be well connected (Easley & Kleinberg, 2012). This is also the case in the co-investor similarity network – there is a clear tightly connected cluster in the middle with smaller disconnected groups left in the peripheral. Relatively large number of disconnected individual investors is due to the relatively high cut off threshold of 30%.

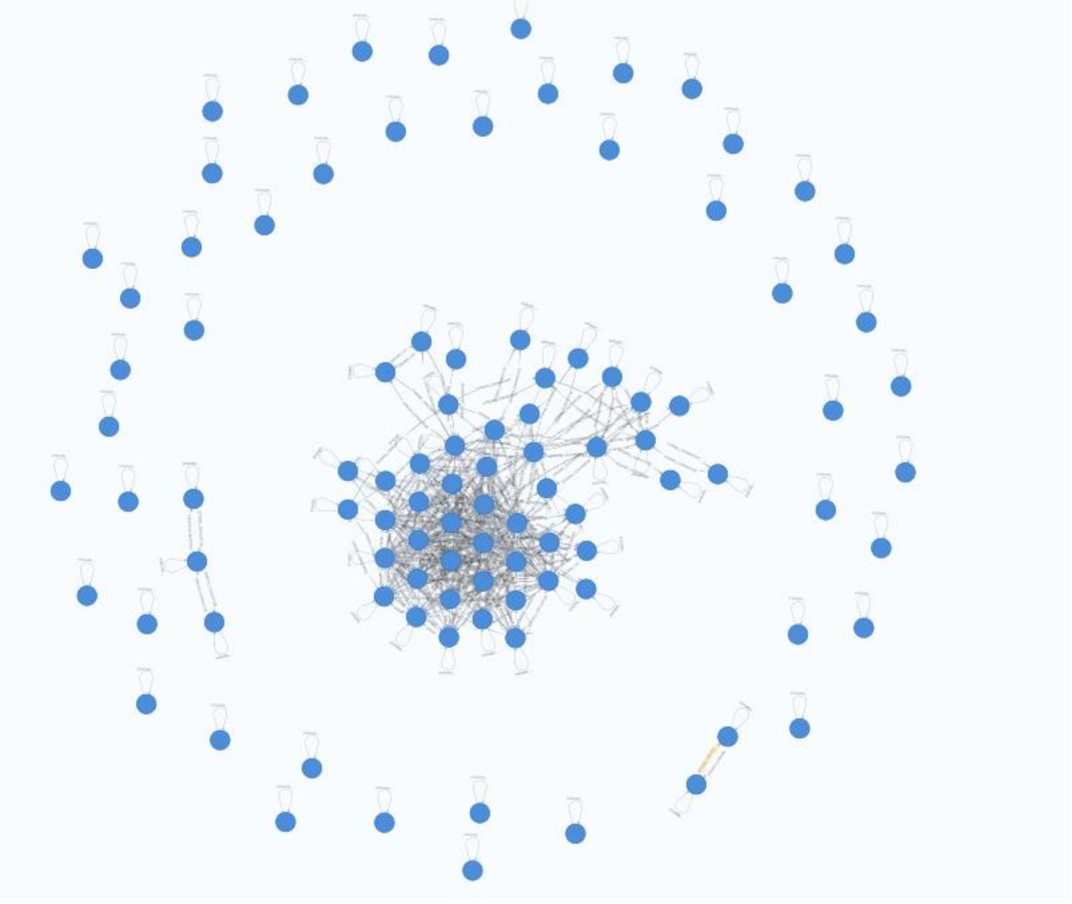


Figure 19. Network structure using 30% similarity cut-off threshold

4.4 Evaluation of the Triadic Closure Property Hypothesis (H1)

4.4.1 Model Update for Triadic Closure Property Hypothesis (H1)

The question remains unanswered whether the selected investors' network has a Triadic Closure property (H1). For more interpretable results the initial model to measure Triadic closure property in the co-investors' network (see Figure 1) has been slightly redesigned (see Figure 20).

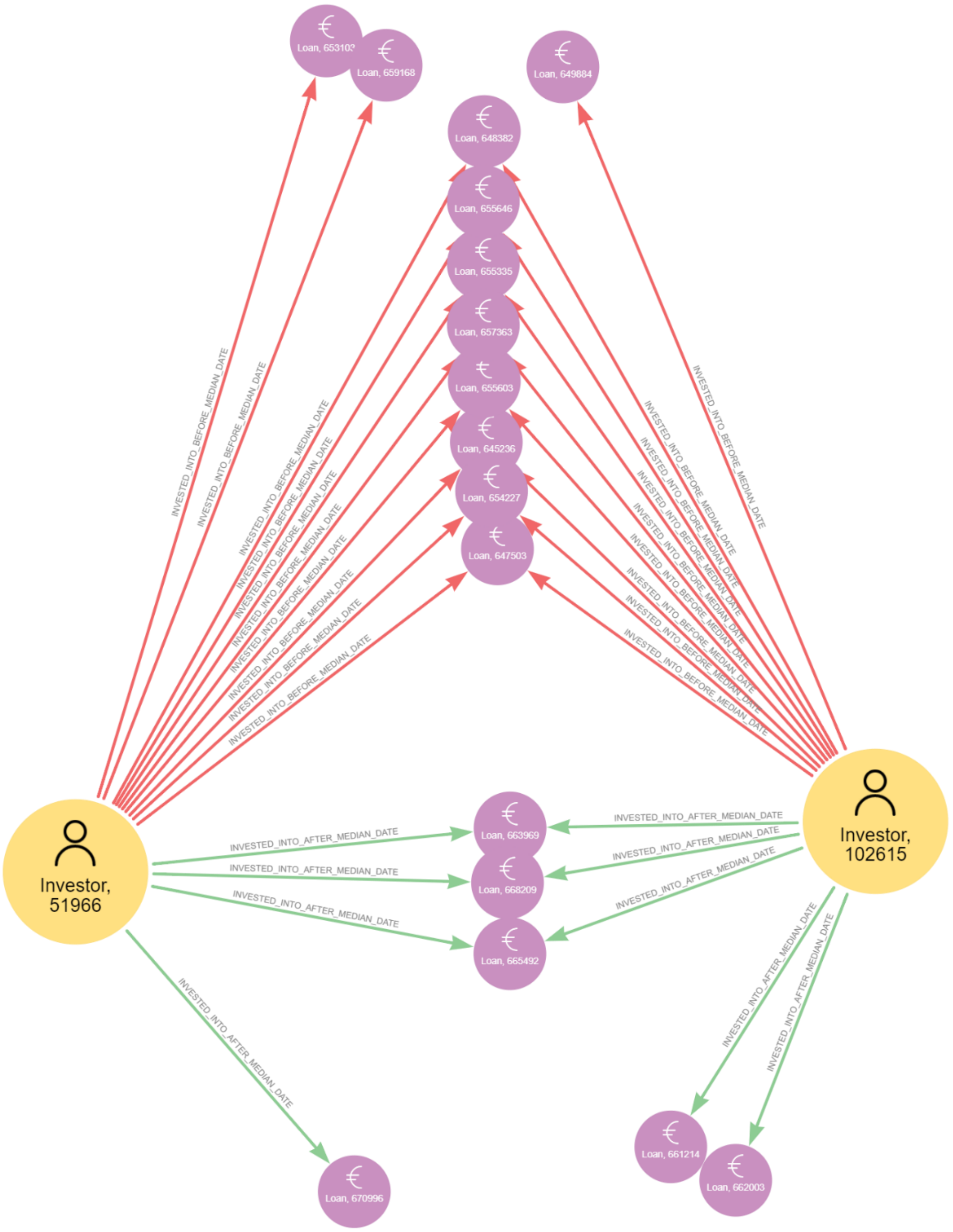

Loan, 65310
Loan, 659168
Loan, 649884
Loan, 648382
Loan, 655646
Loan, 655335
Loan, 657363
Loan, 655603
Loan, 645236
Loan, 654227
Loan, 647503
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
INVESTED_INTO_BEFORE_MEDIAN_DATE
Investor, 51966
Investor, 102615
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
INVESTED_INTO_AFTER_MEDIAN_DATE
Loan, 663969
Loan, 668209
Loan, 665492
Loan, 670996
Loan, 661214
Loan, 662003

Figure 20. Adjusted (loan-focused) Triadic closure property analysis model in investors' network

In this improved analysis model, the Triadic closure between two investors is defined by their tendency to co-invest into similar loans based on their past shared loans rather than their shared co-investors. In the diagram below (see Figure 20) each yellow circle represents an investor with their corresponding investor ID that is given by the platform itself. Each purple circle represents a loan that she or he invested into. Red and green arrows indicate relationships between investors and their loans. Red arrow indicates that an investor invested into a specific loan before the chosen *median date* (which corresponds to 1st of March 2022;

see Figure 23 and paragraph that follows for the precise derivation of this date). The green arrow demonstrates that an investor invested into a specific loan after that *median date*. Hence, the first hypothesis aims to answer the question whether or not the number of loans shared in the past (e.g., purple circles with red arrows pointing to them from both investors in Figure 23) have an impact on the number of loans the two investors would have in common in the future (e.g., purple circles with green arrows pointing to them from both investors in Figure 23).

