Social Network Analysis - Assignment 1

Group 7

Floris Vermeulen Martijn van Iterson Niek Fleerakkers Patryk Grodek
Samir Sabitli

2025-09-12

Table of contents

1 Introduction

With the advent of online peer-to-peer (P2P) lending platforms, the traditional methods of financial intermediation have been usurped by individual choice. The surging popularity and accessibility of services such as Prosper.com has grown to become one of the most important avenues through which individuals can secure micro-loans (Cai, Lin, Xu, & Fu, 2016). Arguably, while the democratisation of credit alleviates numerous pre-existing issues, it serves to exaggerate well-known informational asymmetries associated with determining creditworthiness (Mingfeng Lin, N.R. Prabhala, & Siva Viswanathan, 2009). Per, Chen, Zhou, & Wan (2016), from a traditional financial lens, these asymmetries can contribute to moral hazard and adverse selection, ultimately causing systemic financial losses. This arises from the fact the users are anonymous, challenging the veracity of their information. Further, the unsecured nature of credit makes its collection relatively ineffective. Jointly, these factors characterise the lemons market theory by Akerlof (1978).

However, social networks can act a crucial avenue through which information transfer could be facilitated, alleviate some of the existing asymmetries (Mingfeng Lin et al., 2009). There are several prominent theories and research that attempt to bridge this gap. First, the theory of social capital is particularly salient because it relates directly to the reputation and trustworthiness of the users (Adler & Kwon, 2002). While its definition is not strictly defined, according to Putnam (2015), it can be seen as a means through which social organisation facilitate coordination and trust. Nahapiet & Ghoshal (1998) define three dimensions, including structural, relational, and cognitive. We primarily study the structural component, whereby the network properties relating to tie formation between borrowers is considered. Following this perspective, Chen et al. (2016) argue that social capital is earned through connections to others. However, per the authors, the internet facilitates weak forms of these ties, diminishing the benefits of social capital on P2P platforms, especially for non-friendship networks. Notably, they cite a low interdependence and closed structures to this effect.

With this, we arrive at our first research problem, aiming to understand **to what extent are borrower networks weak in P2P networks?** More specifically, we formulate the hypothesis

$\mathbf{H}_1^a$	Borrower networks based on similarity have a low degree of centralisation

To understand social capital, we employ techniques related to the occurance of dyadic connections rather than studying their structural components. As such, this problem can be studied with the use of the Quadratic Assignment Procedure (QAP) and Conditional Uniform Graph (CUG) tests, whereby the theoretical distributions of network characteristics can be tested.

Our second study attempts to model social capital as a statistical model, specifically considering the structural dimension to social capital. According to the comprehensive review of Bachmann et al. (2011), past studies have considered financial, demographic, and soft informational factors in the financial outcomes of P2P lending, however, these perspectives have not been considered with relation to each other. With respect to this premise, we posit **what is the role of structural and non-structural factors in the formation of social capital?**. This problem is arguably best studied with the use of Exponential Random Graph Models (ERGM) as they are considered to work well in identifying a combination of structural and exogenous patterns in the context of a variety of network specifications (Jackson, 2011).

In studying demographic factors, Pope & Sydnor (2008) suggest that age plays a significant role in funding success. In particular, compared to individuals aged 35-60, those aged 35 or younger have a 40-90 basis point higher success rate. Additionally, those aged 60 and above are 1.1-2.3 percentage points less successful. While this is not a direct relationship with social capital, it may suggest that these individuals may be placed lower in the P2P market, per Putnam (2015).

In studying gender, Barasinska & Schaefer (2010) find that female lenders are less risk averse than male lenders, funding credit with lower interest rates at a higher frequency. In conjunction, Pope & Sydnor

(2008) find that single women pay 0.4% less interest than men. These factors may indicate that edges are more likely to form between borrowers if they are women.

The structural dynamics of the network could be explained by the theory of Relational Herding, as formulated by De Liu, Brass, Lu, & Chen (2015). The authors explain that in the face of uncertainty, actors tend to exhibit a clustering behaviour, even putting aside their own private information. Though, other social dynamics may also explain the phenomenon. Devenow & Welch (1996) posit that herding may also occur if individuals blindly follow others without rational analysis. In the P2P context, Herzenstein, Dholakia, & Andrews (2011) show evidence of strategic herding when borrowers observe that a loan has gotten funding in the past.

