Essays on Behavioral Finance in the Digital Age

Doctoral Thesis

Submitted to the
Department of Business Administration and Economics
Justus-Liebig-University Gießen

by
Daniel Lukasz Czaja
Bahnhofstraße 22
35390 Gießen
Daniel.L.Czaja@wirtschaft.uni-giessen.de

Supervisors:
Prof. Dr. Andreas Walter
Chair of Financial Services
Justus-Liebig-University Gießen

Prof. Dr. Christina E. Bannier
Chair of Banking and Finance
Justus-Liebig-University Gießen

Gießen, 27th of August 2020

Contents

List of Tables, Figures and Appendices

Tables

Figures

Appendices

I. Self-attribution bias and overconfidence among nonprofessional traders

Co-author: Florian Röder
Own share: 50%

This article has been published as:

Czaja, D. and Röder, F. (2020): Self-attribution bias and overconfidence among nonprofessional traders, *The Quarterly Review of Economics and Finance (forthcoming)*.
https://doi.org/10.1016/j.qref.2020.02.003

Previous versions of this paper have been presented at the following conferences and workshops:

- European Annual Meeting of the Financial Management Association (FMA Europe), 2018, Kristiansand, Norway.
- 80th Annual Meeting of the German Academic Association for Business (VHB), 2018, Magdeburg, Germany.
- 35th Annual Conference of the French Finance Association (AFFI), 2018, Paris, France.
- 1st Personal Finance Workshop, 2018, Castle Rauischholzhausen, Germany.
- Social Finance and Financial Technology Conference, 2018, Edinburgh, UK.

Self-attribution bias and overconfidence among nonprofessional traders

Daniel Czaja[a] and Florian Röder[b]

Abstract - We investigate consequences of the self-attribution bias for nonprofessional traders. By applying a textual analysis of more than 44,000 public comments on a large social trading platform, we contribute to empirical literature on investment and trading behavior in three ways: First, we show that one component of the self-attribution bias, the self-enhancement bias, leads to subsequent underperformance. Second, results support the theory that traders become overconfident due to biased self-enhancement. Third, we find that traders' social trading portfolios attract higher investment flows from investors when showing self-enhancement biased behavior.

Keywords: Self-attribution bias; overconfidence; individual investors; trading behavior; social trading

JEL Codes: D14; G11; G41

[a] Institute of Financial Services, University of Giessen, Daniel.L.Czaja@wirtschaft.uni-giessen.de.

[b] Institute of Financial Services, University of Giessen, Florian.R.Roeder@wirtschaft.uni-giessen.de.

1. Introduction

Approximately 17.2 million households in the US own a brokerage account. About 11.1 millions of them make at least one transaction a year (Brien and Panis, 2015). However, literature points out that trading is hazardous to their wealth due to overconfidence (Barber and Odean, 2000). A theoretical model by Gervais and Odean (2001) relates overconfidence to individuals' tendency to overestimate the degree to which they are responsible for their own successes. This tendency is known as the self-enhancement bias. The self-enhancement bias is one component of the self-attribution bias (or the self-serving attribution bias) consisting of both, the self-enhancement bias and the self-protection bias (Gervais and Odean, 2001; Miller and Ross, 1975).

In this paper, we investigate the consequences of the self-attribution bias for nonprofessional traders and investors. Therefore, we look into three major research questions: First, does the self-attribution bias affect future trading performance? Second, does the self-attribution bias (and in particular the self-enhancement bias) trigger overconfidence? Third, do traders attract more investment flows from their investors when they are prone to the self-enhancement bias?

The self-attribution bias is a well-known subject in psychology (e.g., Anderson and Slusher, 1986; Miller and Ross, 1975; Tetlock and Levi, 1982) that recently gained attention in management research as well (e.g., Billett and Qian, 2008; Kim, 2013; Libby and Rennekamp, 2012). Relating to investors, Hoffmann and Post (2014) find evidence for the existence of the self-attribution bias among individual investors. However, it is not clear whether biased self-attribution influences trading performance. In addition, there is no empirical evidence for the relationship between the self-attribution bias and overconfidence among investors, yet. Furthermore, there is no study distinguishing between the effects of the self-enhancement bias and the self-protection bias on individuals' financial decisions. Lastly, we are not aware of any study that investigates if traders' self-attribution biased behavior affects their investors.

To address our research questions, we use data from a social trading platform. On this platform, traders manage social trading portfolios in which investors can invest.[1] In detail, we label individuals managing virtual portfolios on the platform as ’traders’ while we label individuals investing real money in the corresponding structured product as ‘investors’. Following the idea of social networks, traders can write public comments about their transactions addressed to their investors. Traders can freely assess whether and when to write a comment. In addition, traders can determine scope and content of comments at their discretion. Those circumstances allow traders to express their thoughts, attitudes and purposes unforcedly. Our sample covers more than 44,000 public comments of more than 2,000 traders that offer investable social trading portfolios in the period from 2012 to 2016. Based on those comments, we can identify self-attribution biased traders.

We apply a 'bag-of-words'-model based content analysis (Salton and McGill, 1983) to measure the difference between the share of first person pronouns and the share of third person pronouns within a comment to proxy for the self-reference of a trader. To identify traders being prone to the self-attribution bias, we use traders’ self-reference in relation to past performance following the approach of Kim (2013) and Li (2010). We then examine the effects of biased self-attribution on nonprofessional traders and investors by using a time- and portfolio fixed effects panel regression framework. In doing so, we are able to estimate the within variation of one trader over time, which ensures that regressions are robust to trader-specific, time-invariant omitted variables.

Results suggest that the self-enhancement bias leads to future underperformance. Moreover, a trader shows higher future trading frequencies and portfolio turnovers as well as lower portfolio diversification when she is prone to the self-enhancement bias. This relationship is in line with theoretical literature suggesting that overconfident behavior results from biased self-attribution. Traders that excessively attribute high past returns to their own abilities (self-enhancement bias) become overconfident (Gervais and Odean, 2001), and thus, subsequently underperform (Barber and Odean, 2000).

[1] We define a (social trading) portfolio as a virtual portfolio managed by a trader on the investigated social trading platform. A partner of the platform can issue a structured product (certificate) that replicates the performance of the virtual portfolio so that investors can invest real money in the social trading portfolio. Our sample contains investable social trading portfolios only.

In addition, we find that portfolios receive significantly more investment flows from investors when the trader is prone to the self-enhancement bias. As investors might perceive biased traders as more confident, these results are in line with literature suggesting that confidence strengthens individual's social status or perceived level of knowledge and trustworthiness (Anderson et al., 2012; Price and Stone, 2004). However, as traders prone to the self-enhancement bias subsequently underperform, the self-enhancement bias does not only harm the affected trader but also her investors.

Our paper is an important contribution to literature because of at least three reasons. First, we are the first to examine the effects of biased self-attribution on trading performance. Second, we provide first empirical evidence that supports the hypothesis of a link between the self-enhancement bias and overconfident behavior among nonprofessional traders as suggested by theoretical literature (Gervais and Odean, 2001). Third, we present novel findings on how investors react to traders showing characteristics of the self-enhancement bias.

Results are robust to using different return measures, namely market adjusted returns, Carhart (1997) four-factor alphas and Sharpe ratios. Following Heckman (1979), we additionally correct for a potential sample selection bias that results from the fact that not all traders at the platform write comments. Furthermore, results are robust to applying different methods of identifying self-attribution biased traders. Lastly, we address potential reverse causality issues by showing that overconfident trading behavior does not trigger the self-enhancement bias or the self-protection bias, respectively.

2. Hypotheses and related literature

2.1 The self-attribution bias

The self-attribution bias is a well-documented mental process in personality psychology. It refers to the tendency to credit oneself and one's own abilities excessively with past success but to blame others or external factors for failures (Campbell and Sedikides, 1999; Miller and Ross, 1975; Zuckerman, 1979). Consequently, the self-attribution bias can be separated into two components. While the self-enhancement bias refers to the attribution of past success, the self-protection bias denotes the shirking of responsibility for failures. Evidence from psychological literature suggests various explanations for these biases that can be classified either as motivational or cognitive reasoning (Shepperd et al., 2008).

Motivational reasoning refers to self-enhancement and self-presentation. According to this, people ascribe achievements to themselves in order to appear positively to others (Schlenker, 1980). Cognitive reasoning, however, explains the self-attribution bias as a result of cognitive evaluation of achievements (Schlenker, 1980). Based on this, individuals tend to show an illusion of objectivity resulting in the self-attribution bias as they look for explanations with the least amount of effort (Kunda, 1990). Since they have positive expectations, individuals do not question positive results and attribute these to their own abilities. However, they try to find possible explanations other than their own insufficiency to evaluate negative outcomes (Schlenker, 1980).

The economic literature also reports on the self-attribution bias, especially in studies that refer to a management context. Bettman and Weitz (1983) find that managers take credit for positive results, but blame external factors for failures because of motivational and cognitive reasons. Recently, studies about earnings forecast issuance (Baginski et al., 2004; Baginski et al., 2000; Libby and Rennekamp, 2012) as well as mergers and acquisitions (Billett and Qian, 2008; Doukas and Petmezas, 2007; Kim, 2013) show that managers are prone to the self-attribution bias.

Within the field of investing and trading behavior, however, we only know little about the self-attribution bias, yet. Hilary and Menzly (2006) suggest that analysts are affected by the self-attribution bias. Moreover, two studies examine online traders' self-perception of their trading records (Dorn and Huberman, 2005; Hoffmann and Post, 2014). Hoffmann and Post (2014) show that the higher the past returns of individual investors, the more they agree that past performance reflects their investment skills. Dorn and Huberman (2005) provide evidence that biased self-attribution affects the risk attitude of traders.

2.2 The self-attribution bias and overconfidence

Overconfidence describes the tendency of individuals to overestimate their own abilities. In general, the literature suggests that overconfidence significantly influences people's behavior (McCannon et al., 2016). Regarding trading behavior, various studies support this finding. The literature suggests a link between overconfidence and trading frequency (Barber and Odean, 2001; Chen et al., 2007; Glaser and Weber, 2007; Odean, 1998). Additionally, overconfident traders take higher risks (Barber and Odean, 2000; Merkle,

2017) and hold less diversified portfolios (Goetzmann and Kumar, 2008; Merkle, 2017). Moreover, the economic literature finds empirical evidence for a link between biased self-attribution and overconfident behavior among managers in the context of mergers and acquisitions (Billett and Qian, 2008; Doukas and Petmezas, 2007), management forecasting (Libby and Rennekamp, 2012) and public communication (Kim, 2013). Hilary and Menzly (2006) find this relationship among analysts, as well. However, there is no empirical evidence on the link between the self-attribution bias and overconfidence among investors or traders, yet.

Based on the idea of learning, Gervais and Odean (2001) develop a theoretical multi-period market model linking biased self-attribution of traders with subsequent overconfidence. Not knowing about their own abilities, traders draw inferences from successes and failures. Since causal reasoning is biased, the self-attribution bias leads traders to become overconfident. In their model, traders are not overconfident initially, but overconfidence may only result from assessing past trading experience. In the context of trading, we typically assume that overconfidence is rather triggered by past successes than by past failures. This assumption is in line with studies suggesting the self-enhancement bias being more important than the self-protection bias (Fiske and Taylor, 1991; Gervais and Odean, 2001; Miller and Ross, 1975). As overconfident trading behavior leads to subsequent underperformance (Barber and Odean, 2000), we expect that traders perform worse when they are prone to biased self-attribution.

Overall, traders prone to the self-enhancement bias should display overconfident trading behavior and thus, subsequently underperform. Therefore, we formulate the following hypothesis:

H1: Traders underperform when they are prone to the self-enhancement bias (H1a) as they develop overconfident trading behavior (H1b).

2.3 Perception of biased self-enhancement

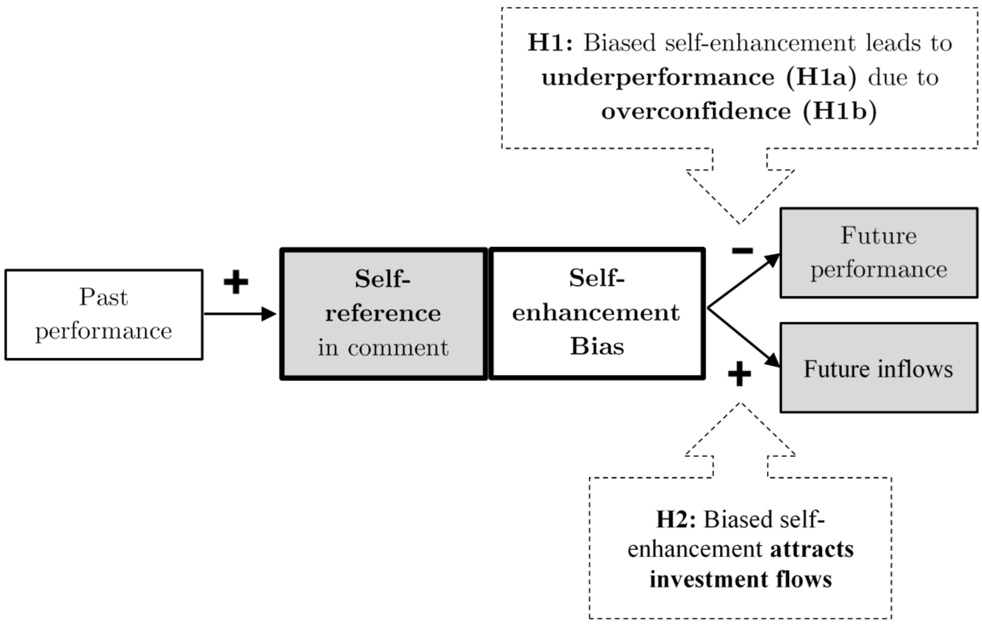
To the best of our knowledge, there is no study investigating the effect of the self-attribution bias on others. Following the concept of motivational reasoning, however, the self-enhancement bias also includes self-presentation: people ascribe achievements to themselves in order to appear positive to others (Schlenker, 1980). Hence, we assume that investors might perceive traders that excessively credit themselves with past successes as more confident compared to others. In this context, psychological research suggests that confidence affects perception and treatment by others (Chance and Norton, 2015). There is evidence that individuals adopt recommendations by confident people more likely than by non-confident ones (Van Swol and Sniezek, 2005). Furthermore, confident people are perceived to be more knowledgeable (Price and Stone, 2004). Thus, higher confidence leads to a higher social status (Anderson et al., 2012). We assume that those findings might also apply to individual investors. Therefore, we formulate the following hypothesis:

H2: Traders attract higher investment flows when they are prone to the self-enhancement bias.

Figure I-1 shows a graphical summary of our hypotheses.

Figure I-1: Hypotheses

This figure gives an overview of the hypothesized relationships.