4.4.2 Statistical Evaluation of the Triadic Closure Property Hypothesis (H1)

The initial exploratory analysis forms the foundations that will help to select the correct statistical tools to approve or reject hypotheses of this Master thesis. For instance, figures from the previous section (i.e., Figure 6, Figure 16, Figure 18) demonstrate that the small number of rich investors have relative large investment portfolios which in turn result in them having more *co-investor* connections than the less wealthy investors. This observation does not pose problems for the evaluation of the first hypothesis, but it will come handy when explaining the observed effects. However, observation of the absolute similarity matrix from the previous section (see Figure 17) does present some challenges. If one was to apply similarity threshold of one loan to the absolute similarity score matrix (see Figure 17), then the result would be the following:

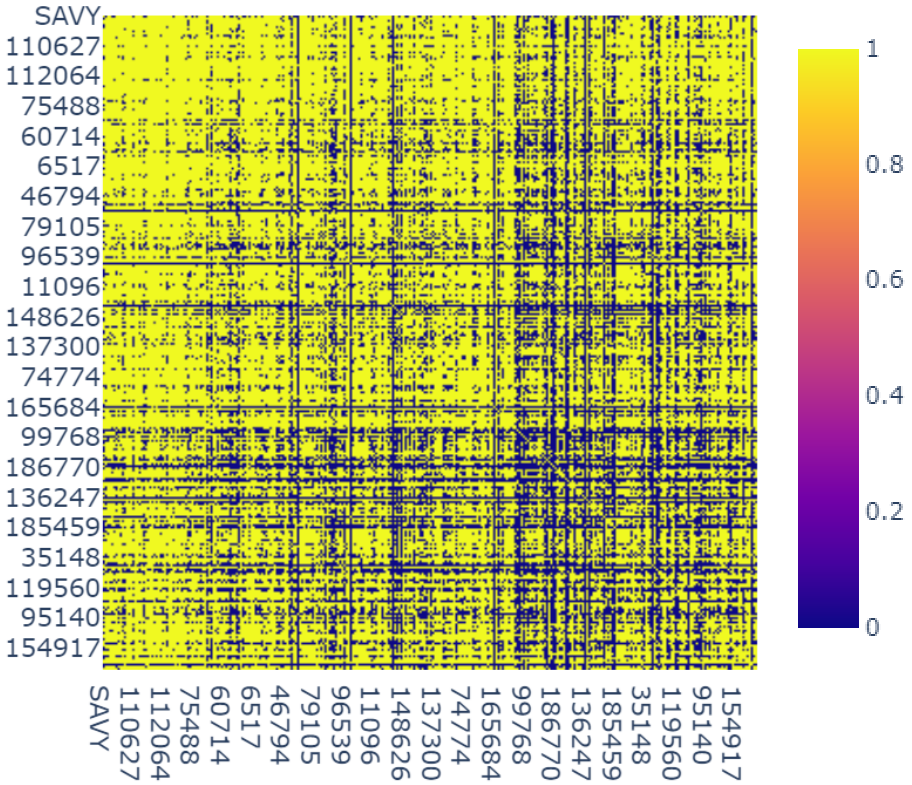


Figure 21. Absolute similarity matrix with a similarity threshold being 1 loan (TOP 1000 wealthiest investors)

Similarly, to the previous diagrams (e.g., Figure 18), Figure 21 shows that two arbitrary investors *x* and *y* are connected provided that they have at least one loan in common. Here yellow colour in the matrix position *(x, y)* indicates that the two investors share at least one loan in common while blue colour indicates that they do not have a shared loan. However, as demonstrated in Figure 21, almost all richest investors have at least one loan in common with the fellow investors. This in turn would lead to the fact that almost all investors would be connected to almost every other investor. This phenomenon that most of the nodes on the network are connected to one large sub-net is not new – previous researchers called the biggest interconnected sub-net as the *giant component* (Easley & Kleinberg, 2012). Prior work that analysed the existence of the Triadic Closure property in the online forums (Easley & Kleinberg, 2012) did not have this problem as the decision whether the person would become a member was binary – a person either joined or did not join the online community given that some of her or his friends had already been members of it. In the case of the co-investors network one can choose a threshold when the two investors can be considered connected. The diagram below (see Figure 22) helps to choose a threshold (the minimum number of loans the two investors need to have to be called true co-investors) that would produce relevant and practical results.

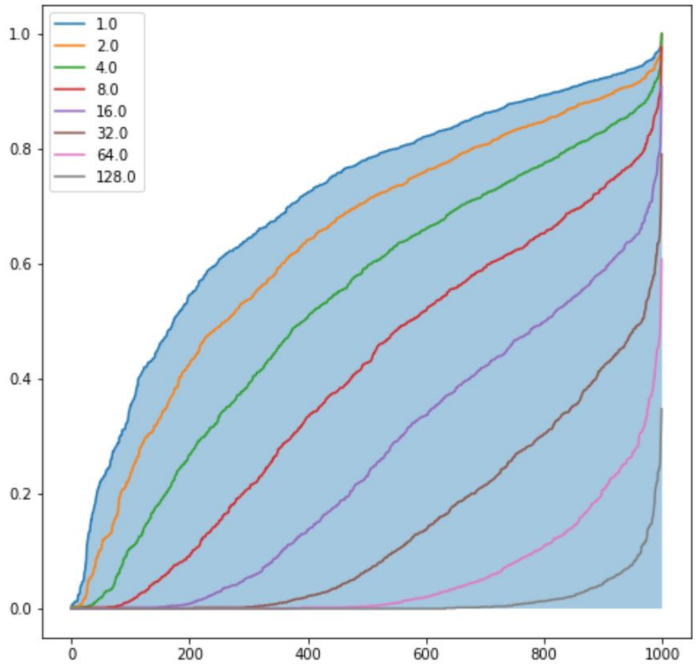


Figure 22. Cumulative distribution function of the proportion of the wealthiest 1000 investors that have at least 1, 2, 4, 8, 16, 32, 64, 128 loans in common with other rich investors (the most connected investors being on the right)

As demonstrated by the blue line in Figure 22, if the similarity threshold was set to 1, then median investor (500th most connected investor) would be connected to almost 80% of other investors. In contrast, if the threshold was set to 128 loans (grey line), then almost no investors would be connected except from the few richest investors. Hence, the middle ground (red line) was chosen – two arbitrary investors are considered to be connected by the *co-investor* relationship provided that they have at least 8 loans in common.

To evaluate the existence of the Triadic Closure property the data about loans is split into two time periods (before and after the selected median date). The first set of investments has all the loans that were financed from the time the lending platform was created until the selected *mid date*. The second half of the selected subset of data has all the loans from the *mid date* until the time the data was collected from the peer-to-peer lending website. This second period is shorter timewise because there were more loans issued on the platform in the recent years (see Figure 24).

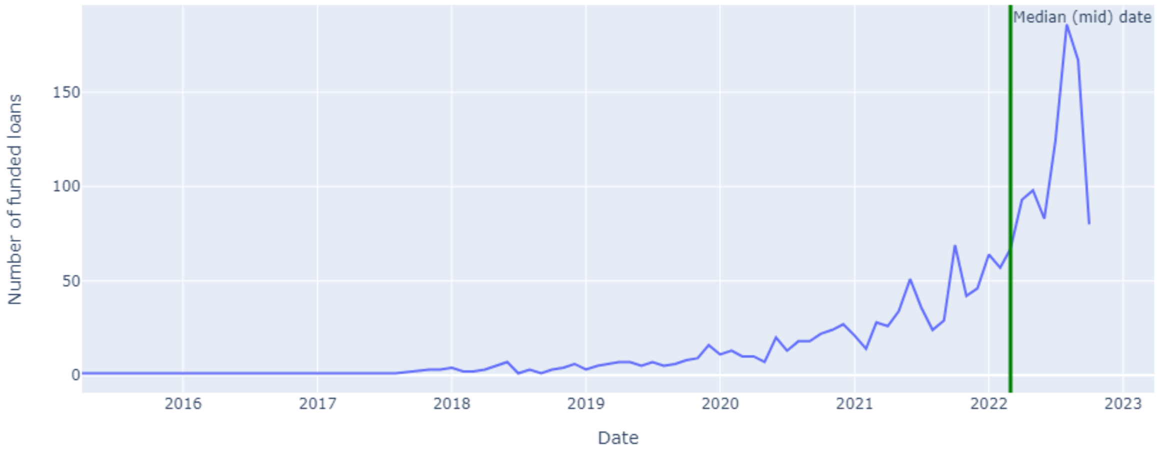


Figure 23. Number of funded loans by month

The selected median date is 1st of March 2022. Finally, after establishing the mid date and splitting the loans into two groups, the effect of Triadic Closure in this case would be measured by the probability that the two investors (m, n) would co-invest after the median date given that they already have x loans in common before the median date. The results of the Triadic Closure calculation are shown the following diagram (see Figure 24).