Further, Podolny (1993) explains that the ways in which information flows in social networks can explain its outcomes. In particular, if the reputation or status of an individual can be observed, they are said to be "pipes" for other actors to transact with them. However, if this reputation is merely *perceived*, then it said to be a "prism". In our context, prism seems more appropriate as users usually cannot directly observe prestige-related characteristics, however, this can be tested by constructing a variable related to latent status.

Hypothesis	Type	Dyad	ERGM Term	Motivation
$\mathbf{H}_2^a$: Younger borrowers tend to form more relationships with each other	E xogenous	Independent	`nodecov()`	The `nodecov()` term captures how the age covariate affects the likelihood of edge formation. Since we want a numeric range, this term is most appropriate.
$\mathbf{H}_3^a$: Women tend to form more relationships with other borrowers	E xogenous	Independent	`nodefactor()`	The `nodefactor()` term captures the tendency of women to form more edges, regardless of the other gender.
$\mathbf{H}_4^a$: Borrowers tend to exhibit herding around each other	En dogenous	Int erdependent	`gwesp()`	The `gwesp()` term captures the tendency for triadic relationships to form. In other words, for ties to form due to the presence of other ties in the dyad's neighbourhood.
$\mathbf{H}_5^a$: Borrowers tend to form ties around other high-status borrowers	En dogenous	Int erdependent	`gwd()` and `nodecov()`	The `gwd()` term describe degree distribution, indicating whether a borrower is popular to the degree specified. It does not necessarily relate to community. It is necessary to also use an indicator of prestige to understand whether pipe or prism dynamics may be more prevalent.

Our research makes notable contributions to existing research on social network analyses on P2P networks. The study primarily expands upon the existing ideas of social capital by analysing it with respect to European P2P lending market and evaluating the interactions between structural and non-structural factors such as demographics. Typically, research tends to consider them independently and utilises predictive, rather than purely statistical models.

Following this, the report will outline the methodology employed. In this section, we will discuss the dataset utilised, specific data processing steps and other considerations relating to the empirical or network environment. Further, we justify the empirical structure of the analysis, evaluating the specific methodologies utilised with respect to the wider quantitative landscape. Subsequently, the results are outlined for each model and hypothesis, alongside their interpretations for our research problem. Finally, we conclude the report by summarising the core research problem, our empirical set-up alongside considerations for future research.

2 Methodology

Within this section, we construct the empirical layout of our research problem, including the dataset and data processing steps.

2.1 Dataset

The study utilises a publicly-retrieved dataset from a leading European P2P platform called Bondora. The dataset contains detailed information on both defaulted and non-defaulted loans given to users between February 2009 and July 2021. Prior to any manipulation, the dataset contains a range of numeric, binary, categorical, and time-series attributes across 85,087 unique users and 179,235 individual loans. The data was originally collected by Siddhartha (n.d.) in 2021 and published on kaggle.com. While the user published the data, they simply downloaded it from a now-defunct page on Bondora's website. Public data is no longer offered by Bondora, access to newer data is not possible.

2.2 Data Processing and Network Construction

To make the data usable for our research, it was processed. The full steps can be seen in the Appendix. First, only attributes relevant to the research were kept for resource efficiency and ease of use. Following this, any rows with missing values were entirely removed to preserve a complete dataset. This step did not remove a significant number of data points. Following this, it is mandatory to reduce the size of the overall dataset to ensure that any analyses conducted can converge and do so in a timely manner. We did this by randomly removing data points until the dataset had 500 remaining observations. While a sampling bias is technically possible as a result, the pseudo-random data reduction should minimize this effect.

To construct the network, we need to specify edges between users. While there are many ways to do this, we prefer an approach that does not introduce unnecessary complexity to models' interpretations and maintains a meaningful semantic relationship between borrowers. Ultimately, we create edges based on how similar borrowers are across several dimensions, computed by cosine similarity. The intuition here arises from the fact that certain behaviours or social processes *groups* borrowers together. This technique has the advantage of discounting vector size for its angle, which reduces the influence of outliers in defining how similar individuals are. However, to avoid a fully connected network that would make ERGM completely redundant, we establish a threshold of `cosine_sim=0.5` to determine whether an edge can exist. This enables us to control how similar individuals must be for our analysis while maintaining a reasonable density.