3. Data, variables and summary statistics

3.1 Data

We use data from a big European social trading platform. The data was provided by the platform upon request. The platform allows traders to manage virtual portfolios in which investors can indirectly invest via exchange traded structured products. In this study, we label individuals managing virtual portfolios on the platform as 'traders' while we label individuals investing real money in the corresponding structured product as 'investors'.

After signing up, traders can publish their investment ideas and start trading in their virtual portfolio. In doing so, all their trading transactions and their trading performance become publicly available. The platform offers a large investment universe ranging from stocks, bonds, mutual funds, ETFs to structured products and even derivatives. Possible investors can signal interest in those social trading portfolios. When there are enough investors interested in a social trading portfolio[2], a structured product (open-ended index certificate) can be issued that replicates its performance. A partner of the social trading platform acts as issuer and index sponsor (in two legal entities) of the structured products. Investors can invest real money in the social trading portfolio by buying the structured product. After the issue of the structured product, the trader still manages the corresponding virtual portfolio and therefore affects the price of the structured product. Trading of those structured products takes place at a regular European exchange. The platform, the issuer and the traders earn fees from the investors.[3] Besides, following the basic principles of a social network, traders can write public comments to communicate with (possible) investors or other traders. Those comments are our main object of investigation. See Appendix I-D for comment examples. The comments are the only way of communication for traders and investors on the social trading platform. For more detailed information about the social trading platform, see Oehler et al. (2016) or Röder and Walter (2019).

One could argue that our dataset is subject to a selection bias, as traders on the social trading platform on average might be more prone to overconfidence or biased self-attribution than other individual investors. We address this issue by applying a fixed portfolio effects model. Estimating the within variation of a trader's variables over time, the absolute level of traders' self-attribution bias or overconfidence is negligible.

[2] More precisely, a trader's social trading portfolio must attract at least ten supporters with a watchlisted capital of at least 2,500 euros. In addition, the social trading portfolio must already exist for at least three weeks.

[3] The issuer earns a fixed annual fee depending on the invested money in the structured product. The platform and the trader share the so called 'performance fee' that depends on the one year performance of the social trading portfolio (based on the high watermark) as well as on the money invested in the structured product.

Our dataset ranges from June 2012 to November 2016 and features daily performance and flow data as well as all public comments written by traders of social trading portfolios that either are or once were investable for investors. Additional, trading data includes all transactions of the social trading portfolios on a daily basis.

The original dataset covers more than 90,000 public comments. To achieve our final sample we make five adjustments. First, we only consider observations of investable social trading portfolios. We make this adjustment to avoid biases resulting from a possible gambling behavior of traders when not being responsible for real money. Second, we measure most of our variables on a 360-days basis. As a result, we lose social trading portfolios that are investable for a time horizon of fewer than 360 days during our sample period. Third, as we focus on nonprofessional traders, we exclude all social trading portfolios managed by professional asset management companies. Fourth, we combine all comments of one portfolio on one day to one observation.[4] Fifth, we exclude all comments with less than three words, as those comments seem not to include relevant information. The final sample covers 44,985 observations of 3,519 social trading portfolios.

3.2 Construction of variables

In our study, we have four different groups of variables. First, we use variables regarding traders' comments that we derive from content analysis. These variables include the time since the last comment of the trader, the length, tone[5] and readability[6] of comments as well as the traders' self-reference within comments. Second, we use social trading portfolio data, including performance (raw return, market adjusted return, Carhart four-factor alpha and the Sharpe ratio), return volatility, investment flows from investors into and out of the social trading portfolio, age of the social trading portfolio and assets under management. Third, we build measures for the self-attribution bias, the self-enhancement

[4] In the following, the term 'comment' denotes all comments of a portfolio on one day.

[5] Following Twedt and Rees (2012), we measure the tone of a comment as the difference of the numbers of positive and negative words relative to the overall number of words of the comment. We classify words as positive, negative or neutral by using the word lists of Bannier et al. (2019). This approach has been applied by several economic studies before (e.g., in Hanley and Hoberg, 2010; Kothari et al., 2009; Loughran and McDonald, 2011; Rogers et al., 2011).

[6] We define the readability measure as the average number of words per sentence plus the percentage of words with more than six letters following Bjornsson (1968).

bias and the self-protection bias. Fourth, we construct proxies for overconfidence to examine whether self-enhancement biased traders show overconfident trading behavior. Those proxies include the number of trading transactions, purchases and sales, number of different securities in the portfolio, portfolio turnover and the maximum of absolute daily returns of the portfolio (Merkle, 2017).

In the following, we describe the construction of our most important variables. Please find a detailed description of all variables in Appendix I-A (Table I-A.1).

To investigate traders' public comments, we use a dictionary based content analysis (Kearney and Liu, 2014; Kim, 2013; Loughran and McDonald, 2011). Applying the 'bag-of-words'-model, we first disaggregate each comment into its single words (Salton and McGill, 1983). Next, we count the number of connoted words in the comment as classified by the business specific word lists of Bannier et al. (2019).

To identify nonprofessional traders being prone to the self-attribution bias, we first measure self-reference (*Self Ref*) within the comments. We follow Kim (2013) and Li (2010) in the construction of this variable using the LIWC (Linguistic Inquiry and Word Count) dictionary by Wolf et al. (2008). In doing so, we define *Self Ref* of social trading portfolio i on day t as the quotient of the number of first person personal pronouns (category 'Self' in the LIWC) minus the number of third person personal pronouns (category 'Other' in the LIWC) and the overall number of words of a comment (in percentage terms). See Appendix I-D for examples of comments and calculation of *Self Ref*.

$$Self\ Ref_{i,t} = 100 * \frac{Number\ Self_{i,t} - Number\ Other_{i,t}}{Number\ Words_{i,t}} \tag{1}$$

We us three different approaches to measure the performance of social trading portfolios: market adjusted return (*Market Adjusted*), the Carhart (1997) four factor alpha (*4F Alpha*) and the Sharpe ratio (*Sharpe*). We define the market adjusted return as the raw return of social trading portfolio *i* minus the return of the MSCI World index in the same period. We obtain the four factor alphas by using international factors provided by the web page of Kenneth R. French (French, 2017). Furthermore, we use the Sharpe ratio to obtain a return measure that is independent from any benchmark. In addition, Doering et al. (2015) show that social trading portfolios produce hedge fund-like returns, while Eling and

Schuhmacher (2007) find that the Sharpe ratio is an appropriate measure for hedge funds' performance. To ensure interpretability of our results in the case of negative returns, we refine the Sharpe ratio as suggested by Israelsen (2005).[7]

Following the literature on mutual funds and hedge funds, we measure investment flows into and out of the corresponding structured product of a social trading portfolio expressed as percentages (e.g., Fung et al., 2008; Huang et al., 2007; Sirri and Tufano, 1998). We define *Net Flows* as euro inflows minus euro outflows into (out of) the structured product of portfolio *i* during the last 360 days divided by assets under management[8] (*AUM*) in *t-360*.

$$Net\ Flows_{i,t} = 100 * \frac{Euro\ Inflows_{i,t}\ - Euro\ Outflows_{i,t}}{AUM_{i,t-360}} \tag{2}$$

To investigate the relationship between the self-attribution bias and subsequent returns as well as investment flows, we need to identify traders being prone to the self-attribution bias.

Evidence suggests that people tend to use more self-inclusive rather than self-exclusive personal pronouns in more positive contexts (Sendén et al., 2014). Moreover, self-reference is an increasing function of past success (Shepperd et al., 2008). In particular, we assume that traders use more self-inclusive (first person pronouns) and less self-exclusive (third person pronouns) personal pronouns when showing good past performance. Therefore, we follow Kim (2013) and Li (2010) in building a measure for the self-attribution bias. We estimate a portfolio fixed effects linear regression of *Self Ref* of social trading portfolio *i* on day *t* on the past 360-days raw return (*Past Performance*) of the respective portfolio. To adjust for possible heteroscedasticity and within-panel correlation, we use robust standard errors clustered by portfolio *i*.

$$Self\ Ref_{i,t} = \alpha + \beta\ Past\ Performance_{i,t} + \varepsilon_{i,\ t} \tag{3}$$

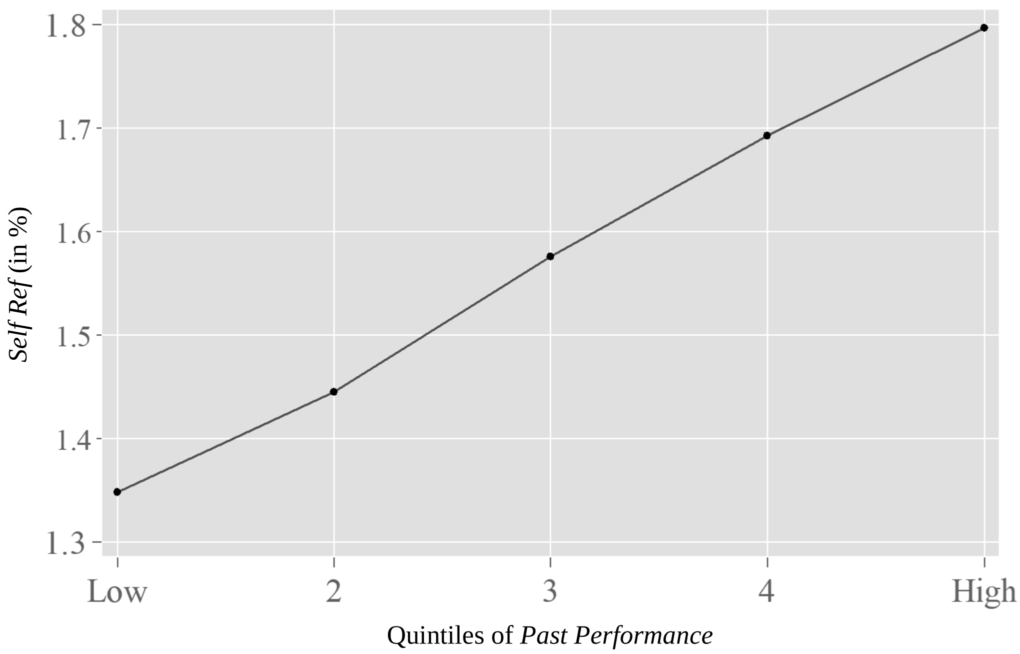
[7] Following Israelsen (2005), we add an exponent to the denominator of the Sharpe ratio (standard deviation of excess return). The exponent is the excess return divided by the absolute value of excess return.

[8] On the investigated social trading platform, investors invest in a structured product that replicates the performance of an underlying social trading portfolio. We define assets under management (*AUM*) as the invested money in the structured product of the underlying portfolio *i* in *t*.

The estimate of the coefficient β is 0.003. The estimation of β is statistically significant at the 10% level. This result implies that traders with good past performance attribute performance to themselves, while they attribute poor past performance to external factors. Our findings conform to Kim (2013) investigating this relationship among CEO interviews. A graphical analysis presented in Figure I-2 confirms the underlying assumption of a positive relationship between *Past Performance* and *Self Ref*.

Figure I-2: Mean of *Self Ref* by *Past Performance* quintiles

We structure the 360-day raw return (*Past Performance)* of investigated social trading portfolios *i* in *t* in quintiles. This figure shows the means of self-reference in the comments (*Self Ref)* among these performance quintiles. *Self Ref* is the quotient of the number of first person personal pronouns (category "Self" in the LIWC) minus the number of third person personal pronouns (category "Other" in the LIWC) and the overall number of words in the comment of portfolio *i* in *t* in percent. The difference between the low performance group and the high performance group is statistically significant at the 1% level.



Based on the positive relationship between *Past Performance* and *Self Ref*, we then define proxies for the self-enhancement bias (*SEB*), the self-protection bias (*SPB*) and the self-attribution bias (*SAB*) as follows:

$$SEB_{i,t} = \begin{cases} 1 \text{ for } Past\ Performance_{i,t} > 0 \wedge \varepsilon_{i,t} > 0 \\ 0 \text{ for } Past\ Performance_{i,t} \leq 0 \vee \varepsilon_{i,t} \leq 0 \end{cases} \tag{4}$$

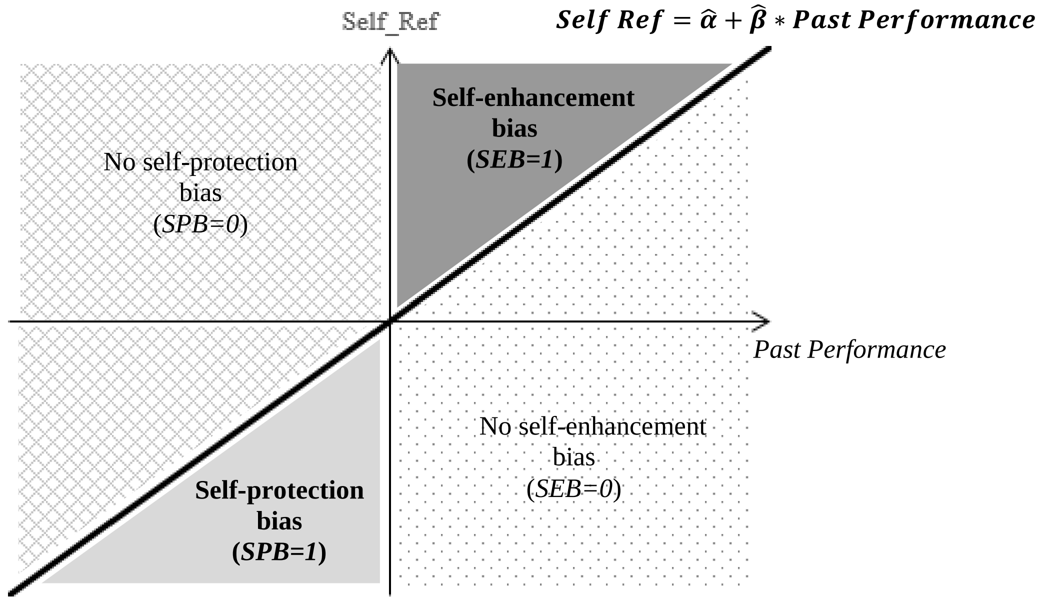
$$SPB_{i,t} = \begin{cases} 1 \text{ for } Past\ Performance_{i,t} < 0 \wedge \varepsilon_{i,t} < 0 \\ 0 \text{ for } Past\ Performance_{i,t} \geq 0 \vee \varepsilon_{i,t} \geq 0 \end{cases} \tag{5}$$

$$SAB_{i,t} = \begin{cases} 1 \; for \; SEB_{i,t} = 1 \vee SPB_{i,t} = 1 \\ 0 \; for \; SEB_{i,t} = 0 \wedge SPB_{i,t} = 0 \end{cases} \tag{6}$$

We identify a trader as self-enhancement biased (*SEB* = 1) when she exhibits excessively high self-referencing behavior ($\varepsilon_{i,t}$ >0) within a comment when her social trading portfolio performance was positive ($Past\ Performance_{i,t} > 0$). We identify a trader as self-protection biased (*SPB* = 1) if she exhibits excessively low self-referencing behavior ($\varepsilon_{i,t}$ <0) within a comment when her portfolio performance was negative ($Past\ Performance_{i,t} < 0$). Lastly, we identify a trader as self-attribution biased (*SAB* = 1) if she is either self-enhancement biased or self-protection biased. See Figure I-3 for a visual presentation of the variable construction.[9]

Figure I-3: Construction of *SAB*, *SEB* and *SPB*

We identify a trader as self-enhancement biased (*SEB*=1) if she shows excessively high self-referencing behavior in the comment when the 360-day raw return of portfolio *i* in *t* is positive. We identify a trader as self-protection biased (*SPB*=1) if she shows excessively low self-referencing behavior in the comment when the 360-day raw return of portfolio *i* in *t* is negative. Lastly, we identify a trader as self-attribution biased (*SAB*=1) if she is either self-enhancement biased or self-protection biased.