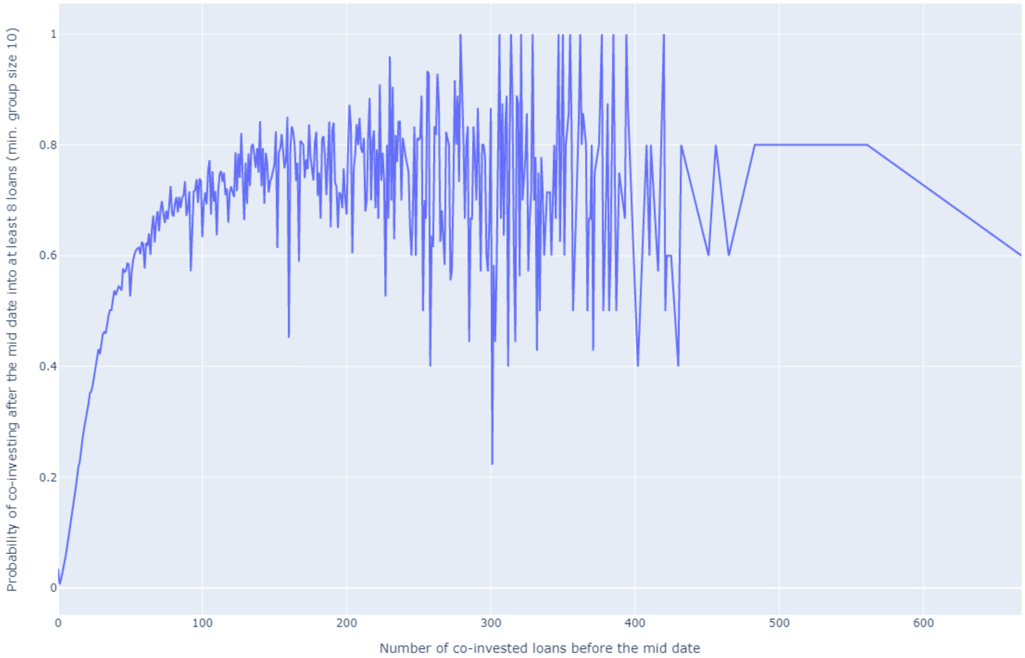


Figure 24. Probability of co-investing after the median date into at least 8 loans given that the two investors have x loans in common before the median date

Each *(x, y)* point on the blue curve in Figure 24 shows the proportion of investors (i.e., *y*) who had *x* loans in common before the median date and at least 8 loans in common after the median date. It is worth mentioning that the minimum group size is 10 investors meaning that every existing point on the right diagram of aforementioned diagram has at least that many investors who coinvested into *x* loans together before the median date. This minimum groups size threshold was introduced because small groups (e.g., of two investors) would skew the results substantially. In addition, as previously mentioned, the selected similarity threshold is at least 8 loans in common during the second half meaning that the two investors are considered unconnected after median date if they do not have that many shared loans. Finally, it is worth emphasising that the investments from the platform itself were ignored from this analysis.

The diagram in Figure 24 represents the same information as the previous diagram but focus on the subsequent diagram is a trendline which combines neighbouring data points using Locally Weighted Scatterplot function.

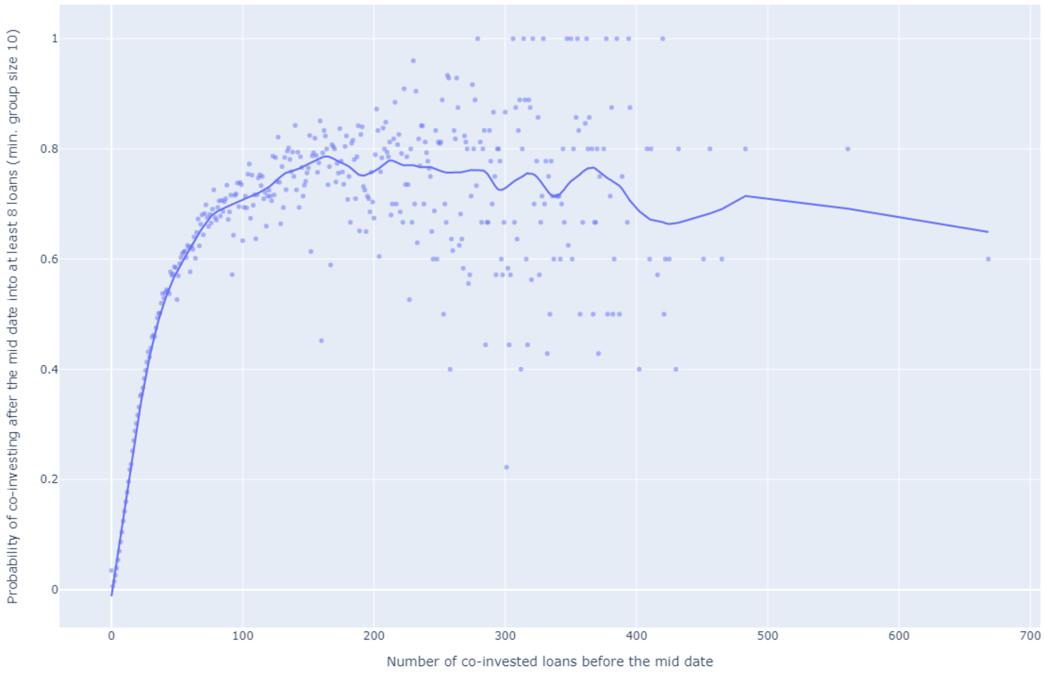


Figure 25. Trendline of the probability of co-investing after the median date into at least 8 loans given that the two investors have x loans in common before the median date

Observed Triadic closure results among the investors (see Figure 24) reveal that probability that investors would coinvest into a similar set of loans during the second half strongly correlates with the number of common investments made before the median date. To be precise, the probability of coinvesting grows almost linearly from 0% to 60% as investors coinvest from 0 to 40 loans. However, then the similarity effect reaches a plateau and probability of coinvesting into at least 8 identical loans fluctuates around 75-80%. Nevertheless, the sharp increase in the probability of coinvesting as the function of the previous common loans suggests that Triadic Closure property exists in the selected investors’ network.

However, the question remains whether the observed clustering effect among investors is not prevalent because of the randomness. To be precise, it could be the case that investors are investing randomly into all the loans, and they just happened to co-invest into a similar subset of loans together. Hence, the hypothesis that triadic closure property is present and investors flock to the same loans could be tested using null-hypothesis test that investors are coinvesting by chance. In other words, given the statistical data about the investors’ portfolios and the number of loans, what is the probability that the above charts would have been produced under the null hypothesis.

4.4.3 Null Hypothesis Testing of the Triadic Closure Property Hypothesis (H1)

In order to test the null hypothesis, it is required to denote some variables and formulas. First of all, let’s denote the number of loans *L* and the number of all investors as *I*. Hence, all the loans are split into two groups (each of size $L/2$). Then, let’s assume that an investor *i* according to the observed data invested into l_i number of loans. Assuming that investor *i* invests at random, the probability that she or he would invest in any given loan is:

$$p_i = \frac{l_i}{L}$$

Assuming that all investors invest at random and hence independently of one another, the probability that the two investors *i* and *j* would co-invest into the same loan is:

$$P\left(i \overset{COINVESTED=1}{\leftrightarrow} j\right) = p_i p_j$$

Hence, under the null hypothesis the two arbitrary investors would have *X* number of loans in common where *X* is a random variable sampled from the Binomial distribution function (denoted *B*):

$$X \sim B\left(\frac{L}{2}, p_i p_j\right)$$

Binomial distribution is applicable here because the fact that two investors would co-invest corresponds to Bernoulli trials where the outcome of the event is either a success (investors co-invest) or failure (the investors do not co-invest).

Similarly, the number of loans in common between two investors *(i,j)* after the median date is described also by the Binomial distribution:

$$Y \sim B\left(\frac{L}{2}, p_i p_j\right)$$

Hence, the probability that the two investors would have k loans in common before and after the median date can be written in the following way:

$$P(X = k, i, j) = C_k^{\frac{L}{2}} (p_i p_j)^k (1 - p_i p_j)^{L-k}$$

$$P(Y = k, i, j) = C_k^{\frac{L}{2}} (p_i p_j)^k (1 - p_i p_j)^{L-k}$$

However, the final goal is to find the probability that the two investors would co-invest into at least 8 loans given that they already have *X* loans in common according to the Bayes' Rule:

$$P(number\ of\ loans\ after\ mid\ date \geq 8\ |number\ of\ loans\ before\ mid\ date = X) =$$

$$= \frac{P(number\ of\ loans\ after\ mid\ date \geq 8, number\ of\ loans\ before\ mid\ date = X)}{P(number\ of\ loans\ before\ mid\ date = X)}$$

Unfortunately, the component in the numerator is difficult to calculate because the two events (i.e., number of loans before and after the median date) are not independent (e.g., an investor who is investing into more loans before the median date would be also more likely to more loans after the median date). Hence, another approach was taken to evaluate the null hypothesis. Using Monte-Carlo simulation methods, 200 random worlds with more than 1700 loans and 8600 investors each were generated. The probability of an investor investing in a random loan was equal to the observed proportion of loans that investor has invested into. Experimentation pipeline is presented in the diagram below (see Figure 26). Each simulation can be encoded in a matrix (A) where column *i* of row *l* stores information whether or not investor *i* invested into loan *l*. Conveniently enough the similarity matrix is obtained by multiplying this encoding matrix (A) with a transposed version of itself (A^T). Moreover, matrix multiplication is a relatively easy task for computers nowadays so the same analysis method could be used in other settings.

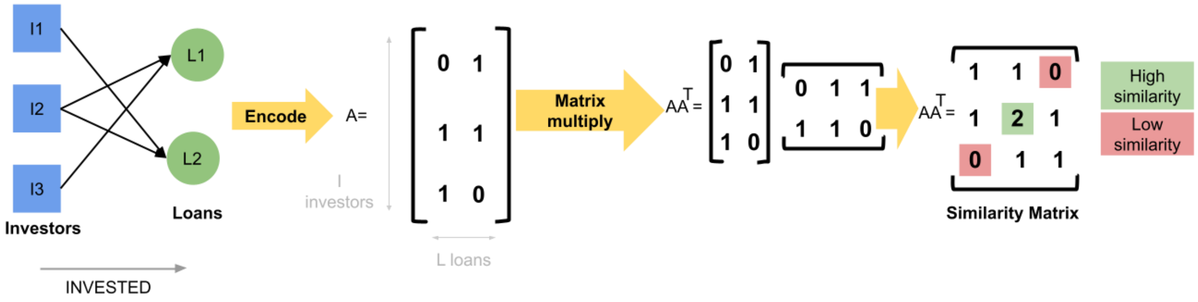


Figure 26. Efficient similarity matrix calculation mechanism to detect similarities in investment strategies between multiple investors

Finally, the results of randomly generated simulations reveal interesting results (see Figure 26). If investors were to invest at random (but preserving the number of loans each investor invest into), then the probability of co-investing would grow much quicker than the growth of the observed actual probability of co-investing (compare green and blue lines). Hence, the Triadic Closure property would also exist if investors were to invest at random. Nevertheless, as shown in Figure 26, the observed and theoretical results almost never coincide (except from one point at the beginning of X axis). Moreover, the observed trendline does not fall under the 95% confidence interval of the trendline if null hypothesis were to be true. As a result, according to the observed data investors do not invest at random, the Triadic Closure property is present among Peer-to-Peer lending investors. However, the simulated experiments demonstrate that the causality of the Triadic Closure property cannot be established – it might be the case that some investors co-invest because they invest at random while others might co-invest because they invest into similar loans.