Defining the dimensions that make an individual similar to another is a difficult task, however, we opted for attributes describing users' loans rather than their demographic or financial characteristics. This is *crucial* because we must isolate these variables from any outcome being studied. Otherwise, our analysis will be severely biased by data leakage and possibly even simultaneous equation bias. Further, we chose to standardise the attributes chosen for cosine similarity to avoid bias introduced by differing variable scales. Ultimately, the edges formed are weighted [1] by how similar individuals are to each

[1] We understand that weighted edges make ERGMs significantly more difficult. We can definitely make them binary for further analysis. We believe that making them weighted can make the analysis more meaningful because it would allow

other. The attributes excluded from this similarity measure were later added to the `igraph` network object as vertex attributes.

2.3 Descriptive Statistics and Network Overview

Having processed the dataset and created a weighted network of similar borrowers, we provide an overview of the descriptive statistics relating to the network and wider dataset. The figure below shows an overview of the network, however, since we rely on sampling for the nodes, the network is not equivalent to the underlying population. Observing the figure, we see that there are indeed prevalent communities, however, there is also significant overlap between them.

Figure 1: Plot of the Bondera Network with and without Community Detection

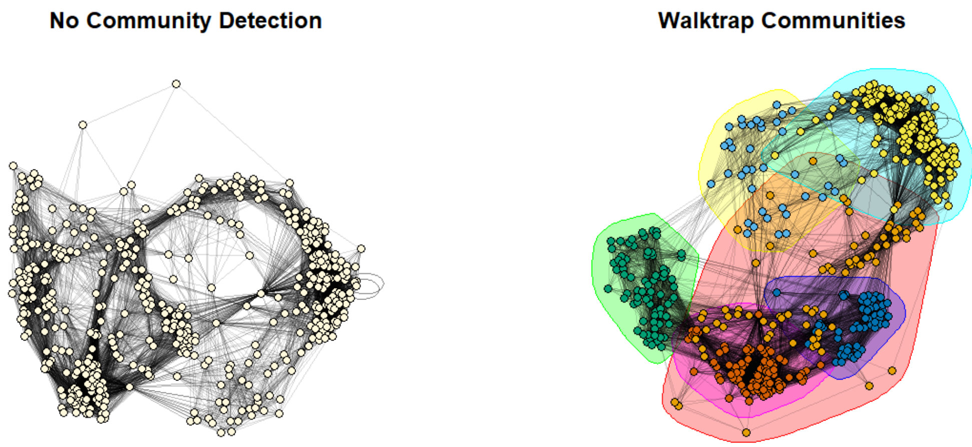


Table 3: Network Descriptive Statistics

Network Measure	Value
Number of Vertices	495
Number of Edges	14,017
Density	0.109
Reciprocity	1 [2]
Transitivity	0.692
Mean Distance	2.112
Dyad Census	Mutual: 13269
	Null: 108,996

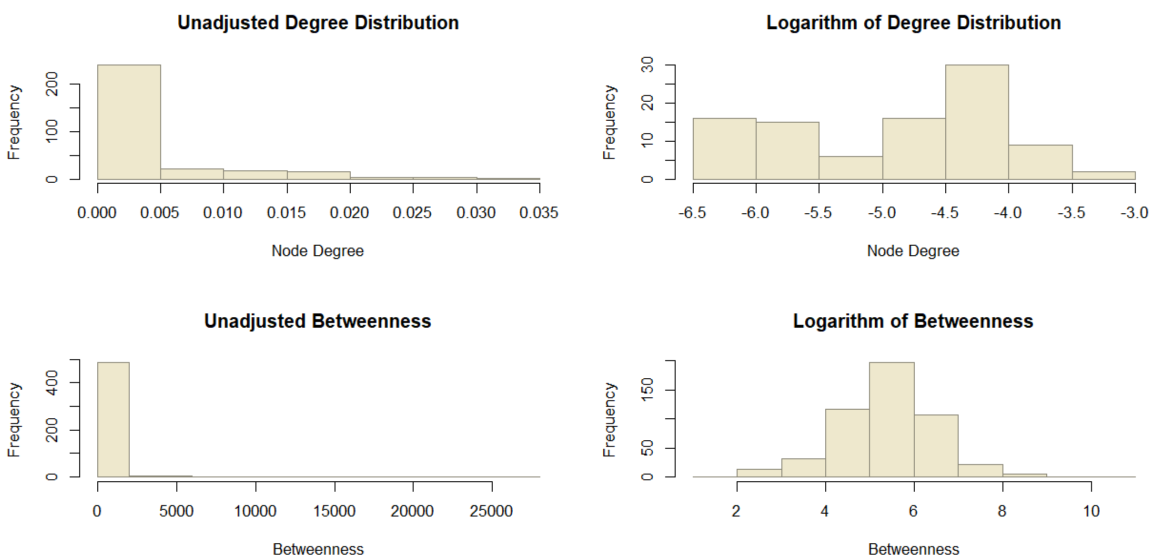
Observing the figure below, we see low betweenness centrality and degree distribution without any logarithmic adjustments. The latter indicates that most nodes have very low connectivity. This may