[9] Note that we applied an alternative method of creating the variables *SEB*, *SPB* and *SAB* without using a regression to examine the robustness of our results. Results are comparable to our main results and are available in Appendix I-C.

In comparison to Kim (2013) and Li (2010), we make two adjustments: first, we use raw returns instead of Carhart four-factor returns (Carhart, 1997) to adjust the measure for the use among nonprofessional traders. We do so because the raw return is the only performance measure that is directly accessible on the main page of every social trading portfolio on the web page of the social trading platform. Additionally, Röder and Walter (2019) find that participants on the investigated social trading platform rely on raw returns rather than on factor model alphas or the Sharpe ratio. Moreover, literature apart from social trading suggests that nonprofessional traders are more likely to follow raw returns rather than factor-adjusted returns (Clifford et al., 2013; Veld and Veld-Merkoulova, 2008). Second, we extend the method insofar that we do not only create a measure for the self-attribution bias but also for the self-enhancement bias and the self-protection bias, separately.

Our methodology to identify self-attribution biased traders could be affected by the following three limitations. First, as we use these self-attribution bias measures in our second stage regressions findings may show an attenuation bias (Cameron and Trivedi, 2005). This bias refers to the underestimation of an estimator because of measurement errors in the independent variables. Therefore, the expected value of an estimator is lower than the actual value of the parameter. In our linear regression, this issue would bias against finding significant coefficients. Second, traders on the social trading platform, on average, might be more prone to overconfidence or biased self-attribution than other individual investors. We address this issue by applying a fixed portfolio effects model in all our main estimations. Estimating the within variation of one trader's variables over time, we measure if a dependent variable differs for a trader being prone to the self-attribution bias (at one point in time) compared to the same trader not being prone to the self-attribution bias (at another point in time). Third, the probability that a trader writes a comment could depend on (time variant) portfolio characteristics. For example, Ammann and Schaub (2017) find that social traders are more likely to write comments, when they show a positive past performance. To address this type of selection bias we apply the two-stage Heckman (1979) correction. Therefore, we first estimate a probit model investigating the determinants of writing a comment.[10] Then, we use the hazard rates of this regression as a control variable in all regressions of our main results.[11]

To investigate the relationship between the self-enhancement bias and overconfidence, we construct proxies for overconfidence, such as trading frequency, turnover, trading volume or degree of diversification. The construction of these variables is based on Merkle (2017). See Appendix I-A (Table I-A.1) for a detailed list of all variables and construction details.

3.3 Summary statistics

Table I-1 shows the summary statistics. As can be referred from panel A, the self-reference (*Self Ref*) within traders' comments is zero for at least 25% of the observations, implying the same number of first and third personal pronouns or no use of personal pronouns in these comments at all. Traders write more often about themselves than about others as more than 25% of observations show a positive sign, while less than 25% are negative. The average number of words in one observation (*Length of Comment*) is 57.36, while the median is 31.00. It follows that most observations include several sentences. The mean of *Tone,* i.e. the mean of the difference between positive and negative words relative to the overall words in the comments, is -0.11 percent. Moreover, the median of *Tone* is zero. Hence, the overall tone of comments is rather neutral. The mean of *Readability* is 0.46 which can be interpreted as moderately difficult (Bjornsson, 1968).

Panel B shows details about portfolio data. The means of *Market Adjusted* as well as *4F Alpha* are high, with values of 4.60 and 4.52 percent, respectively. These high alphas might result from the fact that our sample is restricted to traders that actually write comments. As mentioned before, traders that show a high past performance are more likely to write comments (Ammann and Schaub, 2017). Furthermore, as social trading investors follow past performance (Röder and Walter, 2019), portfolios with higher past performance will survive longer and thus, represent a larger proportion in our sample. We address this selection bias by adding portfolio fixed effects as well as by applying the Heckman (1979) correction as explained in section 3.2 and Appendix I-B.

[10] For regression results see Table I-B.1 in Appendix I-B.

[11] Note that our main results are robust to not controlling for possible selection bias. Results are available upon request.

While the average invested money in the social trading portfolios (*AUM*) is 275,801 euros, the median is only 9,512 euros, which indicates that there are a lot of small portfolios as well as only few large portfolios. This is one reason why percentage *Inflows* and *Net Flows* into and out of the social trading portfolios are relatively high in comparison to mutual fund flows (Sirri and Tufano, 1998), showing values of 5,240 and 3,690 percent of *AUM*, respectively. Additionally, the social trading platform shows an annual growth rate of more than 30% during our sample period, leading to high investment flows in comparison to *AUM*. The five percent percentile of *AUM* is zero. However, we exclude those observations from our estimations as we control for the natural log of *AUM* in our regressions. We do so because traders might behave differently when not being responsible for real invested money.

Panel C contains dummies that identify traders as self-attribution biased, self-enhancement biased and self-protection biased, respectively. We identify traders in 45% of the comments as self-attribution biased. By construction, this variable shows a mean close to 50%. As most of the raw returns are positive, we find slightly more comments being self-enhancement biased than self-protection biased.

Descriptive statistics of overconfidence proxies (Panel D) show that the average social trading portfolio in our sample holds 44 different securities (*# Securities*) and shows a *Turnover* of 4.84% of the current portfolio value every 90 days. These numbers suggest active diversification and moderate trading. However, the average trader in our sample makes 205 trading transactions *(# Transactions*) per 90 days while the median number of transactions is 76. The high number of transactions could result from the circumstance that transactions in social trading portfolios do not cause any transaction costs apart from bid-ask spreads. In summary, traders on the social trading platform make a high number of transactions, however, as most trades have a low trading volume portfolios show only a moderate turnover.

Table I-1: Summary statistics

This table contains the summary statistics of our dataset. We define the variables as follows: *Self Ref* is the quotient of the number of first person personal pronouns minus the number of third person personal pronouns and the overall number of words in comment of portfolio *i* in *t* in percent. *Length of Comment* is the average number of words in the comment of portfolio *i* in *t*. *Tone* is the difference of positive and negative words relative to the overall number of words in the comment of portfolio *i* in *t* in percent. *Readability* is the sum of average number of words per sentence and the percentage of words with more than six letters in the comment of portfolio *i* in *t* divided by 100. *Time-Lag Comment* are the days since the last comment of portfolio *i*. *# Comment* is the number of comments for portfolio *i* until day *t*. *Past Performance* is the 360-day raw return of portfolio *i* in *t* in percent. *Market Adjusted* is the 360-day raw return of portfolio *i* in *t* minus the 360-day raw return of the MSCI World index in *t* in percent. *Sharpe* is the 360-day Sharpe Ratio of portfolio *i* in *t* (negative values adjusted). *4F Alpha* is the 360-day four-factor alpha of portfolio *i* in *t* in percent. *AUM* are the assets under management, i.e. invested euros in the structured product of portfolio *i* on day *t*. *Inflows* is the sum of inflows to (the structured product of) portfolio *i* over the last 360 days divided by the invested euros in portfolio *i* in *t-360* in percent. *Net Flows* is the sum of inflows minus the sum of outflows to/out of (the structured product of) portfolio *i* over the last 360 days divided by the invested euros in portfolio *i* in *t-360* in percent. *Volatility* is the 360-day return volatility of portfolio *i* in *t* in percent. *Issue Age* is the age (since issue of the structured product) of portfolio *i* on day *t* in years. *SAB*, *SEB* and *SPB* are dummies that equal 1 if the comment of portfolio *i* in *t* is identified as self-attribution biased, self-enhancement biased or self-protective biased, respectively. *# Securities* is the average number of securities in portfolio *i* over the last 90 days. *Turnover* is the trading volume of portfolio *i* over the last 90 days divided by the value of virtual portfolio *i* in *t* in percent. *# Transactions*, *# Purchases* and *# Sales* are the number of transactions, purchases and sales, respectively, of portfolio *i* over the last 90 days. *Abs Max Return* is the 90-day maximum absolute daily raw return of portfolio *i* in percent.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	N	Mean	SD	P5	P25	P50	P75	P95
Panel A: Comments								
Self Ref	44,985	1.57	3.24	-0.51	0.00	0.00	2.86	7.69
Length of Comment	44,985	57.36	84.68	5.00	15.00	31.00	68.00	187.00
Tone	44,985	-0.11	3.10	-3.70	0.00	0.00	0.00	2.78
Readability	44,985	0.46	0.15	0.25	0.37	0.45	0.52	0.70
TimeLagComment	44,985	14.58	42.27	1.00	1.00	4.00	9.00	59.00
# Comment	44,985	101.92	102.68	4.00	26.00	70.00	145.00	302.00
Panel B: Portfolio Data								
Past Performance	44,985	6.54	17.90	-23.78	-2.40	7.35	17.57	33.87
Market Adjusted	44,985	4.60	17.40	-24.67	-4.01	5.15	15.11	31.06
Sharpe	44,985	13.75	14.91	-0.00	-0.00	9.51	23.08	42.75
4F Alpha	44,985	4.52	15.81	-22.50	-3.08	4.96	14.12	28.37
AUM	44,985	275,801	1,054,696	0	1,552	9,512	51,021	1,580,882
Inflows	23,193	5,242	102,972	0	13	84	366	4,097
Net Flows	23,193	3,689	78,501	-82	-18.51	19.26	175.82	2,106.08
Volatility	44,985	1.02	1.10	0.36	0.61	0.83	1.12	2.28
Issue Age	44,985	1.33	0.77	0.24	0.80	1.20	1.78	2.81
Panel C: Self-Attribution Bias Dummies								
SAB	44,985	0.45	0.50	0.00	0.00	0.00	1.00	1.00
SEB	44,985	0.24	0.43	0.00	0.00	0.00	0.00	1.00
SPB	44,985	0.21	0.41	0.00	0.00	0.00	0.00	1.00
Panel D: Trading Data								
# Securities	25,482	43.92	41.72	3.00	15.08	30.14	58.61	131.94
Turnover	25,482	4.84	9.40	0.05	0.60	1.79	5.30	18.76
# Transactions	25,482	204.83	382.89	2.00	22.00	76.00	210.00	891.00
# Purchases	25,482	107.18	196.30	1.00	11.00	36.00	111.00	495.00
# Sales	25,482	97.65	195.14	1.00	10.00	35.00	95.00	405.00
Abs Max Return	25,482	3.45	3.39	1.05	1.90	2.67	3.91	8.00

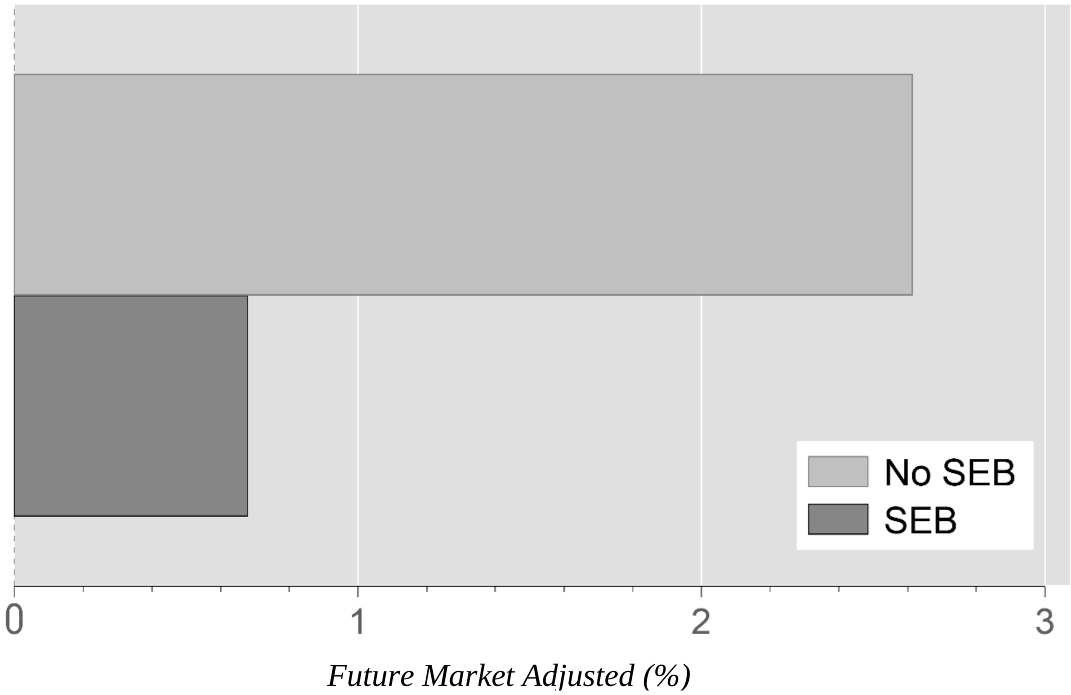
4. Results

4.1 The self-enhancement bias and future performance

We hypothesize that traders perform worse when they are prone to the self-enhancement bias (H1a). Figure I-4 pictures the univariate connection between the self-enhancement bias (*SEB*) and the future market adjusted performance (*Future Market Adjusted*). Consistent with all our investigations of the effect of the self-enhancement bias, we only compare traders that exhibit positive past performance (see *SEB*=1 and *SEB*=0 in Figure **I-3**). We do so to prevent any bias resulting from return momentum, mean reversions or similar effects. However, our results remain robust when we estimate the regressions without this adjustment.[12]

Figure I-4: Mean of future market adjusted returns

This figure illustrates the univariate relationship between self-enhancement bias (*SEB*) and *Future Market Adjusted*, i.e. the future 360-day raw return of portfolio *i* in *t* minus the future 360-day return of the MSCI World index in percent. We only include observations that show positive *Past Performance* (*SEB=1* and *SEB=0*). The difference is statistically significant at the 1% level.



While self-enhancement biased traders show 360-days future market adjusted returns of approximately 0.68% on average, non-biased traders show future returns of about 2.61% on average. The difference is statistically significant at the 1% level. This finding is a first indication that biased self-enhancement leads to future underperformance. To study this relationship in more detail, we use the following linear panel regression framework. In all our main regressions, we cluster standard errors by portfolio *i* and date *t* to adjust for possible heteroscedasticity and both, within-panel and cross-sectional correlation.