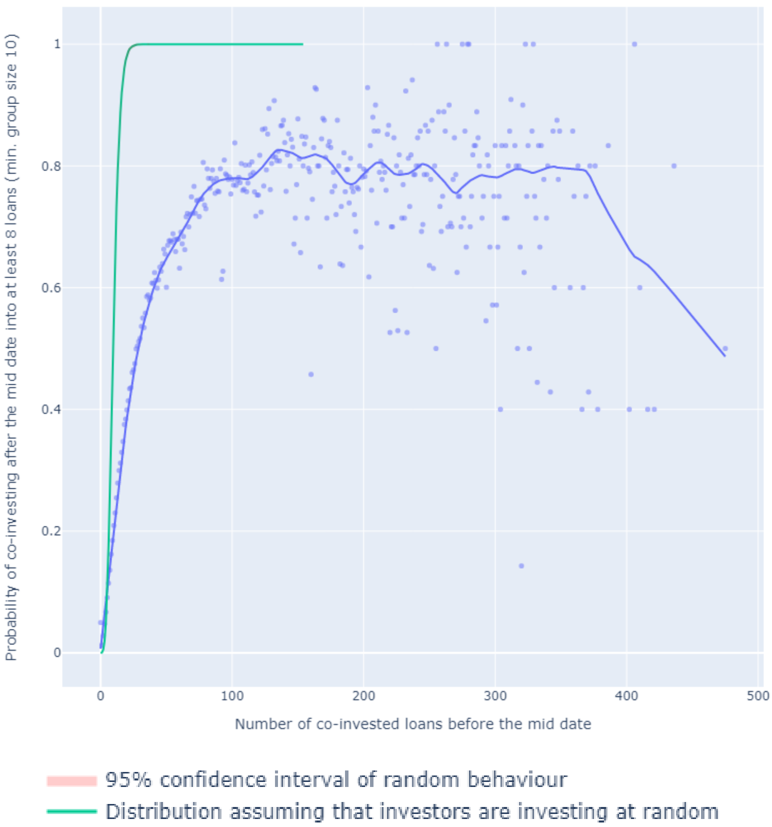


Figure 27. Probability of co-investing after the median date into at least 8 loans given that the two investors have x loans in common before the mid date compared with the null hypothesis (95% confidence interval of random investors' behaviour tightly overlaps with the co-investment probability distribution of random behaviour)

The gap between observed and theoretical trendlines demonstrates that some investors explicitly have lower probability to co-invest with some other investors. For instance, according to Figure 27, investors who had at least 100 loans in common, co-invest with about 80% of their former co-investors after the median date while if they were to invest at random, then they should have at least 8 loans in common with almost 100% probability. Hence, there is a small divergence effect observed here.

4.5 Evaluation of the Group Strategy convergence Hypothesis (H2)

In order to evaluate the last remaining hypothesis (H2), it is first required to establish clusters of investors. As presented in Literature review (see section 2.4 Clustering Techniques to Group Investors) one way to group investors according to their behaviour would be to construct a UPGMA tree. To achieve that, it is first required to encode each investor behaviour using an L-length feature vector (where *L* is the number of loans in the selected data set) (see Figure 28). Data analysis pipeline is presented in Figure 28. Here each investors' behaviour is encoded as a feature vector for his or her investment decisions; 1 would represent that an investor invested into a specific loan while 0 would indicate the opposite of that). This feature matrix would correspond to a distance matrix that would be used in the later steps. However, as the historical investor behaviour change is analysed, similarly to the previous hypothesis, the data is split into two periods (before and after the median date) (see Section 4.4.2 Statistical Evaluation of the Triadic Closure Property Hypothesis (H1)). Then a dendrogram is constructed for each period. Dendrogram is visualisation tool of an unsupervised clustering mechanism that iteratively connects the closest rows in the distance matrix (Brunton & Kutz, 2019). To reduce computational time, only the wealthiest 1000 investors were selected for the dendrogram analysis. In other words, an investor must have had a total portfolio of at least 2100 euros.

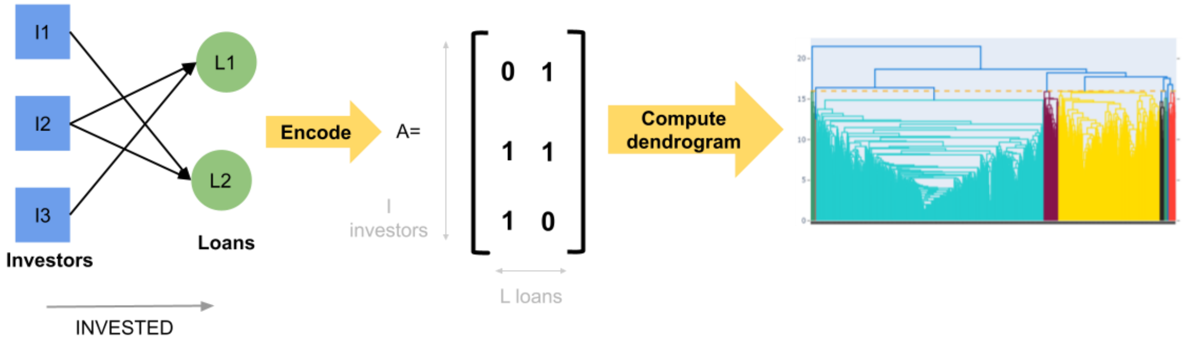


Figure 28. Data analysis pipeline to produce dendrograms

As alluded in Literature review section, one algorithm to construct clusters is called UPGMA. UPGMA algorithm defines the distance between two rows of the distance matrix as a Euclidean distance between all pairs of the elements in the two vectors (Brunton & Kutz, 2019). After grouping two similarly behaving investors, their feature vectors are combined and replaced by the average feature vector of the two feature vectors. Grouping investors using the UPGMA algorithm yields the following results (see Figure 29). It seems that there is one large mostly monolithic group of investors highlighted in green light. Reducing the required similarity threshold from 16 to 13.6 does not split the large cluster in the middle of the diagram (see Figure 30).

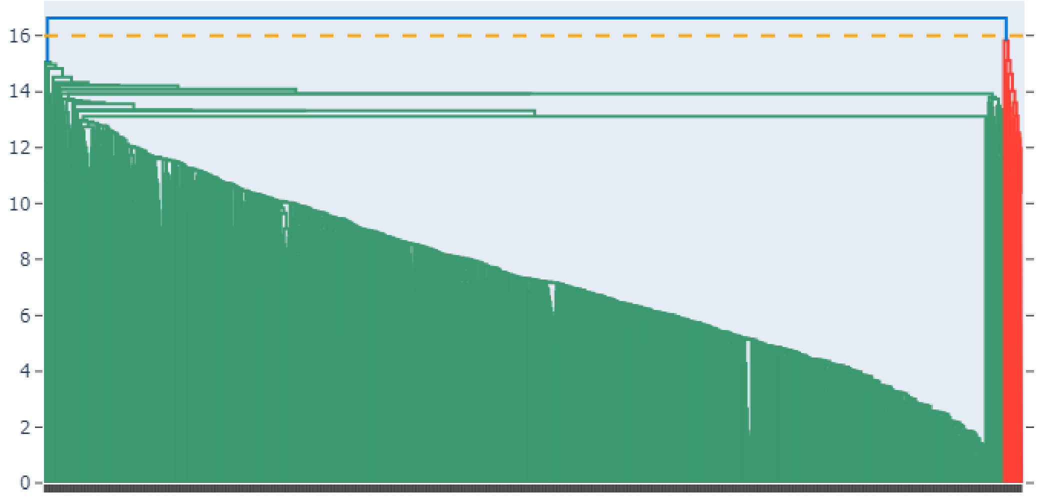


Figure 29. UPGMA-produced investors' behaviour dendrogram according to their investment decisions made before the selected median date (dendrogram similarity distance threshold being 16.0)

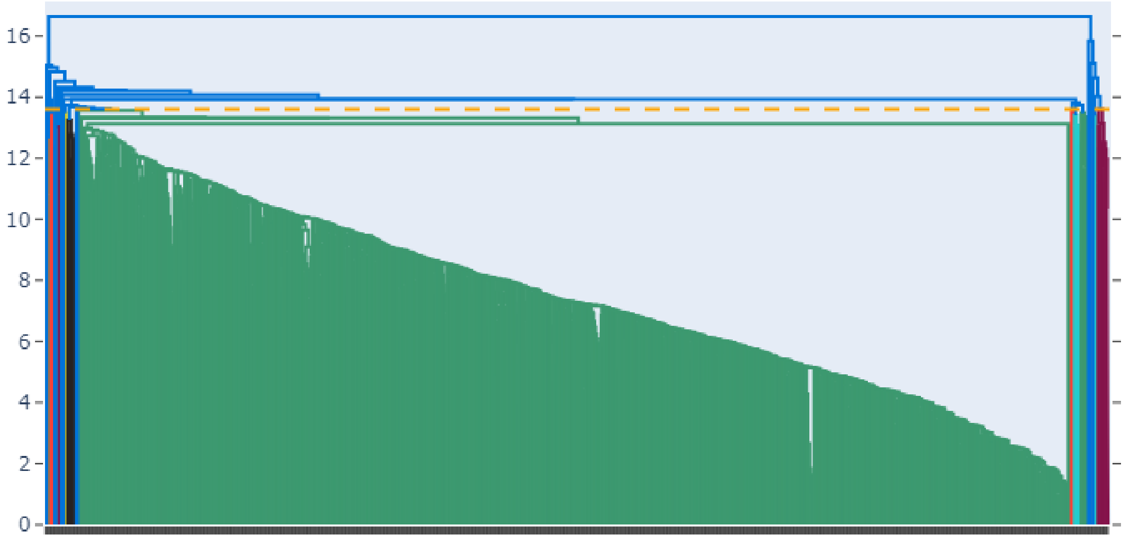


Figure 30. UPGMA-produced investors' behaviour dendrogram according to their investment decisions made before the selected median date (dendrogram similarity distance threshold being 13.6)