us to determine the degree to which a predictor makes someone similar which can be important in determining something like the relational herding or prism/pipe effects. To be discussed.

[2] We recognise that this may not be ideal and we are looking for ways to deal with the 100% reciprocity. Potential point for discussion.

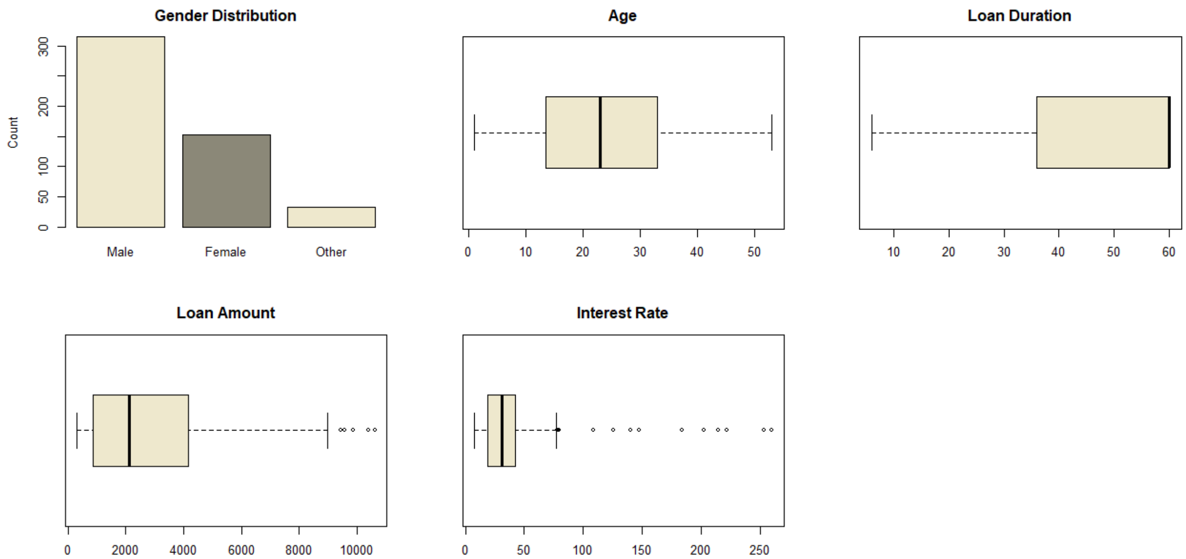
have been a result of the threshold set for the similarity. The former suggests that few nodes act as the bridge between different communities of borrowers and that the few highly connected nodes are crucial for the connection of the wider P2P network.

Figure 2: Degree and Betweeness Distribution of the P2P Network



The figure below shows various aspects of the wider dataset. Firstly, we observe that there are significantly more males within our subsample borrowing than females. Notably, some users did note "Other" for their gender. It is unclear whether this relates to the user being non-binary or whether it was a data error. Generally, borrowers tend to be young, with the median hovering around 22. The interquartile range is approximately 10 years, suggesting that most users lean younger. The duration of users' loans tends to be extremely left skewed, indicating the the majority of users request very long loans. There is quite a large range in terms of the loan amounts and their interest rates. The median loan amount hovers around 2000 euros, with most of the users not borrowing more than 4000. This distribution reaffirms that users prefer loans on the lower end. For most users, the interest rate is tightly dispersed and below 40%. This could suggest that loan providers are quite selective in their clientèle. Notably, some users do exceed 100% interest rates, suggesting that the data has users with extraordinary circumstances.