[12] Results are not included in the paper and are available upon request.

$$Future\ Return_{y,s,i,t} = \alpha_y + \beta_{y,\,s}\ Bias_{s,i,t} + \sum_{j}^{j=J} \gamma_{y,\,j}\ Control_{j,i,t} + \varepsilon_{y,s,i\ t} \tag{7}$$

We regress the different 360-days future return measures *y* (*Future Return*) of the social trading portfolio *i* on day *t* on the different bias dummies *s* (*Bias*) and controls *j* (*Control*). For the sake of completeness, we do not only examine the effect of the self-enhancement bias on future performance, but also the effects of the self-protection bias (*SPB*) and the self-attribution bias (*SAB*) as a whole. As stated before, estimating the effects of *SEB* and *SPB* separately, we only compare positive past performers and negative past performers with each other (*SEB* = 1 versus *SEB* = 0 or *SPB* = 1 versus *SPB* = 0). We control for the following potential determinants of future portfolio performance: past portfolio performance (*Past Performance*), *Tone*[13], *Readability,* and length of the comments (*Length of Comment*) as well as past net flows into the social trading portfolio (*Net Flows*). Moreover, we control for the natural logarithms of days since the last comment of the social trading portfolio *i* (*Ln Time Lag Comment*), number of comments for portfolio *i* until day *t (Ln # Comment),* age of social trading portfolio *i* in years (*Ln Issue Age*), euros invested in portfolio *i* (*Ln AUM*) and past return volatility (*Ln Volatility*). Additionally, the following trading controls are included: number of transactions (*Ln # Transactions*), average number of securities (*# Securities*), turnover (*Turnover*) and the maximum absolute daily raw return (*Max Return*) over last 90 days each. Finally, we include the hazard rates of the first stage regression of the Heckman correction (see section 3.2) as a control variable to control for a potential sample selection bias.

[13] Following Twedt and Rees (2012), we measure the tone of a comment as the difference of positive and negative words relative to the overall number of words of the comment. We classify words as positive, negative or neutral by using the word list of Bannier et al. (2019).

Table I-2: Regression of future performance on biased self-attribution

This table contains ordinary least squares linear regression results of three future performance measures (market-adjusted return, Carhart 4-Factor return and the Sharpe ratio) on *SEB, SPB* and *SAB,* respectively and a comprehensive set of control variables. We define variables as follows: *Future Market Adjusted* is the future 360-day raw return of portfolio *i* in *t* minus the future 360-day return of the MSCI World index in percent. *Future 4F Alpha* is the future 360-day four-factor alpha of portfolio *i* in *t* in percent. *Future Sharpe* is the future 360-day Sharpe Ratio of portfolio *i* in *t* (negative values adjusted). *SAB*, *SEB* and *SPB* are dummies that equal 1 if the comment of portfolio *i* in *t* is identified as self-attribution biased, self-enhancement biased or self-protective biased, respectively. Trading controls are *Max Return, Turnover, Ln # Transactions and Ln # Securities*. We refer to Table I-A.1 in Appendix I-A for the definition of all variables. t statistics in parentheses: * significant at 10%; ** significant at 5%; *** significant at 1%.

VARIABLES	(1) Future Market Adjusted	(2) Future 4F Alpha	(3) Future Sharpe	(4) Future Market Adjusted	(5) Future 4F Alpha	(6) Future Sharpe	(7) Future Market Adjusted	(8) Future 4F Alpha	(9) Future Sharpe
SEB	-0.480**	-0.611***	-0.007***						
	(-2.422)	(-3.406)	(-2.679)						
SPB				0.411	0.166	0.007			
				(0.764)	(0.300)	(0.568)			
SAB							-0.235	-0.218	-0.014**
							(-0.681)	(-0.793)	(-2.243)
Past Performance	-0.482	-0.586	-0.007	1.307	-1.839	0.035	-0.273	-0.834**	0.024
	(-1.008)	(-1.444)	(-1.126)	(1.323)	(-1.603)	(1.311)	(-0.570)	(-2.266)	(1.204)
Tone	0.011	-0.000	-0.000	-0.008	0.035	0.001	0.022	0.024	0.001
	(0.306)	(-0.008)	(-0.283)	(-0.147)	(0.494)	(0.690)	(0.605)	(0.599)	(1.530)
Readability	0.593	-0.270	-0.008	1.265	0.217	0.061	1.462	0.181	0.009
	(0.754)	(-0.307)	(-0.784)	(0.491)	(0.085)	(0.889)	(1.559)	(0.195)	(0.599)
Length of Comment	-0.040	0.041	0.000	-0.069	-0.040	-0.005**	-0.129	-0.010	-0.002
	(-0.253)	(0.308)	(0.109)	(-0.579)	(-0.187)	(-2.183)	(-0.886)	(-0.084)	(-0.985)
Net Flows	0.000	0.000	0.000	-0.000	-0.000	0.000	0.000	0.000	0.000**
	(0.335)	(0.362)	(1.024)	(-0.333)	(-0.842)	(0.871)	(1.211)	(0.533)	(2.248)
Ln TimeLagComment	0.236**	0.318***	0.003	0.507**	0.465**	0.012	0.280**	0.402***	0.006**
	(2.021)	(3.077)	(1.446)	(2.184)	(2.186)	(1.350)	(2.487)	(4.290)	(2.198)
Ln # Comment	-2.174*	-1.742	-0.039	8.165*	9.431	0.148*	-3.680**	-1.266	-0.024
	(-1.669)	(-1.309)	(-1.365)	(1.901)	(1.608)	(1.743)	(-2.283)	(-0.912)	(-0.614)
Ln Issue Age	1.550	33.314	0.174	-161.170*	100.715	-4.920**	-14.227	47.462	-2.964
	(0.036)	(0.860)	(0.295)	(-1.855)	(1.019)	(-1.999)	(-0.345)	(1.397)	(-1.596)
Ln AUM	-6.662	-12.785	-0.146	54.646*	-36.643	1.493*	-0.957	-17.510	0.871
	(-0.452)	(-0.966)	(-0.704)	(1.799)	(-1.056)	(1.819)	(-0.069)	(-1.508)	(1.423)
Ln Volatility	-1.543	9.268***	0.177***	6.466***	2.829	0.376***	2.769	7.465***	0.279***
	(-0.519)	(2.901)	(3.446)	(2.968)	(0.585)	(4.103)	(1.045)	(3.175)	(3.616)
Trading Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
Heckman Correct.	YES	YES	YES	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Portfolio FE	YES	YES	YES	YES	YES	YES	YES	YES	YES
Observations	8,613	8,613	8,613	2,010	2,010	2,010	10,664	10,664	10,664
Adj. Within R^2	0.261	0.193	0.216	0.221	0.257	0.475	0.274	0.258	0.435

Columns 1 to 3 of Table I-2 show the effect of the self-enhancement bias on future performance over a time horizon of 360 days. We document a statistically significant negative relationship between the self-enhancement bias and future performance over all performance measures. Since we use portfolio fixed effects we infer the following interpretation of our results: when we identify a trader as self-enhancement biased, her social trading portfolio shows a 0.48 (0.61) percentage point lower future market adjusted return (four-factor alpha) than when identified as non-biased. These findings support our hypothesis that the self-enhancement bias leads to future underperformance (H1a). However, columns 4 to 6 point out that the self-protection bias seems not to affect future returns. This finding is in line with literature suggesting that the self-enhancement bias has more impact on trading behavior than the self-protection bias (Fiske and Taylor, 1991; Gervais and Odean, 2001; Miller and Ross, 1975). When we combine the self-enhancement bias and the self-protection bias to the self-attribution bias, we find a statistically significant negative relationship only for the Sharpe ratio (columns 7 to 9). We infer that the self-enhancement bias drives this relationship.

4.2 The self-enhancement bias and future trading behavior

In section 4.1, we find that the self-enhancement bias is negatively correlated with future trading performance. We assume that this relationship can be explained by overconfident trading behavior that is triggered by the self-enhancement bias (H1b). Therefore, we examine the effect of the self-enhancement bias on future trading behavior. We use different variables that are associated with overconfidence in financial literature: number of trades, portfolio turnover, number of different securities, return volatility and extreme returns (Barber and Odean, 2000; Goetzmann and Kumar, 2008; Merkle, 2017). We apply the following panel regression approach to estimate the influence of the self-enhancement bias on traders' overconfidence.

$$Future\ Overconfidence\ Proxy_{v,i,t} = \alpha_v + \beta_v\ SEB_{i,t} + \sum_{j}^{j=J} \gamma_{v,j}\ Control_{j,i,t} + \varepsilon_{v,i,t} \tag{8}$$

We regress different proxies for overconfidence *v (Future Overconfidence Proxy)* of the trader of social trading portfolio *i* on day *t* on the self-enhancement bias dummy (*SEB*) and controls *j* (*Control*). Table I-3 shows the results of our regressions. Following our hypothesis (H1b), the table only includes regression results for the self-enhancement bias.[14]

[14] For the sake of completeness, we repeat this investigation with the self-protection bias instead of the self-enhancement bias. For only one of seven regressions, we find a statistically significant relationship. Consequently, the self-protection bias seems not to be an important driver of overconfidence. This is in line with our assumption in section 2.2.

Table I-3: Regression of future trading variables on self-enhancement bias

This table contains ordinary least squares linear regression results of trading activity variables on *SEB* and a comprehensive set of control variables. We define variables as follows: *SEB* is a dummy that equals 1 if the comment of portfolio *i* in *t* is identified as self-enhancement biased. *Future Ln # Transactions*, *Future Ln # Purchases* and *Future Ln # Sales* are the natural logs of the numbers of transactions, purchases and sales, respectively, of portfolio *i* over the next 90 days. *Future Turnover* is the trading volume of portfolio *i* over the next 90 days divided by the value of portfolio *i* in *t* in percent. *Future Ln # Securities* is the natural log of the average number of securities in portfolio *i* over the next 90 days. *Future Ln Volatility* is the natural log of the future 90-days return volatility of portfolio *i* in percent. *Future Abs Max Return* is the 90-day future maximum absolute daily raw return of portfolio *i* in percent. We refer to Table I-A.1 in Appendix I-A for the definition of all variables. t statistics in parentheses: * significant at 10%; ** significant at 5%; *** significant at 1%.

VARIABLES	(1) Future Ln # Transactions	(2) Future Ln # Purchases	(3) Future Ln # Sales	(4) Future Turnover	(5) Future Ln # Securities	(6) Future Ln Volatility	(7) Future Abs Max Return
SEB	0.033**	0.034**	0.030*	0.179***	-0.004*	-0.002	-0.015
	(2.192)	(2.165)	(1.911)	(2.687)	(-1.684)	(-0.278)	(-0.219)
Past Performance	0.046***	0.061***	0.028***	0.618***	0.000	0.025***	-0.150***
	(5.487)	(6.968)	(3.309)	(17.576)	(0.173)	(5.713)	(-4.260)
Tone	0.002	0.004*	0.001	0.007	-0.001***	0.001	0.009
	(0.868)	(1.814)	(0.719)	(0.837)	(-2.602)	(1.204)	(1.133)
Readability	0.031	-0.018	0.067	0.155	0.019**	0.032	0.192
	(0.621)	(-0.344)	(1.305)	(0.705)	(2.355)	(1.160)	(0.870)
Length of Comment	0.011	0.008	0.023**	0.023	-0.001	0.005	0.043
	(1.318)	(0.860)	(2.533)	(0.585)	(-0.509)	(0.946)	(1.107)
Net Flows	-0.000	-0.000	-0.000**	0.000***	0.000	0.000*	-0.000
	(-1.356)	(-0.242)	(-2.110)	(4.748)	(0.156)	(1.830)	(-0.404)
Ln TimeLagComment	-0.025***	-0.019***	-0.023***	0.017	-0.002	0.001	0.040
	(-4.024)	(-2.856)	(-3.517)	(0.629)	(-1.597)	(0.301)	(1.439)
Ln # Comment	-0.092***	-0.061	-0.034	0.527***	0.024***	0.149***	0.857***
	(-2.595)	(-1.627)	(-0.925)	(3.409)	(4.171)	(7.813)	(5.537)
Ln Issue Age	-4.360***	-5.705***	-3.004***	-57.992***	-0.143	-1.613***	12.387***
	(-5.871)	(-7.334)	(-3.922)	(-18.420)	(-1.239)	(-4.140)	(3.930)
Ln AUM	1.596***	2.068***	1.078***	19.273***	-0.006	0.741***	-4.058***
	(6.303)	(7.795)	(4.123)	(17.888)	(-0.161)	(5.559)	(-3.762)
Ln Volatility	-0.184***	-0.097	-0.169***	-1.539***	-0.011	-0.374***	1.221***
	(-3.148)	(-1.588)	(-2.801)	(-6.005)	(-1.141)	(-11.790)	(4.757)
Trading Controls	YES	YES	YES	YES	YES	YES	YES
Heckman Correction	YES	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES	YES
Portfolio FE	YES	YES	YES	YES	YES	YES	YES
Observations	8,372	8,267	8,226	8,613	8,613	8,613	8,613
Adj. Within R^2	0.0319	0.0379	0.0232	0.0793	0.466	0.0942	0.0484

We find a statistically significant positive relationship between the self-enhancement bias (*SEB*) and diverse overconfidence proxies: number of transactions (*Future Ln # Transactions*), number of purchases (*Future Ln # Purchases*) and number of sales (*Future Ln # Sales*) (each in natural logs) as well as portfolio turnover (*Future Turnove*r). In detail, self-enhancement biased traders execute approximately 3.3% more transactions (column 1) than non-biased traders. Since higher trading frequencies lead to lower trading performance due to transaction costs (Barber and Odean, 2000), this could be one reason for why the self-enhancement bias leads to future underperformance. However, as there are no transaction costs despite the bid-ask spreads, transaction costs on the social trading platform are low in comparison to trading costs on common online brokerage platforms. Nevertheless, traders on the explored platform tend to trade much more frequently in comparison to traders at online brokers (e.g., Glaser and Weber, 2009).

Apart from trading frequencies and volumes, we find a statistically significant negative relationship between the self-enhancement bias and the log numbers of different securities in a social trading portfolio (*Future Ln # Securities*). When a trader is self-enhancement biased, the number of different securities in her portfolio is approximately 0.4% lower. As literature suggests that overconfidence leads to lower diversification (Merkle, 2017), this is another indication for the self-enhancement bias triggering overconfidence.

Unlike Dorn and Huberman (2005), we find no evidence that the self-enhancement bias leads to higher return volatilities or to more extreme returns (columns 7 and 8). In contrast to our study, the authors use survey data to identify biased traders. Additionally, the authors do not include portfolio or time fixed effects. We suggest that these differences in the study design could explain different results.