However, it turns out that such clustering result might have been introduced by a common so-called chaining problem where UPGMA algorithm fails to produce descriptive clusters (Lee, Chang, Carrion, & Zhang, 2017). A different technique (called *complete linkage*) produces more descriptive results and clusters of similar size (see Figure 31 and Figure 32). This method produced two large clusters (highlighted in yellow and light blue colours) with several smaller clusters. In the sorted distance matrix (see Figure 32) a pattern of similarly behaving investors also emerged (see large square in the bottom left corner). It seems that more than 50% of investors have high similarity with one another and fall into the large group category.

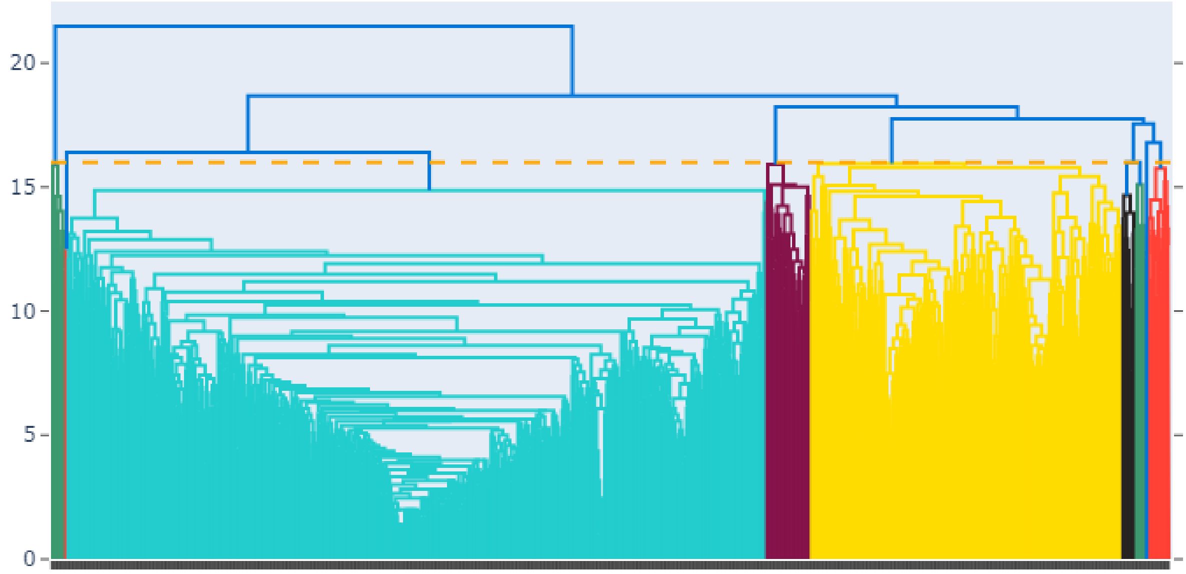


Figure 31.Complete-linkage-produced investors' behaviour dendrogram according to their investment decisions made before the selected median date

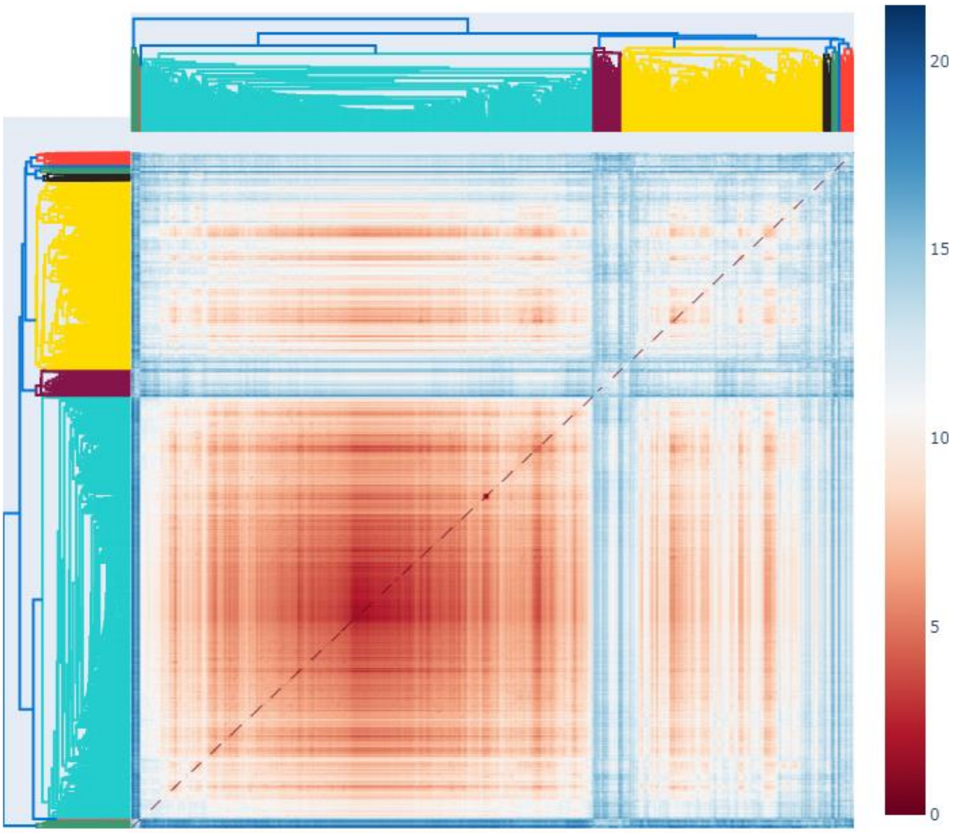


Figure 32. Complete-linkage-produced investors' behaviour dendrogram with distance matrix according to their investment decisions made before the selected median date

After the selected median date, the two main groups became more similar in size suggesting that some investors left the first initially larger cluster.

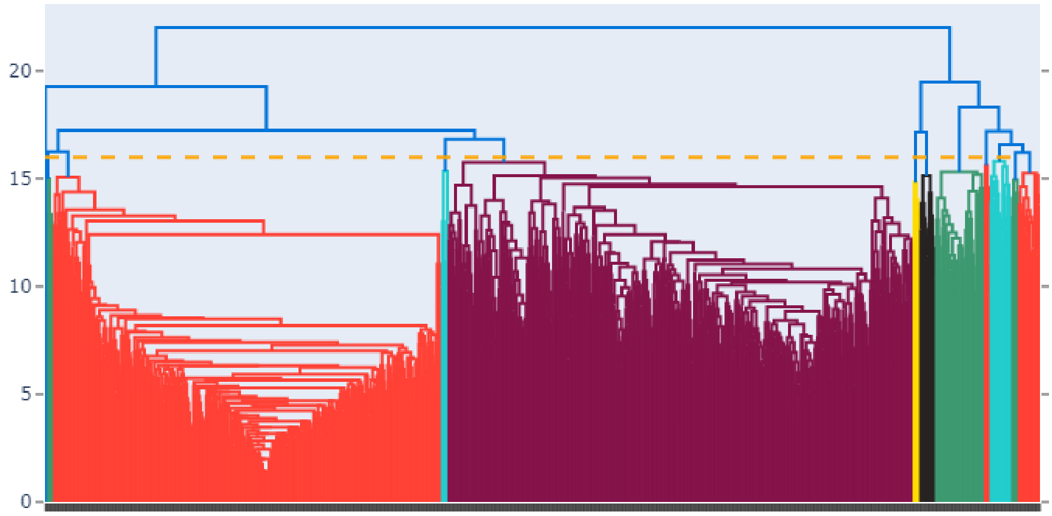


Figure 33. Complete-linkage-produced investors' behaviour dendrogram according to their investment decisions made after the selected median date