Figure 3: Descriptive Statistics of the wider Dataset



References

Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *Academy of Management Review*, *27*(1), 17–40.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235–251). Elsevier.

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., … Funk, B. (2011). Online peer-to-peer lending – a literature review. *Journal of Internet Banking and Commerce*, *16*.

Barasinska, N., & Schaefer, D. (2010). Does gender affect funding success at the peer-to-peer credit markets? Evidence from the largest german lending platform. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1738837

Cai, S., Lin, X., Xu, D., & Fu, X. (2016). Judging online peer-to-peer lending behavior: A comparison of first-time and repeated borrowing requests. *Information & Management*, *53*(7), 857–867. https://doi.org/10.1016/j.im.2016.07.006

Chen, X., Zhou, L., & Wan, D. (2016). Group social capital and lending outcomes in the financial credit market: An empirical study of online peer-to-peer lending. *Electronic Commerce Research and Applications*, *15*, 1–13. https://doi.org/10.1016/j.elerap.2015.11.003

De Liu, Brass, D. J., Lu, Y., & Chen, D. (2015). Friendships in online peer-to-peer lending. *MIS Quarterly*, *39*(3), 729–742. Retrieved from https://www.jstor.org/stable/26629629

Devenow, A., & Welch, I. (1996). Rational herding in financial economics. *Papers and Proceedings of the Tenth Annual Congress of the European Economic Association*, *40*(3), 603–615. https://doi.org/10.1016/0014-2921(95)00073-9

Herzenstein, M., Dholakia, U. M., & Andrews, R. L. (2011). Strategic herding behavior in peer-to-peer loan auctions. *Journal of Interactive Marketing*, *25*(1), 27–36. https://doi.org/10.1016/j.intmar.2010.07.001

Jackson, M. O. (2011). An overview of social networks and economic applications. In *Handbook of social economics* (Vol. 1, pp. 511–585). Elsevier. https://doi.org/10.1016/B978-0-444-53187-2.00012-7

Mingfeng Lin, N.R. Prabhala, & Siva Viswanathan. (2009). *Social networks as signaling mechanisms: Evidence from online peer-to-peer lending*. Retrieved from https://pages.stern.nyu.edu/~bakos/wise/papers/wise2009-p09_paper.pdf

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, *23*(2), 242–266.

Podolny, J. M. (1993). A status-based model of market competition. *American Journal of Sociology*, *98*(4), 829–872. https://doi.org/10.1086/230091

Pope, D., & Sydnor, J. (2008). What's in a picture? Evidence of discrimination from prosper.com. *Monetary Economics eJournal*, *46*. https://doi.org/10.1353/jhr.2011.0025

Putnam, R. D. (2015). Bowling alone: America's declining social capital. In *The city reader* (pp. 188–196). Routledge.

Siddhartha, M. (n.d.). Bondora peer to peer lending loan data. Retrieved October 9, 2025, from https://www.kaggle.com/datasets/sid321axn/bondora-peer-to-peer-lending-loan-data