Overall, our findings are in line with our hypothesis H1b. Along with our results from section 4.1, empirical evidence supports the theoretical multi-period market model developed by Gervais and Odean (2001).[15]

4.3 The self-enhancement bias and investment flows

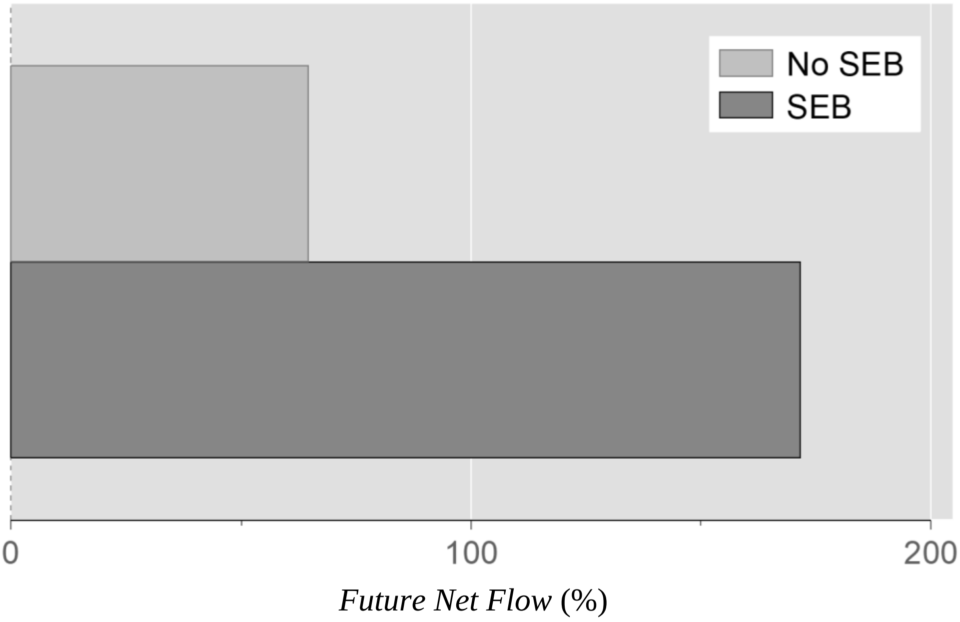
Hitherto, we focused on the effect of the self-enhancement bias on the trader. In this section, we examine if traders' self-enhancement biased behavior also affects their investors. Therefore, we investigate if the self-enhancement bias of a trader is related to flows to the trader's social trading portfolio (more precisely: flows to the structured product that replicates the performance of the underlying social trading portfolio). We hypothesize that traders attract higher investment flows when they are prone to the self-enhancement bias (H2). Figure I-5 shows percentage net flows (*Net Flows*) to social

[15] The examination of the relationship between overconfidence and the self-attribution bias could raise reverse causality issues. Therefore, we estimate regressions using our measures for the self-enhancement bias, the self-protection bias and the self-attribution bias as dependent variables and include past overconfidence measures as independent variables. Evidence suggests that overconfidence does not trigger the self-attribution bias.

trading portfolios, distinguishing between self-enhancement biased (*SEB*=1) and unbiased (*SEB*=0) traders.

Figure I-5: Mean of future relative net flows

This figure illustrates the univariate relationship between self-enhancement bias (*SEB*) and future percentage net flows (*Future Net Flow*) into or out of the portfolio over the next 90 days. We only include observations that show positive *Past Performance* (*SEB=1* and *SEB=0*). The difference between *SEB* group and *No SEB* group is statistically significant at the 1% level.



Results suggest that portfolios attract significantly higher investment flows when the trader is self-enhancement biased. We use the following panel regression framework to examine this relationship in more detail:

$$Flow\ Variable_{k,i,t} = \alpha_k + \beta_k\ SEB_{i,t} + \sum_{j}^{j=J} \gamma_{k,j}\ Control_{j,i,t} + \varepsilon_{k,i,t} \tag{9}$$

We regress the different flow variables *k* (*Flow Variable*) to the structured product of the social trading portfolio *i* on day *t* on the self-enhancement bias dummy (*SEB*) and controls *j* (*Control*). We estimate regressions over different time horizons of future flows, namely 90 days, 180 days and 360 days beginning on the day after the comment. We examine net flows (*Net Flows*) as well as inflows (*Inflows).*

Table I-4: Regression of future investment flows on self-enhancement bias

This table shows ordinary least squares linear regression results of future inflows as well as of future net flows on *SEB* and a comprehensive set of control variables. We define variables as follows: *SEB* is a dummy that equals 1 if the comment of portfolio *i* in *t* is identified as self-enhancement biased. *Future Net Flows* is the sum of inflows minus the sum of outflows to/out of (the structured product of) portfolio *i* over the last 90, 180 and 360 days, respectively, divided by the invested money to portfolio *i* in *t* in percent. *Future Inflows* is the sum of inflows to (the structured product of) portfolio *i* over the last 90,180 and 360 days, respectively, divided by the invested money to portfolio i in t in percent. Trading controls are *Max Return*, *Turnover*, *Ln # Transactions* and *Ln # Securities*. *Ln Issue Age* is multiplied by 100. We refer to Table I-A.1 in Appendix I-A for the definition of all variables. t statistics in parentheses: * significant at 10%; ** significant at 5%; *** significant at 1%.

VARIABLES	(1) Future Net Flows (90d)	(2) Future Net Flows (180d)	(3) Future Net Flows (360d)	(4) Future Inflows (90d)	(5) Future Inflows (180d)	(6) Future Inflows (360d)
SEB	15.665*	19.418**	2.492	23.401**	43.061**	35.202*
	(1.954)	(2.121)	(0.435)	(2.432)	(2.156)	(1.742)
Past Performance	37.355***	28.741***	10.133***	39.515***	65.070***	51.606***
	(8.013)	(5.442)	(3.113)	(7.061)	(5.647)	(4.495)
Tone	-1.035	-1.045	-0.480	-1.651	-3.050	-3.095
	(-0.998)	(-0.919)	(-0.668)	(-1.326)	(-1.229)	(-1.221)
Readability	0.415	-1.521	8.590	12.176	22.027	50.065
	(0.015)	(-0.048)	(0.440)	(0.357)	(0.317)	(0.727)
Length of Comment	5.662	4.822	4.539	7.012	9.577	14.686
	(1.260)	(0.955)	(1.417)	(1.300)	(0.869)	(1.300)
Net Flows	0.001***	0.001*	0.000	0.001***	0.002**	0.001
	(3.548)	(1.725)	(0.618)	(3.094)	(2.134)	(1.357)
Ln TimeLagComment	4.333	4.785	3.110	1.247	5.507	6.822
	(1.247)	(1.224)	(1.287)	(0.299)	(0.646)	(0.801)
Ln # Comment	-11.662	-30.308	-35.366***	-7.123	-37.393	-133.492***
	(-0.612)	(-1.403)	(-2.657)	(-0.311)	(-0.793)	(-2.843)
Ln Issue Age	-30.228***	-23.856***	-9.361***	-31.195***	-53.364***	-42.062***
	(-7.311)	(-5.080)	(-3.237)	(-6.285)	(-5.208)	(-4.124)
Ln AUM	898.211***	568.302***	36.329	912.511***	1,466.859***	851.767**
	(6.321)	(3.531)	(0.367)	(5.349)	(4.177)	(2.439)
Ln Volatility	-38.753	-36.413	24.619	-43.969	-47.726	25.528
	(-1.192)	(-0.987)	(1.063)	(-1.127)	(-0.593)	(0.312)
Trading Controls	YES	YES	YES	YES	YES	YES
Heckman Correction	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Portfolio FE	YES	YES	YES	YES	YES	YES
Observations	6,742	7,382	7,771	6,742	7,382	7,771
Adj. Within R^2	0.0860	0.127	0.262	0.0753	0.0681	0.132

In Table I-4, we still find a statistically significant relationship between the self-enhancement bias and investment flows for five of six regressions. When a trader is self-enhancement biased, she receives 15.67 (19.42) percentage points higher *Net Flows* in proportion to assets under management in the next 90 days (180 days). On a time horizon of 360 days, results for *Net Flows* are not statistically significant any more (column 3).

However, it is not surprising that investors rather react on public comments in the short run than in the long run.[16]

Overall, traders seem to attract more net flows and inflows when they are prone to the self-enhancement bias. This finding supports our hypothesis (H2) and is in line with literature suggesting that the confidence of an individual strengthens its social status or perceived level of knowledge and trustworthiness (Anderson et al., 2012; Price and Stone, 2004). However, as traders prone to the self-enhancement bias subsequently underperform, the self-enhancement bias does not only harm the affected trader but also her investors.

5. Conclusion

Our study of biased self-attribution among nonprofessional traders contributes to financial literature in at least three ways. First, we show that biased self-enhancement leads to future underperformance, while the self-protection bias seems not to harm traders' performance. We recommend future studies that examine the self-attribution bias to measure the self-enhancement bias and the self-protection bias separately, as well. Second, results suggest that overconfidence arising from biased self-enhancement is a possible driver of traders' future underperformance: self-enhancement biased traders develop overconfident trading behavior such as higher trading frequencies and portfolio turnover as well as lower portfolio diversification. Although Gervais and Odean (2001) assume this connection in their theoretical model, our paper is the first empirical study supporting their hypothesis. Further studies regarding the role of the self-attribution bias as well as the resulting overconfident behavior among professional traders and portfolio managers would be worthwhile. Third, we find that self-enhancement biased behavior does not only harm the affected traders' performance but also possible investors: Traders being self-enhancement biased receive more investment inflows in comparison to being non-biased. It follows that the self-enhancement bias is not only harmful to affected individuals but

[16] Again, we repeat this investigation with the self-protection bias instead of the self-enhancement bias. We find no statistically significant relationship between the self-protection bias and future investment flows to or out of the social trading portfolio. The result suggests that investors do neither prefer nor disfavor a portfolio manager when she shows signs of the self-protection bias.

also for a third party. Here, future research should be undertaken to explore whether this issue also affects institutional investors.

In our main regressions, we use fixed portfolio effects, fixed time effects as well as double-clustered standard errors. Besides, we show robustness by using various performance measures. Furthermore, we control for a possible sample selection bias by using the Heckman (1979) correction.

One could argue that the use of personal pronouns in comments is not driven by the self-attribution bias but by self-marketing of the traders. However, according to literature the self-attribution bias is also driven by self-presentation and the purpose to portray oneself positively to others (Arkin et al., 1980; Schlenker, 1980; Shepperd et al., 2008). Consequently, traders who intentionally use an abnormal number of first person personal pronouns in their comments after past trading success are (per definition) also self-attribution biased. Notwithstanding, one limitation of our study is that we are not able to separate traders that intentionally use an abnormal number of first person pronouns from those who do it unintentionally. In future investigations, it might be possible to use a different approach or setting in which the self-attribution bias will be isolated from trader's self-marketing.

Altogether, the self-enhancement bias and connected overconfidence negatively influences the performance of nonprofessional traders and even investors that interact with them. The findings of this study have a number of implications. With regard to traders, a reasonable approach to tackle biased assessment of past performance could be to increase traders' financial literacy. Providing any type of financial education by policy makers might diminish behavioral biases that affect investing decisions. With regard to investors, our findings also suggest several courses of action. When investors use delegated portfolio management services, they should not be deflected by a managers' self-presentation, but rather assess mangers' capabilities as well as past performance in a more comprehensive way by mainly including quantitative and objective measurements into their assessment. Again, financial education that increase investors' financial literacy might be expedient.

6. References

Ammann, M., & Schaub, N. (2017). The impact of internet postings on individual investors, in American Finance Association 2018 Annual Meeting, available online: https://pdfs.semanticscholar.org/e77f/b59def2418c9e164eadd68d9e73fefd271f4.pdf

Anderson, C. A., & Slusher, M. P. (1986). Relocating motivational effects: a synthesis of cognitive and motivational effects on attributions for success and failure. *Social Cognition*, *4*(3), 270–292.

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735.

Arkin, R. M., Appelman, A. J., & Burger, J. M. (1980). Social anxiety, self-presentation, and the self-serving bias in causal attribution. *Journal of Personality and Social Psychology*, *38*(1), 23–35.

Baginski, S. P., Hassel, J. M., & Kimbrough, M. D. (2004). Why do managers explain their earnings forecasts? *Journal of Accounting Research*, *42*(1), 129–142.

Baginski, S. P., Hassell, J. M., & Hillison, W. a. (2000). Voluntary causal disclosures: tendencies and capital market reaction. *Review of Quantitative Finance and Accounting*, *15*, 371–389.

Bannier, C., Pauls, T., & Walter, A. (2019). Content analysis of business communication: Introducing a German dictionary. *Journal of Business Economics*, *89*(1), 79–123.

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, *55*(2), 773–806.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261–292.

Bettman, J. R., & Weitz, B. A. (1983). Attributions in the board room: Causal reasoning in corporate annual reports. *Administrative Science Quarterly*, *28*(2), 165–183.

Billett, M. T., & Qian, Y. (2008). Are overconfident CEOs born or made? Evidence of self-attribution bias from frequent acquirers. *Management Science*, *54*(6), 1037–1051.

Bjornsson, C. H. (1968). *Lesbarkeit durch Lix*. Stockholm: Pedagogiskt Centrum.

Brien, M. J., & Panis, C. W. A. (2015). *Brokerage accounts in the united states*.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics*. Cambridge University Press: New York.

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*(1), 23–43.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, *52*(1), 57–82.

Chance, Z., & Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, *6*, 104–107.

Chen, G., Kim, K. A., Nofsinger, J. R., & Rui, O. M. (2007). Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. *Journal of Behavioral Decision Making*, *20*(4), 425–451.

Clifford, C. P., Fulkerson, J. A., Jordan, B. D., & Waldman, S. R. (2013). *Risk and fund flows*.

Doering, P., Neumann, S., & Paul, S. (2015). *A primer on social trading networks – institutional aspects and empirical evidence*.

Dorn, D., & Huberman, G. (2005). Talk and action: What individual investors say and what they do. *Review of Finance*, *9*(4), 437–481.

Doukas, J. A., & Petmezas, D. (2007). Acquisitions, overconfident managers and self-attribution bias. *European Financial Management*, *13*(3), 531–577.

Eling, M., & Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking and Finance*, *31*(9), 2632–2647.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition (2nd ed.)*. New York: McGraw-Hill.

French, K. R. (2017). Retrieved from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

Fung, W., Hsieh, D. A., Naik, N. Y., & Ramadorai, T. (2008). Hedge funds: Performance, risk, and capital formation. *Journal of Finance*, *63*(4), 1777–1803.

Gervais, S., & Odean, T. (2001). Learning to be overconfident. *Review of Financial Studies*, *14*(1), 1–27.

Glaser, M., & Weber, M. (2007). Overconfidence and trading volume. *GENEVA Risk and Insurance Review*, *32*(1), 1–36.

Glaser, M., & Weber, M. (2009). Which past returns affect trading volume? *Journal of Financial Markets*, *12*(1), 1–31.

Goetzmann, W. N., & Kumar, A. (2008). Equity portfolio diversification. *Review of Finance*, *12*(3), 433–463.

Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *The Review of Financial Studies*, *23*(7), 2821–2864.

Hastorf, A. H., Schneider, D. J., & Polefka, J. (1970). *Person perception*. Addison Wesley,Reading, Mass.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Hilary, G., & Menzly, L. (2006). Does past success lead analysts to become overconfident? *Management Science*, *52*(4), 489–500.

Hoffmann, A. O. I., & Post, T. (2014). Self-attribution bias in consumer financial decision-making: How investment returns affect individuals' belief in skill. *Journal of Behavioral and Experimental Economics*, *52*, 23–28.

Huang, J., Wei, K. D., & Yan, H. (2007). Participation costs and the sensitivity of fund flows to past performance. *Journal of Finance, 62*(3), 1273–1311.

Israelsen, C. (2005). Refining the Sharpe ratio. *Journal of Asset Management*, *5*(6), 423–427.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Kim, Y. H. (Andy). (2013). Self attribution bias of the CEO: Evidence from CEO interviews on CNBC. *Journal of Banking and Finance*, *37*(7), 2472–2489.

Kothari, S. P., Li, X., & Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, *84*(5), 1639–1670.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Li, F. (2010). *Managers' self-serving attribution bias and corporate financial policies*.

Libby, R., & Rennekamp, K. (2012). Self-serving attribution bias, overconfidence, and the issuance of management forecasts. *Journal of Accounting Research*, *50*(1), 197–231.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65.

McCannon, B. C., Asaad, C. T., & Wilson, M. (2016). Financial competence, overconfidence, and trusting investments: Results from an experiment. *Journal of Economics and Finance*, *40*(3), 590–606.

Merkle, C. (2017). Financial overconfidence over time: Foresight, hindsight, and insight of investors. *Journal of Banking and Finance*, *84*, 68–87.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *Journal of Finance*, *53*(6), 1887–1934.

Oehler, A., Horn, M., & Wendt, S. (2016). Benefits from social trading? Empirical evidence for certificates on wikifolios. *International Review of Financial Analysis*, *46*, 202–210.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57.

Röder, F., & Walter, A. (2019). What drives investment flows into social trading portfolios? *The Journal of Financial Research*, 42(2), 383-411.

Rogers, J. L., Buskirk, A. Van, & Zechman, S. L. C. (2011). Disclosure tone and shareholder litigation. *The Accounting Review*, *86*(6), 2155–2183.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations*. Monterey, California: Brooks/Cole Publishing Company.

Sendén, M. G., Lindholm, T., & Sikström, S. (2014). Biases in news media as reflected by personal pronouns in evaluative contexts. *Social Psychology*, *45*(2), 103–111.

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring causes of the self-serving bias. *Social and Personality Psychology Compass*, *2*, 895–908.

Sirri, E. R., & Tufano, P. (1998). Costly search and mutual fund flows. *Journal of Finance*, *53*(5), 1589–1622.

Tetlock, P. E., & Levi, A. (1982). Attribution bias: on the inconclusiveness of the cognition - motivation debate. *Journal of Experimental Social Psychology*, *18*, 68–89.

Twedt, B., & Rees, L. (2012). Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy*, *31*(1), 1–21.

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *The British Journal of Social Psychology*, *44*(3), 443–461.

Veld, C., & Veld-Merkoulova, Y. V. (2008). The risk perceptions of individual investors. *Journal of Economic Psychology*, *29*(2), 226–252.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. *Diagnostica*, *54*(2), 85–98.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. *Journal of Personality*, *47*(2), 245–287.

7. Appendix

Appendix I-A: Descriptions of variables

Table I-A.1 contains descriptions and construction details of all variables used in this paper.

Table I-A.1: Descriptions of variables

This table contains descriptions and construction details of all variables used in this paper.

Panel A: Comments	
Self Ref	Quotient of the number of first person personal pronouns (category "Self" in the LIWC) minus the number of third person personal pronouns (category "Other" in the LIWC) and the overall number of words in the comment of portfolio *i* in *t* in percent.
Length of Comment	The average number of words in the comment of portfolio *i* in *t*.
Tone	Difference of positive and negative words as classified by Bannier, Pauls, and Walter (2019) relative to the overall number of words within the comment of portfolio *i* in *t* in percent.
Readability	Readability measure of the comment of portfolio *i* in *t* following Bjornsson (1968) divided by 100. It is defined as the average number of words per sentence plus the percentage of words with more than six letters.
TimeLagComment	Days since the last comment of portfolio *i*.
# Comment	Number of comments for portfolio *i* until day *t*.
Panel B: Portfolio Data	
Past Performance	360-day raw return of portfolio *i* in *t* (we exclude outliers above the 97.5 percentile and under the 2.5 percentile) in percent.
Market Adjusted	360-day raw return of portfolio *i* in *t* minus the 360-day return of the MSCI World index in *t* in percent.
Sharpe	360-day Sharpe Ratio of portfolio *i* in *t*; negative values adjusted as suggested by Israelsen (2005).
4F Alpha	360-day four-factor alpha of portfolio *i* in *t* in percent.
Volatility	360-day return volatility of portfolio *i* in *t* in percent.
Inflows	Sum of inflows to (the structured product of) portfolio *i* over the last 360 days divided by the invested euros in portfolio *i* in *t-360* in percent.
Net Flows	Sum of inflows minus the sum of outflows to/out of (the structured product of) portfolio *i* over the last 360 days divided by the invested euros in portfolio *i* in *t-360* in percent.
Issue Age	Age (since issue of the structured product) of portfolio *i* on day *t* in years.
AUM	Assets under management (i.e., invested euros in the structured product) of portfolio *i* on day *t*.
Panel C: Self-Attribution Bias Dummies	
SAB	Dummy equaling 1 if the comment of portfolio *i* on day *t* is identified as self-attribution biased, and zero otherwise.
SEB	Dummy equaling 1 if the comment of portfolio *i* on day *t* is identified as self-enhancement biased, and zero otherwise.
SPB	Dummy equaling 1 if the comment of portfolio *i* on day *t* is identified as self-protection biased, and zero otherwise.
Panel D: Overconfidence Proxies	
# Transactions	Number of transactions of portfolio *i* over the last 90 days.
# Purchases	Number of purchases of portfolio *i* over the last 90 days.
# Sales	Number of sales of portfolio *i* over the last 90 days.
Turnover	Trading volume of portfolio *i* over the last 90 days divided by the value of the virtual portfolio *i* in *t* in percent.
# Securities	Average number of securities in portfolio *i* over the last the 90 days.
Max Return	Maximum absolute daily raw return of portfolio *i* over the last 90 days in percent.

Appendix I-B: Heckman correction

In our study, we focus on traders' comments. However, not all traders write comments on the social trading platform. While some traders write comments frequently, others do not write comments at all. In result, a sample selection bias could occur. We apply the two-stage Heckman (1979) correction to address this issue. First, we estimate a probit model regression of determinants of writing a comment. For this regression, the dependent variable equals one if the trader of portfolio *i* writes a comment on day *t*, and zero otherwise. Hence, sample size is much higher than in our main results. Second, we use hazard rates obtained from this regression as a control variable in all following regressions.

Table I-B.1: Determinants of writing a comment (Heckman correction)

This table shows the probit regression results of the first stage of the two-stage Heckman (1979) selection model. The dependent variable is *Dummy Comment* equaling one if the trader of portfolio *i* writes at least one comment on day t and zero otherwise. We refer to Table I-A.1 in Appendix I-A for a definition of the other variables. We use the hazard rates of this regression as control variable in all regressions in our main results to correct for a potential sample selection bias. t statistics in parentheses: * significant at 10%; ** significant at 5%; *** significant at 1%.

	(1)
VARIABLES	Dummy Comment
Past Performance	0.468***
	(3.562)
Net Flows	0.002***
	(2.682)
Ln Issue Age	-0.409***
	(-8.601)
Ln Volatility	0.004
	(0.138)
Ln AUM	0.139***
	(10.422)
Pseudo R^2	0.0688
Observations	762,760

Results in Table I-B.1 indicate that the probability of writing a comment is positively related to past portfolio performance, past net flows as well as assets under management. This is in line with Ammann and Schaub (2017) who find that successful social traders are more likely to write comments. Furthermore, portfolio age is negatively related to the probability of writing a comment. Given that, traders seem to be more motivated to write comments at the beginning of their social trading career.

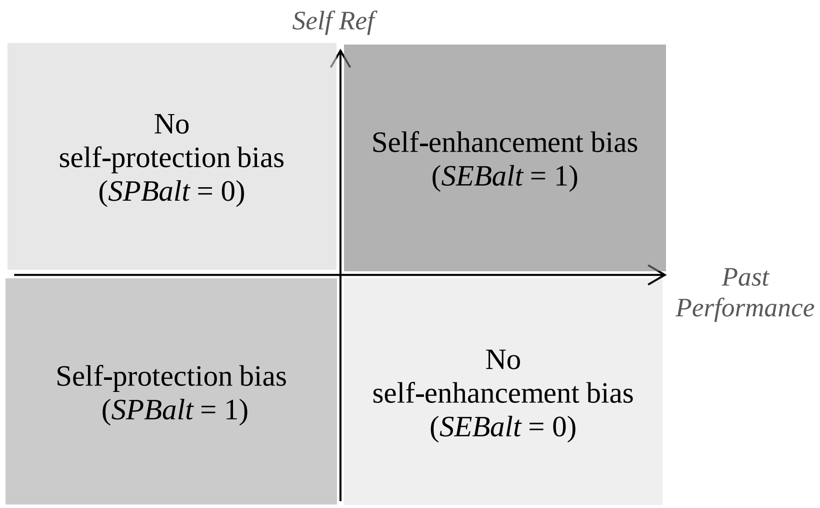
Appendix I-C: An alternative measure for biased self-attribution

In this appendix, we use an alternative way to identify traders being prone to the self-attribution bias. We do so because the definition of the self-attribution bias differs among different sources. While Gervais and Odean (2001, p. 1) define biased self-attribution as the behavior to "overestimate the degree, to which we are responsible for our own success", Hastorf et al. (1970, p. 73) describe it as the "attributing success to our own dispositions". The main difference is that the first definition states that there is a justified level to which past success is attributable to oneself. From this perspective, only people showing a level of self-attribution higher than justified are classified as self-attribution biased. In the second definition, however, any attribution of past success to oneself is attributed to biased self-attribution. While the measure for the self-attribution bias in our main specification relies on the first definition, we use a measure for the self-attribution bias based on the second definition in this appendix.

Following the second definition of the bias, every positive value of *Self Ref* that follows on positive past returns is associated to biased self-attribution (self-enhancement bias). Additionally, every negative value of *Self Ref* following on negative past returns is associated to biased self-attribution, as well (self-protection bias). See Figure I-C.1 for a graphical presentation.

Figure I-C.1: Construction of alternative measures of biased self-attribution

Every positive value of *Self Ref* that follows on positive past returns is associated to self-enhancement bias (*SEBalt=1*). Additionally, every negative value of *Self Ref* that follows on negative past returns is associated to the self-protection bias (*SPBalt=1*).



Based on these new measures for biased self-attribution, self-enhancement and self-protection, we re-estimate our main regressions presented in tables I-2 to I-4. Overall, results qualitatively remain the same and are available upon request. Overall, results qualitatively remain the same and are available upon request.

Appendix I-D: Examples of comments on the investigated social trading platform

To illustrate our approach of calculating self-referencing within a trader's comment, we provide two examples from our sample.[17] We highlight all first person personal pronouns as bold and underline all third person personal pronouns. Additionally, we calculate the self-referencing measure for each example the way it is done in our paper.

$$Self\ Ref_{i,t} = 100 * \frac{Number\ Self_{i,t} - Number\ Other_{i,t}}{Number\ Words_{i,t}} \tag{D.1}$$

Example 1:

__I__ am very happy about the performance of __my__ portfolio. Although there was a noticeable loss with Tesla to be absorbed, <u>it</u> had only a small impact on the overall portfolio due to a weighting of only 3%. __My__ stabilizers work the way __I__ wanted, although __I__ am aware that Dialog Semiconductor is quite speculative as a stabilizer. Since __I__ am really convinced by the security, __I__ see more opportunities than risks, especially after the noticeable share price loss before __my__ initial investment.

→ $Self\ Ref = 100 * \frac{8-1}{83} = 8.43$

Example 2:

Today also, markets are drifting down. <u>They</u> remain under pressure since political events as well as fear of a new wave of bank and government bankruptcies (Argentina, France, Netherlands) spread fear and panic. Currently, <u>it</u> seemed to __me__ that the DAX has developed a massive weakness. Compared with the DOW, the DAX still shows underperformance.

→ $Self\ Ref = 100 * \frac{1-2}{56} = -1,79$

[17] We translated the examples from German to English.

II. Signaling in initial coin offerings - the key role of entrepreneurs' self-efficacy and media presence

Co-author: Florian Röder
Own share: 50%

A previous version of this paper has been presented at the following conference:

- Cryptocurrency Research Conference, 2019, Southampton, UK.

Signaling in initial coin offerings - the key role of entrepreneurs' self-efficacy and media presence

Daniel Czaja[a] Florian Röder[b]

Abstract - By analyzing data of more than 1,000 Initial Coin Offerings (ICOs) obtained from seven different ICO information platforms, we investigate the effectiveness of projects' quality signals (human capital, entrepreneurs' self-efficacy, ambiguity reduction and level of media presence) with regard to ICO funding success. Results imply that media presence and entrepreneurs' self-efficacy are effective signals of project quality in the ICO market and thus, can foster funding success. Project initiators that communicate (more actively) via social media collect more funds than those who do not. Analogously, entrepreneurs appearing self-efficacious with regard to the quality of their venture receive more funds.

Keywords: Initial Coin Offering, token sale, cryptocurrency, signaling, crowdsale

JEL-Codes: L26, G11, M13

[a] Institute of Financial Services, University of Giessen, Daniel.L.Czaja@wirtschaft.uni-giessen.de

[b] Institute of Financial Services, University of Giessen, Florian.R.Roeder@wirtschaft.uni-giessen.de

1. Introduction

The development of the Initial Coin Offering (ICO) market in the recent years highlights its increasing importance for entrepreneurs and investors. Raised funds increased from less than 0.03 billion USD in 2015 to more than 15 billion USD in the first half of 2018 (EY, 2018). The popularity of ICOs virtually led to a hype among investors. The announcement of an ICO by the traditional photography company Kodak in January 2018[1], for example, suddenly increased Kodak's popularity among investors. By the end of the announcement day, Kodak's stock price jumped up by about 120 percent. Hence, the use of an ICO for capital formation seems to be a positive signal for potential investors per se. Nevertheless, the success of ICOs highly varies among the different projects: while some ICOs attract several hundred millions of USD, many are not able to raise any funds at all.[2] Therefore, investors seem to distinguish between the different projects. However, given the fact that ICO investors cannot directly observe project quality, ICO projects need to effectively signal venture quality to encourage investors to participate in the token sale.