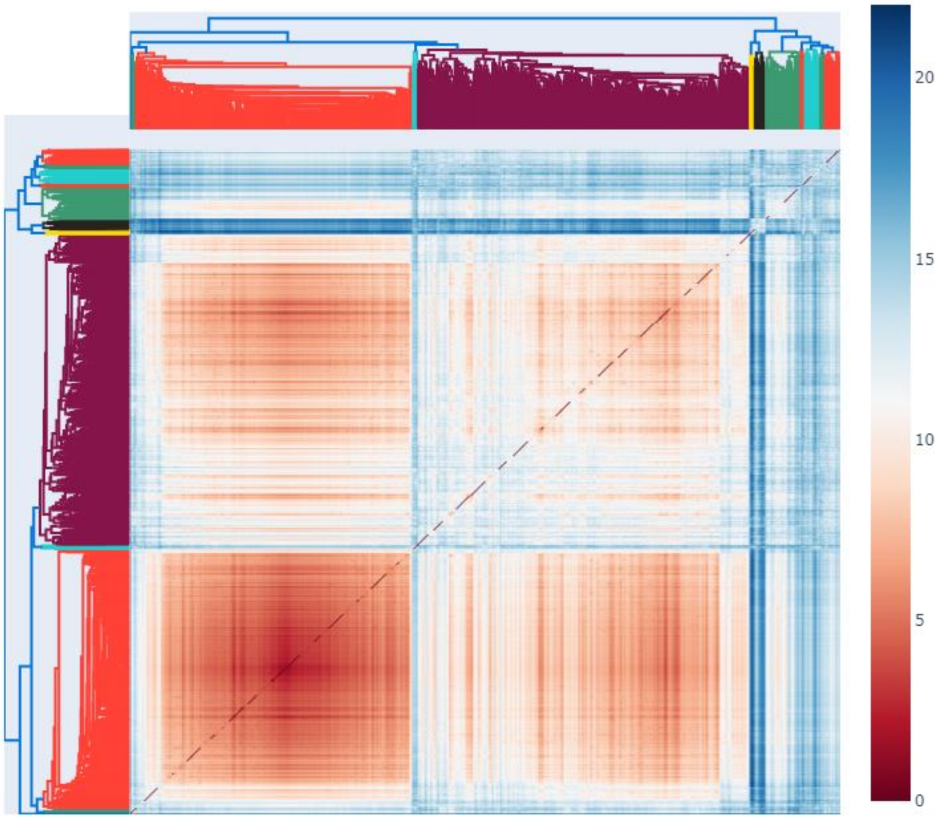


Figure 34. Complete-linkage-produced investors' behaviour dendrogram with distance matrix according to their investment decisions made after the selected median date

5. Discussion

This section provides a holistic overview of the literature review and contrasts them with research findings. Then, both theoretical and practical implications are described. Finally, the limitations of this study are discussed, and future recommendations are provided.

5.1 Discussion about the Literature Review

Previous academic work focused on the borrowers and Peer-to-Peer lending platforms but less on the investors. However, investors are equally important for the longevity and growth of the Peer-to-Peer lending ecosystem. For Peer-to-Peer lending platforms to thrive, they have to secure reliable source of funding. As this analysis demonstrates, medium and big investors form the backbone of the total money supply for the selected lending platform (see Figure 5). In order to attract new investors and keep the existing ones, lending platforms should provide analytical tools that would help individual investors to make better investment decisions. Lack of proper knowledge or incorrect loan risk assessment were among the most common reasons why individual investors tend to underperform in Peer-to-Peer lending platforms (CarloSerrano-Cincas et al., 2015, p. 1). Ready-made Peer-to-Peer lending market analysis tools would greatly reduce the barrier of entry for novice investors. Moreover, such tools could be a differentiating factor for the new lending platforms whose number has grown recently according to the Bank of Lithuania (Bank of Lithuania, 2022). Other industries, such as cryptocurrencies, do have extensive market analysis tools[9].

Moreover, the concept of the giant component has re-occurred multiple times in the observed data. At first, a relative similarity network with a 30% similarity threshold had one major component with several other smaller disconnected components (see Figure 19). Then, a giant component was observed in the UPGMA dendrogram investor similarity tree. This practical observation demonstrates that disconnected clusters of equal size rarely occur in nature. Hence, more flexible clustering algorithms are needed to identify clusters even among the connected elements (e.g., in a situation when almost all investors are connected to one another via multiple different connections and relations with one another).

Finally, the initially selected UPGMA clustering methodology produced completely different results than another clustering method called *complete linkage*. Hence, the results of the second hypothesis are inconclusive.

[9] For example, https://bscscan.com/ - a blockchain state analysis website where users can investigate the total amount of tokens each individual cryptographic wallet has

5.2 Overview of Findings

To begin with, this thesis identified that the analysed lending platform ("Savy") had an unexpectedly large portion of the entire visible investment portfolio (see Figure 5). Moreover, the distribution of credit scores among all borrowers follows a distribution function similar to the Bell curve (also referred to as Normal distribution), with the lowest credit score being almost non-existent in the secondary Peer-to-Peer lending market (see Figure 12).

As far as hypothesis testing is concerned, the steep slope in the co-investing probability diagram (see Figure 27) suggests that there is a Triadic Closure property among the investors. However, a comparison with simulated investments yields an interesting fact that if investors were to invest at random, their similarity should be even higher. This is apparent from Figure 27 because the observed tendency to converge to some loans was lower than that of the similarity between simulated investments. This suggests that there might be certain market forces that implicitly or explicitly force investors to converge less than potentially could. One such mechanism is the minimum investment size - it prohibits smaller investors from spreading their portfolios thinly among many investments. Another market restriction is loan size - in the experiment it was assumed that all willing investors would be able to invest into any given loan. However, in reality this may not always be the case. The investor might be online or might not have an investment robot turned on when the prospect loan is announced on the platform. Furthermore, in other cases investments from individual investors on Peer-to-Peer lending platforms are accepted on a first-come-first-served basis. Hence, if investors were to offer more money than there are loans in both the primary and secondary market, then market saturation would prohibit certain investors from being able to invest. However, the actual resolution method of how the platform resolves these issues when multiple investors compete for the same is not known, as it is a proprietary mechanism.

Another observation of the lower-than-random similarity distribution in Figure 27 shows that investors do not overlap at random – they explicitly tend to invest not like some other investors. Hence, this is proof that some investors flock to clusters. But as demonstrated in previous sections (see Figure 19) and previous work of other researchers (Easley & Kleinberg, 2012) the naturally occurring networks (e.g., co-investors network) tend to be largely monolithic and require specific rules to split the *giant component* into smaller sub-networks.

5.3 Theoretical Implications

The Triadic Closure evaluation diagram of the observed investors’ behaviour (see Figure 27) strongly corresponds with a similar diagram from the other study which investigated the probability that a person would join an online forum given that some of her or his friends were already members of it (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006). To be precise, the previous analysis also observed a relatively steep increase in the probability of a member connecting to an online forum provided that the number of his or her friends in that community grew from zero to around 10-15 (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006). As noted previously, subsequent common friends in the online community had diminishing impact (Easley & Kleinberg, 2012) and the probability of joining an online community reached the plateau in that study as well. This previous works did not question the causality relationship of the Triadic Closure. As the simulated experiments of the null hypothesis for H1 hypothesis reveals that some effects of the observed increase in Triadic Closure could be explained by uneven distribution of the number of loans investors invest into. Even though the aforementioned null hypothesis result diagram (see Figure 27) differs from the results of the observed data, the shape and growth of this function helps one explain some effects of the Triadic Closure. This thesis emphasises through experimental data that people (i.e., investors in this case) are highly likely to become connected if one of them has many links (i.e., invests in high proportion of loans). Visual comparison of the experimental results (see Figure 27) suggests that roughly 80% of the tendency to connect was explained via effect (see a proportion observed and null hypothesis trendlines as the number of shared loans increases in Figure 27). The remaining tendency to connect (or not to connect) could be explained by the clustering or group behaviour of the investors. Hence, if one wants to connect to some specific cluster or community, it might be better off to have more random connections and hope that at least one of them would connect to the desired community rather than have a few but specialised links.

5.4 Managerial Implications

As far as the managerial implications are concerned, this thesis provides several insights into the Peer-to-Peer lending market as well as demonstrates several interesting techniques that could help businesses to understand actual user behaviour through electronic means.

As demonstrated in Figure 6 and Figure 8, wealth distribution in the selected lending platform is largely uneven. Hence, by knowing this fact, the platform owners could have sufficient funds by keeping (or attracting) a handful of the richest investors to keep the lending platform operational. In contrast, knowing such platform owners' ability, smaller individual investors should regularly track the wealth consolidation on the platform as the platform owners could offer preferential treatment to larger investors and thus introduce uneven conditions for competition among the investors. This thesis also presents data analysis tools (see Figure 3) that an investor could use to reconstruct the investment behaviour of the selected individual investor. For example, given the uneven wealth distribution among the investors (see Figure 5), it might be interesting to investigate the investment trends of the richest investors (e.g., investors named '1' or '42920') on the platform. Filtering investments made by the investor '42920' reveals that her or his monthly investments were steadily increasing and peaked at more than thirteen thousand euros a month in the summer of 2022 (see Figure 35).

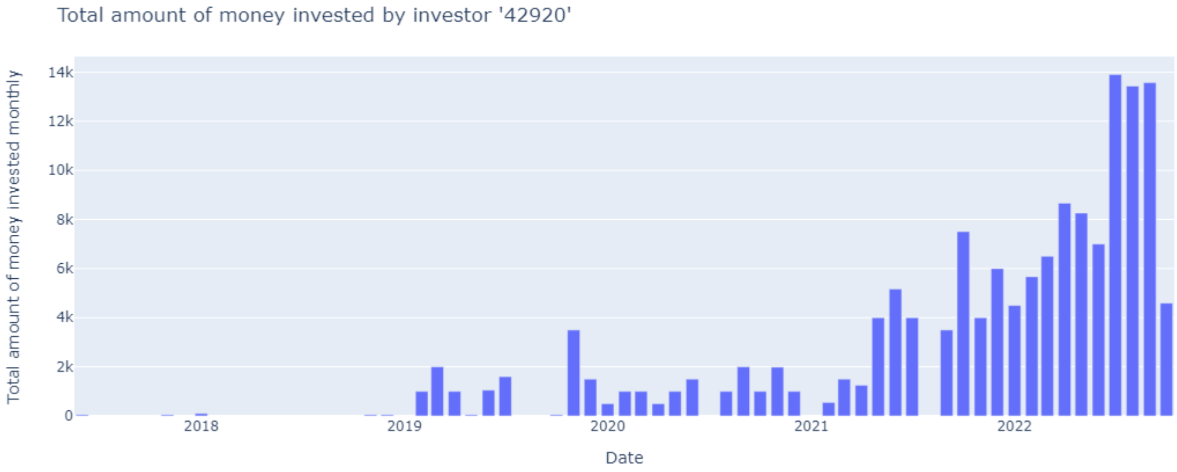


Figure 35. Amount of money investor '42920' invested every month

As demonstrated in the previous section, Triadic Closure property does exist among investors in Peer-to-Peer lending platforms. Consequently, lending platforms could utilise this information by offering the loan-recommendation system for their current individual investors. Nowadays, it is a common practice for social network platforms such as Facebook or LinkedIn to recommend their visitors' close friends the page visitors may know. Similarly, every individual investor that invests in Peer-to-Peer lending platforms could get recommendations to invest based on his or her past investment decisions and actions of close co-investors. For individual investors, such a loan recommendation mechanism would facilitate the search process while it could be a differentiating feature for the platforms themselves.