A Source Code - Data Processing

```r

# ---------------------------------------------------------------------------- #
# Reset WD
#setwd("../..")

# Import the Bondora P2P Dataset
bondora_raw <- read.csv("dataset/LoanData_Bondora.csv",
                        header = TRUE)
cols <- colnames(bondora_raw)

# ---------------------------------------------------------------------------- #

# Subset only Columns we need
keep_cols <- c("LoanId", "UserName","NewCreditCustomer","LanguageCode",
               "Age", "Gender", "Country", "Amount", "Interest",
               "LoanDuration","UseOfLoan", "Education", "MaritalStatus",
               "NrOfDependants", "Rating", "Restructured",
               "NoOfPreviousLoansBeforeLoan", "MonthlyPayment")
bondora <- bondora_raw[keep_cols]

# Remove Rows with NAs -> Complete Dataset Preferred
bondora_complete <- na.omit(bondora)
sum(is.na(bondora_complete))

# Observe Class of Each Attribute
sapply(bondora_complete, class)

# Make Binary Indicators Binary
new_customer_mapping <- c("True" = 1, "False" = 0)
bondora_complete$NewCreditCustomer <- new_customer_mapping[
  bondora_complete$NewCreditCustomer]

# Replace User inputs of Blank Dependants with Zero
bondora_complete$NrOfDependants[bondora_complete$NrOfDependants == ""] <- NA
bondora_complete$NrOfDependants[is.na(bondora_complete$NrOfDependants)] <- 0

# Make the Column Numeric
bondora_complete$NrOfDependants <- as.numeric(bondora_complete$NrOfDependants)
bondora_complete$NrOfDependants[is.na(bondora_complete$NrOfDependants)] <- 0

# Make Restructured Binary
bondora_complete$Restructured <- new_customer_mapping[
  bondora_complete$Restructured]

# Randomly Remove Observations until Desired Size is Reached
set.seed(42)
sample_size <- 500
sample_indices <- sample(1:nrow(bondora_complete), sample_size)
bondora_sample <- bondora_complete[sample_indices, ]

# ---------------------------------------------------------------------------- #

# Choose Feature Subset for Similarity Metric
similarities <- c("LoanDuration","Amount","MonthlyPayment", "NewCreditCustomer",
                  "NoOfPreviousLoansBeforeLoan","LanguageCode")
bondora_similar <- bondora_sample[similarities]

# Standardise Numeric Features in the Similarity Set
bondora_similar_scaled <- scale(bondora_similar)

# Compute Cosine Similarity
cosine_sim <- function(X) {
  # numerator: dot product
  sim <- X %*% t(X)
```

```
  # denominator: product of norms
  norms <- sqrt(rowSums(X^2))
  sim <- sim / (norms %*% t(norms))

  return(sim)
}

similarity_matrix <- cosine_sim(bondora_similar_scaled)

# Get the Usernames for the Random Lenders
vertex_names <- as.character(bondora_sample$UserName)

# Get upper triangle indices
ut <- which(upper.tri(similarity_matrix), arr.ind = TRUE)

# Filter by threshold
ut <- ut[similarity_matrix[ut] >= threshold, ]

# Create edge list
p2p_bondera <- data.frame(
  from = vertex_names[ut[,1]],
  to   = vertex_names[ut[,2]],
  weight = similarity_matrix[ut],
  stringsAsFactors = FALSE
)

# Difference between attributes present and not present
att_diffs <- setdiff(keep_cols, similarities)

# Merge Data Frames to Ensure other Attributes Appear in Edge List
p2p_bondera <- merge(p2p_bondera, bondora_sample[att_diffs],
                     by.x = "from", by.y = "UserName", all.x = TRUE)
colnames(p2p_bondera)[4:14] <- paste0("from_", colnames(p2p_bondera)[4:14])

p2p_bondera <- merge(p2p_bondera, bondora_sample[att_diffs],
                     by.x = "to", by.y = "UserName", all.x = TRUE)
colnames(p2p_bondera)[15:25] <- paste0("to_", colnames(p2p_bondera)[15:25])

# ----------------------------------------------------------------------------- #

bondera_sample_atts$name <- bondera_sample_atts$UserName

p2p_bondera_network <- igraph::graph_from_data_frame(
  d = p2p_bondera[c('from','to','weight')], directed = FALSE)

walktrap_comm <- snafun::extract_comm_walktrap(p2p_bondera_network)
snafun::g_summary(p2p_bondera_network)

par(mfrow = c(1, 2))

plot(p2p_bondera_network,
     main = "No Community Detection",
     edge.arrow.size = 0.3,
     edge.color = rgb(0,0,0, alpha = 0.15),
     vertex.frame.color = "black",
     vertex.label = NA,
     vertex.frame.size = 3,
     vertex.size = 5,
     vertex.shape = "circle",
     vertex.color = "cornsilk",
     edge.curved = FALSE,
     layout = igraph::layout.fruchterman.reingold)

plot(walktrap_comm, p2p_bondera_network,
     main = "Walktrap Communities",
```

```r
    edge.arrow.size = 0.3,
    edge.color = rgb(0,0,0, alpha = 0.15),
    vertex.frame.color = "black",
    vertex.label = NA,
    vertex.frame.size = 3,
    vertex.size = 5,
    vertex.shape = "circle",
    vertex.color = "cornsilk",
    edge.curved = FALSE,
    layout = igraph::layout.fruchterman.reingold)

# Add Vertex Attributes
igraph::V(p2p_bondera_network)$Age <- bondera_sample_atts$Age[
  match(igraph::V(p2p_bondera_network)$name, bondera_sample_atts$name)]

bondora_sample$Education <- factor(bondora_sample$Education,
                                   levels = c(1, 2, 3, 4, 5),
                                   labels = c("Primary","Basic","Vocational",
                                              "Secondary", "Higher"))

bondora_sample$Gender <- factor(bondora_sample$Gender,
                                levels = c(0,1,2),
                                labels = c("Male","Female","Other"))

bondora_sample$MaritalStatus <- factor(bondora_sample$MaritalStatus,
                                       levels = c(1,2,3,4,5),
                                       labels = c("Married","Cohabitant",
                                                  "Single","Divorced","Widow"))

# Export the Network Object and Other Relevant Objects
saveRDS(p2p_bondera_network, 'resources/objects/p2p_network.RDS')
saveRDS(bondora_sample, 'resources/objects/bondora_sample.RDS')

# Descriptives
snafun::g_summary(p2p_network)
bootcamp::descriptives(bondora_df)

# ---------------------------------------------------------------------------- #

barplot(table(bondora_sample$Gender),
        col = c("cornsilk2", "cornsilk4", "black"),
        main = "Gender Distribution",
        ylab = "Count")

par(mfrow = c(2, 3), mar = c(4, 4, 3, 1))

barplot(table(bondora_sample$Gender),
        col = c("cornsilk2", "cornsilk4"),
        main = "Gender Distribution",
        ylab = "Count")

boxplot(bondora_sample$Age, main = "Age",
        col = "cornsilk2", horizontal = TRUE)
boxplot(bondora_sample$LoanDuration, main = "Loan Duration",
        col = "cornsilk2", horizontal = TRUE)
boxplot(bondora_sample$Amount, main = "Loan Amount",
        col = "cornsilk2", horizontal = TRUE)
boxplot(bondora_sample$Interest, main = "Interest Rate",
        col = "cornsilk2", horizontal = TRUE)
```