Surprisingly, a systematic understanding of effective quality signals in the ICO context is still lacking. Therefore, this paper provides theoretical considerations of different quality signals in the context of ICOs and evaluates the effectiveness of different quality signals for ICO funding success. More specifically, we investigate the relevance of venture quality (in terms of human capital), level of uncertainty (in terms of entrepreneurs' self-efficacy and ambiguity reduction) and the level of familiarity (in terms of media presence) for ICO founding amount.

Our study makes a major contribution to research on early stage financing by providing theoretical considerations and empirical evidence of the effectiveness of various types of signals that are sent out by ICO initiators. A large body of literature has examined the association between information provided by entrepreneurs and investors' funding in different contexts of fundraising. Based on the signaling theory (Spence, 1973), previous studies provide empirical evidence which signals are effective in the sense of fostering

[1] See https://kodakone.com/press/detail-page/kodak-and-wenn-digital-partner-to-launch-major-blockchain-initiative-and-cryptocurrency.html.

[2] See our summary statistics in Table II-1.

investments of potential investors (Ahlers et al., 2015; Ahlstrom and Bruton, 2006; Cosh et al., 2009; Kromidha and Li, 2019; Prasad et al., 2000). However, the transferability of previously published research on entrepreneurial signaling to ICOs is problematic as every method of capital formation has its own idiosyncrasies (Barbi and Mattioli, 2019; Duffner et al., 2009; Giudici et al., 2018). Therefore, it is necessary to investigate the capital formation process of ICOs separately.

An ICO can be defined as a crowdsale that takes place on a blockchain. In particular, project initiators generate so-called tokens on a blockchain and then usually sell those tokens to investors in exchange for other established cryptocurrencies such as Ethereum or Bitcoin. Issued tokens grant purchasers a utility of some kind (e.g., access to a future good or service) or a share of a future cash flow generated by the issuing ICO project. In general, the emerging crypto market is characterized by both, low regulation and high information asymmetries. While some countries, such as China, banned ICOs entirely (PBC, 2017), national legislations of other countries, such as USA and Switzerland, assess token sales on a case-by-case basis (FINMA, 2018; SEC, 2017). Other countries, such as Russia, in turn, seem to foster ICOs by hardly regulating the ICO market at all (MinFin, 2018).

To address our research question, we use data of more than 1,000 ICOs that we identify on seven popular online ICO information platforms.[3] We obtain data on raised funds from those ICO information platforms to assess ICO funding success. Additionally, to avoid potential reverse causality issues, we collect data on human capital, ambiguity reduction and entrepreneurs' self-efficacy directly from ICOs' white papers that were published before the actual ICO event. Those white papers provide information about the underlying project to potential investors. Moreover, we gather data on projects' media presence before ICO from eight different social media platforms. Our final sample covers the period from July 2014 to January 2018 and includes an ICO funding volume of approximately 8.7 billion USD on aggregate.

Results suggest that both, the level of ICOs' media presence and entrepreneurs' self-efficacy, are positively related to ICO funding success and thus are effective signals for

[3] More precisely, the most popular ICO platforms are icodata.io, icotracker.net, icobazaar.com, tokendata.io, icobench.com, smithandcrown.com and icodrops.com.

project quality. More precisely, ICOs that are accompanied by the usage of various social media channels as well as high social media activity receive more funds from investors. Second, ICO characteristics that signal entrepreneurs' self-efficacy, such as low bonus granted, high share of tokens retained by the ICO initiators and short token sale period, are positively related to ICO funding amount. In contrast, our results do not provide evidence in favor of a clear benefit of projects' human capital and ambiguity reduction in the promotion of ICO funding success.

We apply a test proposed by Oster (2019) on whether the presence of omitted variables could bias our main results. The findings of this test show that our results seem not to be driven by omitted attributes and characteristics of ICOs that are not captured in our main model specification.

2. Institutional framework

2.1 Distributed ledger technology, blockchain and initial coin offerings

The Distributed Ledger Technology (DLT) refers to an emergent database concept. In particular, data is consensually recorded and shared across multiple data stores known as ledgers. As all ledgers have to contain the same data records, new additions to data by members (nodes) of this distributed network are recorded on each ledger eliminating the need for a central authority (Yu et al., 2018). In detail, each independent ledger update is shared in the underlying peer-to-peer network and then, to ensure validity of a new entry (i.e., to prevent simultaneous transactions on the same asset or to prevent cyber-attacks, such as distributed denial-of-service attacks), a consensus algorithm is used. Each distributed ledger network has its predefined cryptographic validation method. Once a consensus is reached, all nodes add this new entry to their ledger. Thus, each node has an identical copy of the entire data at any point in time.[4] Moreover, distributed ledgers can be distinguished by two features. First, distributed ledgers are either permissionless or permissioned (Trump et al., 2018). While in permissioned networks nodes need a permission from the responsible entity (i.e., the creator of the distributed ledger) to change ledger entries, data updates in permissionless networks are allowed in principle. Second, distributed ledgers can be differentiated concerning the access to the network. In detail,

[4] See Natarajan et al. (2017) for a more comprehensive description.

while anyone can access public ledgers, private ledgers are only accessible by approved nodes.

Blockchain is a specific type of the DLT. It is the underlying technology used by the vast majority of projects conducting an ICO. Blockchain is characterized by an append-only data structure (i.e., ledgers can only be altered by extension) that exists in the form of a chain of blocks. The key feature of the blockchain technology is the implementation of cryptography. Every new addition (block) to the digital ledger that stores information about transactions, for instance information concerning time, money amount or transaction partners, is ‘hashed’ (Natarajan et al., 2017). More specifically, a cryptographic hash function transforms information about transactions to a bit string of fixed size (hash) by applying a mathematical algorithm (Halevi and Krawczyk, 2006). As the hash function is non-invertible, subsequent modifications of the information about transactions results in a different hash and therefore, manipulations are easy to detect. Every block contains the hash of the previous block, information on the considered transaction and an additional timestamp. As a result, a chain of blocks is formed. Thus, given blocks cannot be altered ex post without altering all subsequent blocks of the chain.

Recently, a considerable number of new ventures employs the blockchain technology for capital formation. Known as ICOs, mainly technology startups generate and sell so-called tokens via blockchain in exchange for traditional fiat money or established cryptocurrencies, such as Bitcoin and Ethereum (Roosenboom et al., 2020). More precisely, tokens are entries on a blockchain. ICO initiators determine token amount, token value and other special conditions (e.g., a bonus scheme for early investors). Then, ICO initiators sell the generated tokens in a predetermined ICO period. All terms and conditions as well as the automatic execution of the token sale are implemented in so-called smart contracts. More specifically, when an investor transfers money to the ICO project’s digital address, i.e. node in a blockchain, she automatically receives an amount of tokens in accordance with the smart contract’s terms and conditions. As described above, all transactional data is stored in the underlying blockchain. The creation and sale of tokens takes place either on an existing blockchain, such as Ethereum, which is most common for ICOs, or on a new blockchain that is especially created for the ICO.

Distributed tokens usually offer an incentive for investors. According to the type of incentive, there is a distinction between so-called 'utility tokens', 'security tokens' and 'currency tokens' (Ante et al., 2018; Howell et al., 2019). The first one represents some form of utility that is granted to token holders, i.e. access to future products or services of the ICO project. Typically, either only token holders can use ICO project's future products or services or if basic features of the services are accessible to everyone, some additional premium features of the services are exclusively available for token holders. Security tokens, on the other hand, are comparable to stocks or bonds and represent a share of the ICO project or a claim on future ICO project's cash flows. However, the profit-sharing mechanism lacks a legal basis, which makes it basically impossible for investors to assert any legal claims. Lastly, some tokens neither represent utility nor profit claims but instead, solely function as digital currencies (currency token).

2.2 The process of initial coin offerings

The starting point of a typical ICO is the preparation of a white paper. A white paper is a document written by the ICO initiators that usually promotes and explains the underlying products or services, introduces the project team and describes a business plan. Additionally, it mostly includes token sale characteristics, such as token amount, distributed share of tokens, sale period, possible bonus schemes, as well as a description on how collected funds will be used (Adhami et al., 2018; Chen, 2019).

Simultaneously to the release of a white paper or shortly thereafter, ICO initiators use social media, especially the platforms of Twitter and BitcoinTalk[5] to promote their project. The first social media presence constitutes the starting point of a marketing campaign that typically lasts until the end of the token sale period. Usually, marketing activities include almost exclusively activities on social media channels, such as presenting project's updates, images and videos as well as communication with potential investors. Moreover, many initiators introduce their project on ICO information platforms. Typically, all marketing activities of an ICO are limited to online channels.

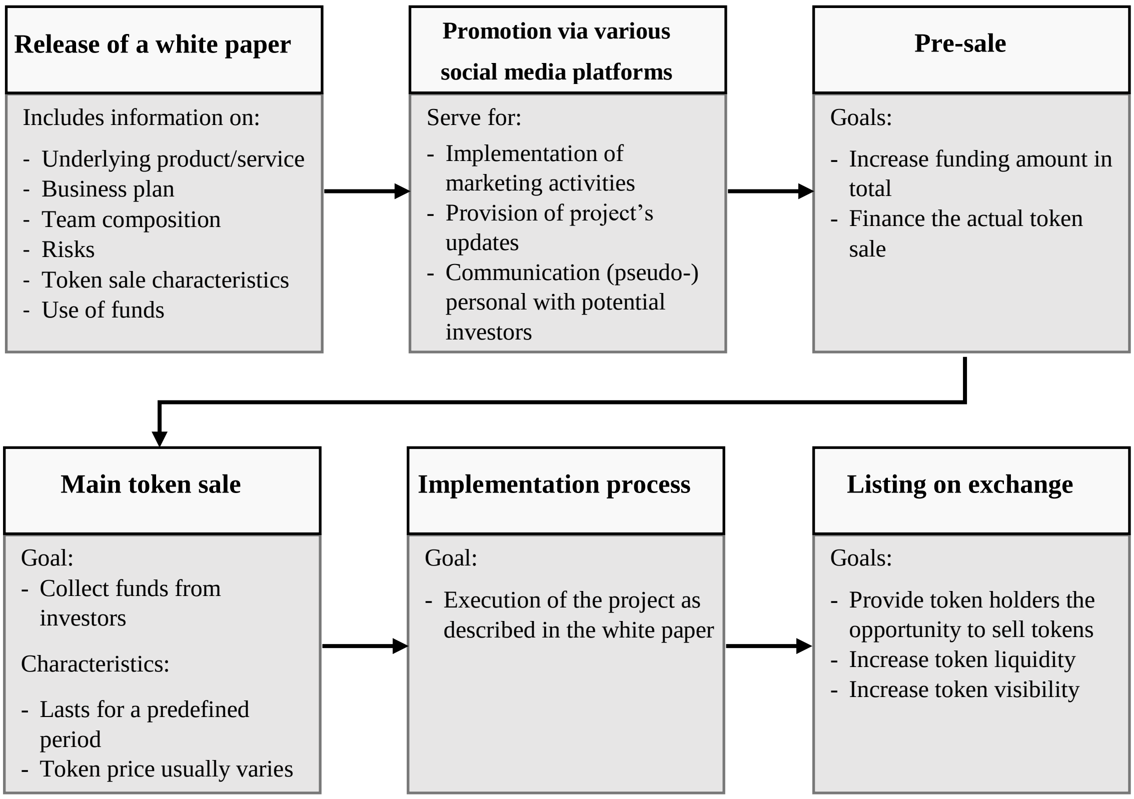
Many ICO initiators prepend a so-called pre-sale (or pre-ICO) prior to the actual token sale period. The goal of a pre-sale is to attract additional attention from investors, increase

[5] See www.bitcointalk.org.

total funding amount or to finance the subsequent main token sale (i.e., technical implementation as well as marketing expenses). Typically, a pre-sale is characterized by very high granted bonuses. After the pre-sale, the main token sale starts and lasts for a predefined period. During this period, the token price usually varies due to a predefined bonus scheme that rewards early investors. Once the token sale period is over and the ICO is successful, ICO initiators begin with the implementation process of their project plans as described in the initial white paper. Some successful ICOs strive for a listing of its distributed tokens on a cryptocurrency exchange, such as Binance[6] or Coinbase[7]. Once a token is listed on an exchange, token holders can start trading their tokens on the secondary market. Figure II-1 summarizes the typical ICO process.

Figure II-1: Schematic ICO process

This figure shows the typical stages of an ICO process.



[6] See www.binance.com.

[7] See www.coinbase.com.

2.3 Legal framework

With increased public attention, regulators worldwide have started to deal with ICOs and to provide regulatory frameworks for token sales. However, the current state of progress of implementation varies by country (Barsan, 2017; Dobrauz-Saldapenna and Klebeck, 2019; Hacker and Thomale, 2018). Especially, regulators have varying views on the legal characterization of cryptocurrencies and tokens, respectively. In consequence of diverging features of tokens, such as the distinction between utility, security and currency tokens, some regulators characterize tokens as commodities (Bolotaeva et al., 2019; Enyi and Le, 2017), while others consider them to be property (IRS, 2014). In the following, we give a brief overview about the legislation on ICOs in the five countries with the highest total amounts raised[8], i.e. United States, Russian Federation, Switzerland, Singapore and China.

In the United States, legal classification of an ICO is based on the classification of issued tokens. As the first step, thus, the American exchange supervisory authority (SEC) assesses whether an issued token has to be classified as a security. For this purpose, the SEC applies the Howey test, the standard test for the classification of financial products in the US (Murphy, 1946). According to this test, an issued token has to be classified as a security if the token constitutes a 'contract, transaction or scheme whereby a person invests his money in a common enterprise and is led to expect profits solely from the efforts of the promoter or a third party' (Murphy, 1946, no. 2). In case of a positive test result, tokens are required to be registered with the SEC and are subject to US security laws (Debler, 2018; Maume and Fromberger, 2019). On the other hand, there is no special regulation of the handling of utility tokens, as those are not classified as securities. In summary, US legislation regulates the legality of a token sale on a case-by-case basis (SEC, 2017).

Analogously to US legislation, for Swiss authorities, the classification of a token constitutes the first step of the assessment which existing laws are applicable. On February 16, 2018, the Swiss Financial Market Supervisory Authority (FINMA) published guidelines on the regulatory framework for ICOs (FINMA, 2018). According to these guidelines, the FINMA distinguishes between 'payment tokens', 'utility tokens'

[8] See www.icowatchlist.com/statistics/geo for data on ICO statistics by country.

and ‘asset tokens’. Only ‘asset token’ that ‘represent assets such as a debt or equity claim on the issuer’ (FINMA, 2018, p. 3) are treated as securities and therefore, are subject to security laws.