Furthermore, the thesis demonstrates an innovative way to efficiently calculate user behaviour by using Monte-Carlo simulations that are built on the limited available data (see Figure 26). One use case of the efficient user similarity calculation mechanism is that companies and individuals can de-anonymise users according to their past behaviour. For example, as demonstrated in this thesis, it was possible to reconstruct an approximated portfolio size of each investor given a limited amount of available data about loans. If one knows only a few loans that an arbitrary investor invested into, then combining this information with the derived wealth distribution data, it would be possible to tell the exact portfolio size of a targeted investor (even though he or she did not give consent to know this information). Hence, businesses and especially platforms need to ensure *k-anonymity* of the data they share about their users (Pierangela & Latanya, 1998). To be precise, sensitive data (e.g., financial data in this case) should be hidden in the group, meaning that at least *k* investors need to have the same specific property (e.g., level of wealth) so that it would be impossible to uniquely identify a specific individual (Pierangela & Latanya, 1998). This could be achieved in the analysed lending platform by two means. One option would be to completely remove information about the fellow co-investors on the secondary market, but this could have business implications as individual investors would not have the mechanisms to see if the platform is popular among other investors by inspecting co-investors' lists in loan descriptions. Another and the more preferred option would be to categorise investors into more coarse groups. To be precise, instead of showing that an investor *x* invested *y* amount of euros into the loan *z*, it could be shown that this investor invested an amount of money that falls inside some range (e.g., between 100 and 200 euros into one loan). Increasing the so-called bin size would ensure that more investors would fall into the category, and investor de-anonymisation would be less difficult.

Efficient user similarity matrix calculation could be used in many other fields. For instance, the data analysis pipeline presented in Figure 26 could also be used to process data that is stored on the Blockchain. For example, a single person can create multiple cryptographic wallets to store their cryptocurrency and the generation of the new wallet does not come at any substantial cost both in terms of time and money. Hence, if someone would like to match which similarly behaving cryptographic wallets correspond to the same owner, one could achieve that by comparing different behavioural patterns of different cryptographic wallet addresses and using the similarity analysis technique presented in Figure 26.

5.5 Limitations of the Study

As far as the limitations of this study are concerned, it is worth emphasising that a lot of data from the secondary loan market of the selected lending platform could be invisible. This is due to the fact that loans that were not being resold on the secondary market were not captured by the data collection program as they are not visible to individual investors either. However, a large number of co-investors per loan (which fluctuates around a few hundred per loan) suggests that such cases when the loan is not visible because it is not being resold should be minimal as there is a high chance that at least one co-investor would be reselling that loan at any given time. Nevertheless, the data might be skewed towards the riskier loans because it could be the case that investors do not tend to resell those loans that are being repaid on time. Furthermore, other investors might be more willing to quickly rebuy those loans that are being repaid on time. However, the sample size (around 1700 loans with more than 8000 investors) should have mitigated such anomalies at least to some extent.

Another pitfall of this study is that conclusions were derived from a single Peer-to-Peer lending platform, as only one lending platform from Lithuania approved the legal request to analyse their historical data. Different platforms might have different loan parameters meaning that the results of different lending platforms, even from the same country, could yield different or incomparable results. Hence, a future study of other lending platforms could compare and contrast the findings of this study.

Moreover, the proprietary credit ranking algorithm of the selected platform is not known to the public. Consequently, it is unclear how different borrowers' parameters determine her or his assigned borrower's grade. Moreover, it is unclear whether or not the credit score rating formula has changed over the years.

Finally, the null hypothesis testing of the first hypothesis (H1) could be slightly wrong as it made a simplifying assumption that all investors started investing at the same time. Further investor cohort analysis would be required to determine the impact of this oversimplification. In addition, the contrasting results of the null hypothesis, the first hypothesis also demonstrates the limitations of the simulated experiments that aim to model real-life behaviour.

5.6 Future Research Recommendations

Future research projects could apply ideas presented in this thesis to a wider or deeper range of situations. The first recommendation would be to compare user behaviour on multiple lending platforms. This is because different lending platforms have different lifespans, which would result in more senior platforms potentially having more defaulted loans. Moreover, different lending platforms apply different fees (e.g., money transfer fees outside the platform), which in turn could result in slightly different investor dynamics. Moreover, one could expand the geography of the analysis. This extension would allow the researcher to compare the effect of legal basis to the behaviour of individual investors in the Peer-to-Peer lending platforms. Another improvement to this study would be to take several secondary market snapshots throughout the time. Not only would this allow to capture more data, but the additional data would also serve the purpose of allowing the researchers to investigate loan velocity (i.e., the speed at which different loans are being resold on the secondary market).

To get more reliable results, one should collaborate with the lending platforms to acquire information about the loans that are not being traded at the moment in the secondary Peer-to-Peer market. Furthermore, more granular information about the loan, such as loan performance notifications, would help the researchers to find even more interesting insights. For instance, another Lithuanian lending platform "Paskolų Klubas" has notifications about the borrower meaning that this textual data could be incorporated into the analysis. For example, due loan payments are reflected as notifications in the loan history. Moreover, "Paskolų Klubas" also provides the expected and actual loan payback schedules. This would allow one to evaluate the probability of a borrower paying back the loan. Such information would come in handy when differentiating between different clusters of investors. For instance, one might observe that a specific cluster might share common investment patterns that, in turn, yield a detrimental outcome for the investment return.

6. Conclusions

6.1 Conclusions of Literature Review

In conclusion, the Peer-to-Peer lending market is relatively new. Nevertheless, several interesting studies have been done that analyse different aspects of borrowers and lenders. Previous studies identified the existence of herding behaviour among investors. In some cases, herding behaviour was beneficial, but it reduced the effective portfolio return rate in other situations. Moreover, among some investors in other countries, discriminative behaviour was apparent when issuing loans to new borrowers. Nevertheless, the Literature Review section identifies that there is a research gap in the studies that quantitatively measure the effect of group behaviour. Hence, the Triadic Closure property is selected as an appropriate data analysis construct in Social Sciences that can be used to measure the tendency for the group to converge.

6.2 Conclusions of Research Methodology

In order to reproduce the steps, one would need to collect secondary market data of the selected Peer-to-Peer lending platform. The recommended option that is presented in this study uses Python programming language to collect the data using the Selenium Web-Scrapping library. For investors' network analysis, one might find Neo4j analysis tools (e.g., *Neo4j Desktop*[10], *Neo4j Bloom*[11]) particularly useful.

6.3 Conclusions of Empirical Research Results

An interesting fact arose that the selected lending platform controls a significant proportion of all the money that is being issued to the borrowers. Uneven wealth distribution among the investors in a Peer-to-Peer lending platform might raise some tensions in the way the platform is governed as a relatively small proportion of total capital is funded by a small proportion of investors.

Triadic closure property does exist in the selected co-investors' network. However, the causality of this phenomenon could not be established. This indecisiveness comes from the fact that randomly simulated investor behaviour produced similar yet different Triadic Closure results. Similarly, the interpretation of the results of the second hypothesis depends

[10] https://neo4j.com/download/

[11] https://neo4j.com/product/bloom/

on the selection of the clustering algorithm. While the UPGMA clustering method produced one monolithic cluster, the complete-linkage method produced more clusters of similar size. Assuming that the complete-linkage method was more appropriate in this situation, one of the largest identified clusters shrank in size, suggesting that some investors might have switched their investment strategy over the period of one year.

6.4 Conclusions of Discussion

Future research projects could compare the obtained results among different lending platforms. Moreover, there are several recommendations that would increase the reliability and generalisability of these results. For instance, one could acquire Peer-to-Peer lending data from the platforms themselves. However, not all individual investors have the required knowledge, technical skills, or time to perform competitor analysis on a daily basis. Hence, improved tooling should be provided to individual investors in Peer-to-Peer lending platforms. Furthermore, this thesis proposes an additional legal framework that would help to protect the privacy of individual investors. If recommendations for governance of Peer-to-Peer lending platforms were ignored, then individual investors could conduct similar competitor analysis using the techniques presented in this thesis. Unfortunately, the analysis of shifting group-internal strategies greatly depends on the selected clustering method. Hence, future research work is required to compare the results obtained by different clustering techniques. Finally, this thesis demonstrates a real-life application of a similarity detection algorithm applied in practice. This unsupervised clustering technique can be applied both in different Peer-to-Peer lending platforms as well as in different settings (such as Blockchain).