data_processing_bondora.R

B Source Code - Network Analysis

```r
# ------------------------------------------------------------------------- #
# Import the Network and Other Object
p2p_network <- readRDS('resources/objects/p2p_network.RDS')
bondora_df <- readRDS('resources/objects/bondora_sample.RDS')

# ------------------------------------------------------------------------- #

# Degree and Betweenness Distribution
deg_dist <- snafun::g_degree_distribution(p2p_network)
bet_dist <- snafun::v_betweenness(p2p_network)

par(mfrow = c(2, 2))

hist(deg_dist,
     main = "Unadjusted Degree Distribution",
     xlab = "Node Degree",
     col = "cornsilk2",
     border = "cornsilk4")

hist(log(deg_dist),
     main = "Logarithm of Degree Distribution",
     xlab = "Node Degree",
     col = "cornsilk2",
     border = "cornsilk4")

hist(bet_dist,
     main = "Unadjusted Betweenness",
     xlab = "Betweenness",
     col = "cornsilk2",
     border = "cornsilk4")

hist(log(bet_dist),
     main = "Logarithm of Betweenness",
     xlab = "Betweenness",
     col = "cornsilk2",
     border = "cornsilk4")
```

network_analysis.R

C Technology Statement

During the preparation of this work, we used ChatGPT in order to generate select parts of the R script utilised to process the dataset. Specifically, the tool was used to transform the processed dataset into a format that `igraph` would accept as a network object. No AI tool was utilised to write parts of the report. The following parts of the assignment were affected/generated by AI tool usage: **DATASET;** the data described within this section was processed partly by some code drafted by ChatGPT and edited by the group. After using this tool/service, **Samir Sabitli** evaluated the validity of the tool's outputs, including the sources that generative AI tools have used, and edited the content as needed. As a consequence, **Samir Sabitli** takes full responsibility for the content of their work.