Appendices

Appendix A. List of parameters in the collected data sample

Feature name	Comment
loan__number	
loan__number_of_completed_instalments	
loan__number_of_all_instalments	
loan__is_loan_paid_on_time	
loan__announcement_date_primary_market	
loan__annual_interest_rate_percent	
loan__more_info_link	
loan__purpose	
loan__borrowing_amount	
loan__borrowing_duration	
loan__loan_status	
loan__SAVY_credit_score	Issued by SAVY itself
loan__risk_of_default	Issued by SAVY itself
loan__payment_day	
loan__list_of_coinvestors	
borrower__age	
borrower__gender	
borrower__city	
borrower__main_source_of_income	
borrower__duration_of_stable_income	
borrower__employment_area	
borrower__employment_duration	
borrower__education	
borrower__city_of_residence	
borrower__family_status	
borrower__number_of_dependent_people	
borrower__owned_property	
borrower__unpaid_loans	
borrower__last_registered_paid_laon	
borrower__income	
borrower__family_income	
borrower__disposable_income	
borrower__liabilities_to_income_ratio	
borrower__liabilities_to_other_institutions	
investment__announcement_date_secondary_market	
investment__invested_amount	
investment__remaining_total_amount_of_money_to_be_paid	
investment__remaining_principal_to_be_paid	
investment__remaining_interests_to_be_paid	
investment__price_in_secondary_market	

Appendix B. Distribution of Loan Durations

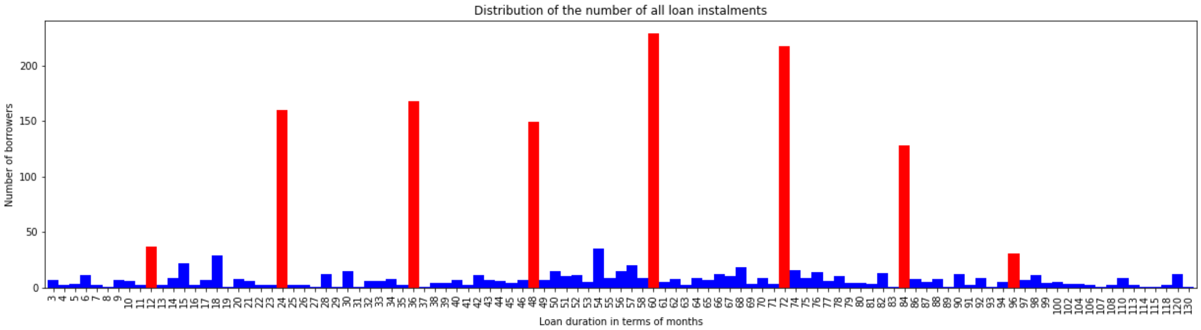


Figure 36. Distribution of loan durations

References

Angerer, M., Brem, A., Kraus, S., & Peter, A. (2017). Start-up Funding via Equity Crowdfunding in Germany - A Qualitative Analysis of Success Factors. *The Journal of Entrepreneurial Finance, 19*(1), 1. Retrieved from https://vb.ism.lt/permalink/f/18athsg/TN_cdi_proquest_journals_1871398050

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group Formation in Large Social Networks: Membership, Growth, and Evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 44–54). Philadelphia, PA, USA: Association for Computing Machinery.

Balyuk, T., & Davydenko, S. (2019). Reintermediation in FinTech: Evidence from Online Lending. *Michael J. Brennan Irish Finance Working Paper Series*.

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, 797-817.

Bank of Lithuania. (2020, September 10). *Lietuvos bankas.* Retrieved from Tarpusavio skolinimo platformos operatorių veiklos apžvalga: https://www.lb.lt/lt/leidiniai/tarpusavio-skolinimo-platformos-operatoriu-veiklos-apzvalga

Bank of Lithuania. (2022, September 11). *Financial market participants*. Retrieved from https://www.lb.lt/en/sfi-financial-market-participants?type=23&market=1&ordering=decisions_count.desc

Bank of Lithuania. (2022, April 1). *Lietuvos bankas*. Retrieved from Naujausi ekonominiai rodikliai: https://www.lb.lt/lt/eap-naujausi-ekonominiai-rodikliai

Brunton, S., & Kutz, J. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control.* Cambridge: Cambridge University Press.

CarloSerrano-Cincas, C., Gutiérrez-Niet, B., Serrano-Cinca, Carlos Gutiérrez-Nieto, Begoña López-Palacios, L., & Serrano-Cinca, Carlos Gutiérrez-Nieto, Begoña López-Palacios, L. (2015). Determinants of Default in P2P Lending. *PLoS One, 10*(10), e0139427-e0139427. Retrieved from https://doi.org/10.1371/journal.pone.0139427

Chen, D., Lai, F., & Lin, Z. (2014). A trust model for online peer-to-peer lending: a lender's perspective. *Information technology and management*, 239-254.

Chen, X., Hu, X., & Ben, S. (2021). How Individual Investors React to Negative Events in the FinTech Era? Evidence from China's Peer-to-Peer Lending. *Journal of Theoretical and Applied Electronic Commerce Research*, 52-70.

Chen, X.-h., Jin, F.-j., Zhang, Q., & Yang, L. (2016). Are investors rational or perceptual in P2P lending? *Information Systems and e-Business Management*, 1617-9854.

Compeau, P., & Pevzner, P. (2014). Bioinformatics Algorithms: An Active Learning Approach. In P. Compeau, & P. Pevzner, *Bioinformatics Algorithms: An Active Learning Approach* (pp. 25-27,68-118). Active Learning Publishers.

Cummins, M., Lynn, T., Mac an Bhaird, C., & Rosati, P. (2019). Addressing Information Asymmetries in Online Peer-to-Peer Lending. In M. Cummins, T. Lynn, C. Mac an Bhaird, & P. Rosati, *Disrupting Finance: FinTech and Strategy in the 21st Century* (pp. 15--31). Springer International Publishing.

Easley, D., & Kleinberg, J. (2012). *Networks, Crowds, and Markets - Reasoning About a Highly Connected World.* Cambridge University Press.

Golovkina, A. (2022). *Factors affecting consumer loans interest rate in peer-to-peer lending platforms in Lithuania.* Vilnius: ISM Vadybos ir ekonomikos universitetas. Prieiga per eLABa – nacionalinė Lietuvos akademinė elektroninė bibliotek.

Haewon, Y., Byungtae, L., & Myungsin, C. (2012). From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. *Electronic commerce research and applications*, 469-483.

Kreditai.info. (2019, July). *Tarpusavio skolinimosi platformos (P2P) Lietuvoje ir Europoje*. Retrieved from Kreditai.info: https://kreditai.info/straipsniai/663-tarpusavio-skolinimosi-platformos-p2p-europoje

Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic commerce research and applications*, 495-503.

Lee, M., Chang, H., Carrion, C., & Zhang, L. (2017). An approach to grouping traffic signals for coordination using clustering methods., (pp. 792-797).

Maudos, J., & Fernandez de Guevara, J. (2004). Factors explaining the interest margin in the banking sectors of the European Union. *Journal of Banking & Finance*, 2259-2281.

Mild, A., Waitz, M., & Wöckl, J. (2015). How low can you go? — Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of business research*, 1291-1305. Retrieved from https://doi.org/10.1016/j.jbusres.2014.11.021

Mingfeng , L., Nagpurnanand , R., & Siva , V. (2013). Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *Management Science*, 17-35.

Pierangela, S., & Latanya, S. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Pope, D., & Sydnor, J. (2011). What's in a Picture? Evidence of Discrimination from Prosper.com. *The Journal of human resources*, 53-92.

Ravina, E. (2019, February 15). Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets. Retrieved from http://dx.doi.org/10.2139/ssrn.1107307

Sadzius, T., & Sadzius, L. (2018). Crowd funding Regulation in the Baltic Region. *International Journal of Business and Social Science*, 99-110.

Slavin, B. (2007, June 21). *Peer-to-peer lending – An Industry Insight.* Retrieved April 21, 2022, from www.bradslavin.com: https://www.bradslavin.com/wp-content/uploads/2007/06/peer-to-peer-lending.pdf

Stanislovaitis, V., Alešiūnas, A., Baltagalvis, A., Gražys, R., Jakienė, S., Jakštonis, J., . . . Šekštelis, A. (2018). *Įgyvendink idėją per kickstarter : sutelktinio finansavimo vadovas.* Inovitas.

State Tax Inspectorate Under the Ministry of Finance of the Republic of Lithuania. (2021). 2021 M. GYVENTOJŲ GAUTŲ FINANSINIŲ PRIEMONIŲ PARDAVIMO AR KITOKIO PERLEIDIMO NUOSAVYBĖN IR IŠVESTINIŲ FINANSINIŲ PRIEMONIŲ REALIZAVIMO PAJAMŲ APMOKESTINIMO IR DEKLARAVIMO YPATUMAI. Retrieved 04 03, 2022, from https://www.vmi.lt/: https://www.vmi.lt/evmi/documents/20142/737112/VMI--.pdf

Statista. (2022, February). *Annual development of the S&P 500 index from 1986 to 2021*. Retrieved from Statista.com: https://www.statista.com/statistics/261713/changes-of-the-sundp-500-during-the-us-election-years-since-1928/

Varnaitė, I. (2018). Identify the success factors of the initial coin offerings from the investors perspective. Retrieved from Access via eLABa – Lithuanian national electronic academic library: https://vb.ism.lt/permalink/f/lu96of/ELABAETD26895419

Verslo žinios. (2022, January). *Rekvizitai.lt*. Retrieved from Bendras finansavimas. UAB SAVY: https://rekvizitai.vz.lt/imone/bendras_finansavimas/

Wang, H., & Greiner, M. (2010). Herding in Multi-winner Auctions. *ICIS 2010 Proceedings*, (p. 235).

Wang, X., & Wang, L. (2019). Investment Intention Towards Online Peer-to-Peer Platform: A Data Mining Approach Based on Perceived Value Theory. In *Parallel and Distributed Computing, Applications and Technologies* (pp. 90-99). Singapore: Springer Singapore.

Zhang, J., & Liu, P. (2012). Rational Herding in Microloan Markets. *Management science,* 892-912.