In Russia, the central bank of the Russian Federation is responsible for the regulation of ICOs. In January 2018, the Ministry of Finances published a first draft of a law regulating digital financial assets, called the Digital Assets Regulation Bill (MinFin, 2018). In accordance with this draft, tokens should be classified as property. Another feature of this draft is that qualified investors can unrestrictedly participate in ICOs while retail investors have only a restricted right to participate.[9] Although there had been several other drafts since then, as of the end of March 2020, there is still no special regulation for token sales in Russia, i.e. Russian authorities do not regulate ICOs at all (Partz, 2020).

Singaporean regulatory authority (MAS), on August 1, 2017, issued guidance on how they will regulate issued tokens that fall under the Securities and Futures Act (SFA) (MAS, 2017b). In this statement, the MAS announced that it would apply existing security laws if a token falls within the definition of a security. Also in 2017, the MAS warned investors against investing in ICOs due to fraudulent conduct by a high number of ICO initiators (MAS, 2017a). In summary, regulation of ICOs is based on a case-by-case assessment by Singaporean authorities as well.

In China, seven central government regulators issued an announcement on September 4, 2017 wherein they prohibited ICOs entirely to protect Chinese investors from fraudulent conduct by ICO initiators (Deng et al., 2018; PBC, 2017). Until then, Chinese regulators did not regulate ICOs at all.

Overall, the legal characterization and regulation of token sales vary markedly for the individual countries. While some countries, such as China and South Korea, take a very restrictive approach by entirely prohibiting ICOs, other countries, such as Russia, do not regulate ICOs at all. Consequently, regulation significantly influences the regional distribution of conducted ICOs.

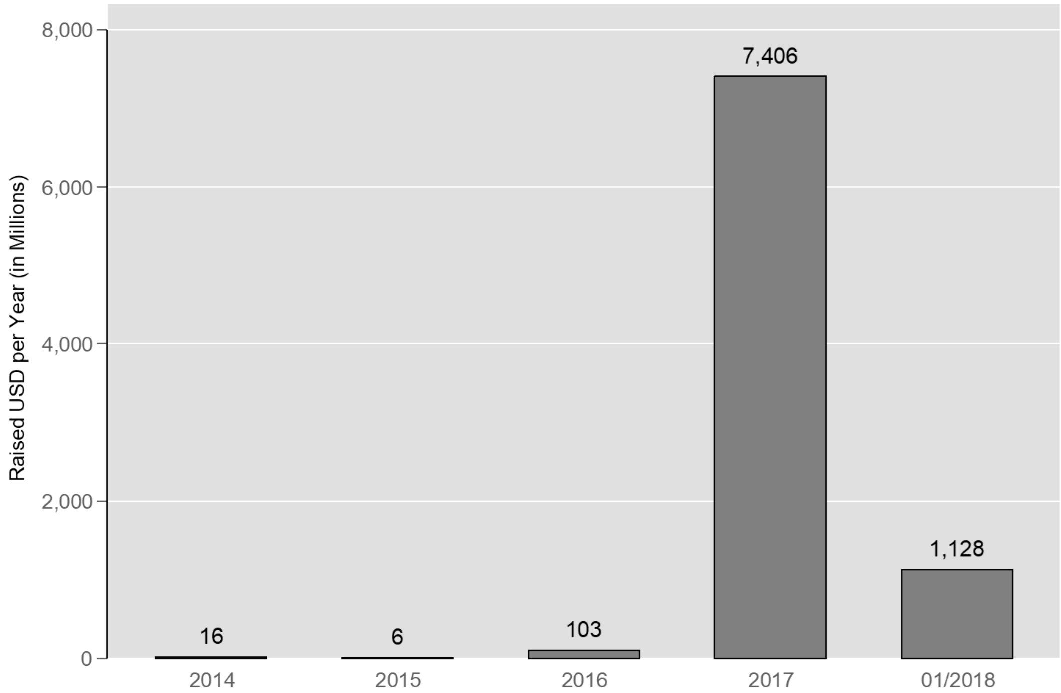
[9] If necessary, Russian authorities reserve the right to prohibit a token sale on a case-by-case basis.

2.4 Market overview

Figure II-2 provides an overview of the development of the ICO market for our sample from July 2014 to January 2018. Presented numbers are in line with other public sources, such as the research report by the accounting firm Ernst and Young (2017).

Figure II-2: Development of the ICO market from July 2014 to January 2018

This figure features all ICO proceeds for our sample comprising ICOs that were conducted between July 2014 and January 2018. The first bar summarizes all proceeds for our whole sample. Further bars illustrate all proceeds in the respective periods indicated below.



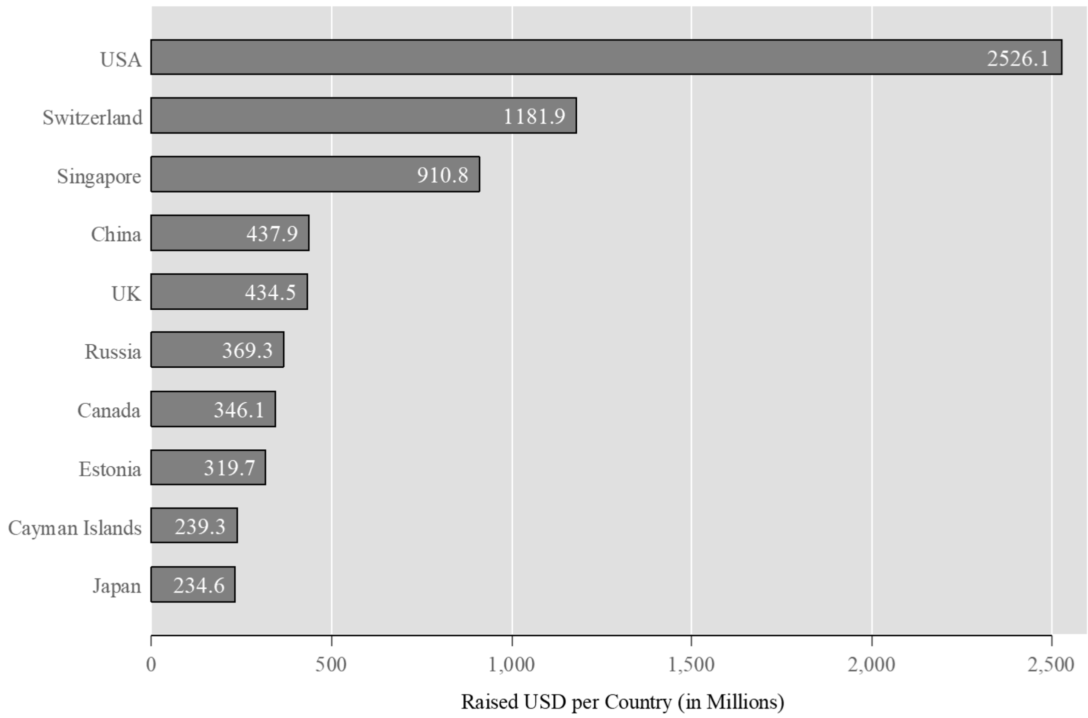
After a total amount of about 125 million USD from 2014 to 2016, ICOs collected more than seven billion USD in 2017 and more than one billion USD in the first month of 2018, and thus, constitute a fast-growing funding source on the global financial markets.

The global ICO market is characterized by a wide geographical dispersion. Figure II-3 presents the distribution of ICO projects' origin for our underlying data sample. As can be seen from Figure II-3, besides western countries, such as the USA, the UK and Switzerland, also Asian countries as well as Russia play an essential role in the ICO market. Moreover, ICOs are popular in offshore financial centers, such as the Cayman Islands. Beside for legal and regulatory reasons, this fact could be an indication that many ICO are conducted for reasons of tax avoidance, money laundering or other fraudulent intents (Tiwari et al., 2020). A study prepared by the ICO advisory firm Statis Group reports that about eleven percent of ICO investments fell prey to fraudulent projects ('scams') (Dowlat, 2018). Moreover, Huang et al. (2019) provide a more detailed overview about the geography of ICOs.

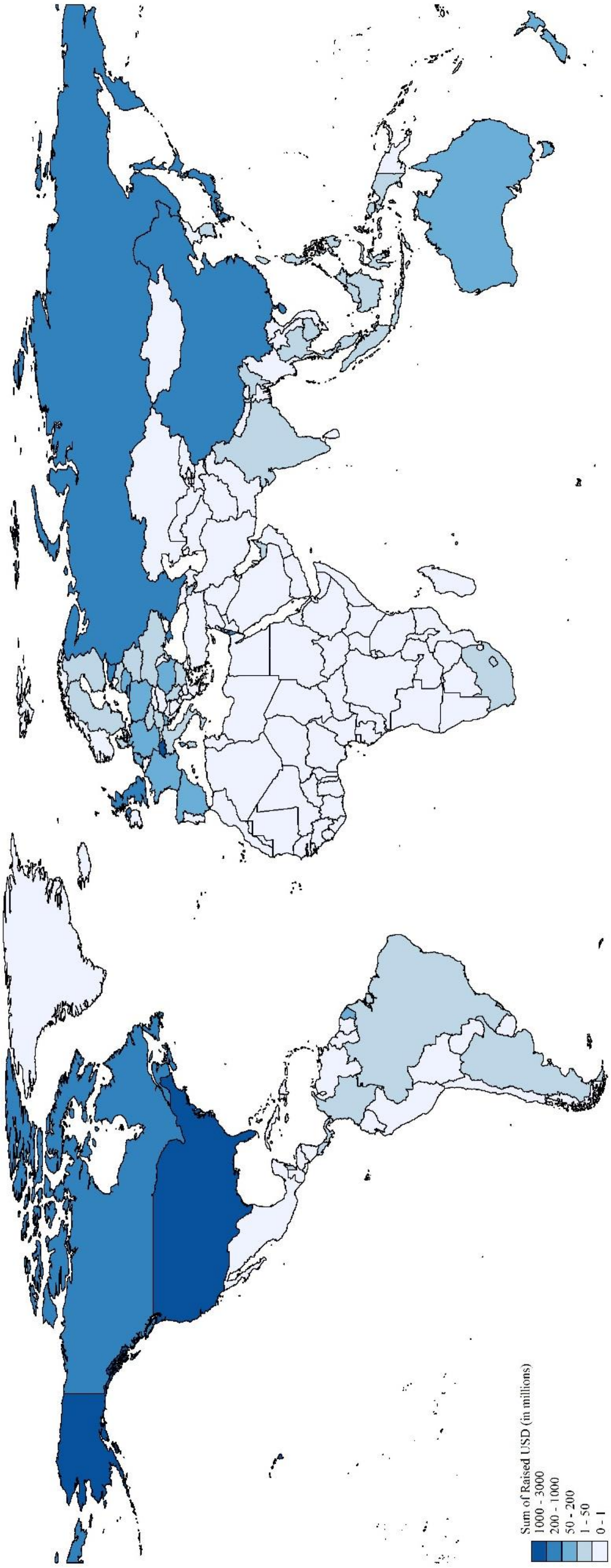
Figure II-3: Raised funds of ICO projects by country

Panel A shows the top ten leading countries worldwide by value of funds raised through ICOs in the period from July 2014 to January 2018. Panel B presents the geographical dispersion of ICO projects on the basis of total funds raised in the period from 2015 to January 2018.

Panel A: Top Ten Countries



Panel B: Geographical dispersion of ICO projects



2.5 Comparison of ICOs to conventional crowdfunding

As stated above, an ICO can be defined as a form of early-stage financing that uses the distributed ledger technology, which, depending on the token form, grants monetary or non-monetary rewards to the backers. The forms of conventional capital formation closest to ICOs are reward-based and equity-based crowdfunding. As defined by Belleflamme et al. (2014, p. 588), 'Crowdfunding involves an open call, mostly through the Internet, for the provision of financial resources'. In reward-based crowdfunding, capital seeking projects provide backers non-monetary rewards or products in exchange for funding (Ahlers et al., 2015). Accordingly, ICOs issuing utility tokens can be considered as a form of reward-based crowdfunding. On the other hand, in equity-based crowdfunding, funders receive an amount of equity or bond-like shares in the underlying project (Ahlers et al., 2015). With this definition, ICOs issuing security tokens can be regarded as a form of equity-based crowdfunding.

Nevertheless, ICOs and conventional crowdfunding campaigns differ in various aspects. In the following, we present the most important differences.

In general, in conventional crowdfunding the investment process is centralized on crowdfunding platforms that act as intermediaries (Belleflamme et al., 2014). Since crowdfunding platforms services are mostly directed to domestic investors and projects, the investor base in crowdfunding rather has a local character (Giudici et al., 2018). Moreover, reward-based crowdfunding is often characterized by social ties between investors and fundraisers (Giudici et al., 2018). In ICOs, however, using the DLT, investors allocate their financial resources directly to the project initiators. Consequently, project initiators and investors do not depend on any (local) intermediary platform. Therefore, given a particular project, we assume a wider geographical dispersion of ICO investors compared to crowdfunding investors.

With regard to the typology of fundraisers, ICOs and conventional crowdfunding are similar. More specifically, startups and young companies usually make use of conventional crowdfunding, i.e. equity-based and reward-based crowdfunding, to foster the growth of their venture (Paschen, 2017). Although some established companies conduct an ICO, however, the majority of ICO projects are at an early stage as well.

According to the differences in the typology of investors and fundraisers, there are great disparities in terms of number of campaigns and the average funding per campaign. As stated above, ICO initiators collected more than seven billion USD in 2017, which constitutes a value similar to the transaction value of equity-based and reward-based crowdfunding campaigns taken together (Statista, 2019a, 2019b). However, while less than one thousand ICO campaigns are responsible for this high transaction value in the emerging crypto market, the transaction value in the conventional crowdfunding market is generated by about 38 thousand equity-based crowdfunding campaigns and about 5.2 million reward-based crowdfunding campaigns (Statista, 2019c, 2019d). As a result, the average ICO campaign from our sample collected about 8.6 million USD, whereas, equity-based crowdfunding campaigns and reward-based crowdfunding campaigns collected 78,867 USD and 765 USD on average, respectively.

A consideration of the geographical distribution of ICO projects and conventional crowdfunding campaigns reveals differences as well. With regard to reward-based crowdfunding, about 80 percent of total funds were collected in China, while another 10 percent were collected in the US (Statista, 2019a, 2019b). These numbers suggest a high level of market concentration. In equity crowdfunding, China and the US constitute the most important markets as well. However, as campaigns in China are accountable for about 21 percent of the market volume and campaigns in the US for about 17 percent, the level of market concentration is essentially lower. With regard to the regional distribution of ICOs as presented in Figure II-3, the leading role of the US becomes apparent. The US account for about 29 percent of total funds in our sample. Furthermore, while the Swiss and the Singaporean market globally play a key role as well, the Chinese market is less important. However, in contrast to conventional crowdfunding that, according to Li (2016), is scarcely regulated in China, ICOs have been entirely banned in 2017 (PBC, 2017). In summary, besides the availability of financial resources, the geographical distribution of ICOs and conventional crowdfunding campaigns, respectively, is mainly driven by regulatory requirements in the different countries.

3. Theoretical background and hypotheses

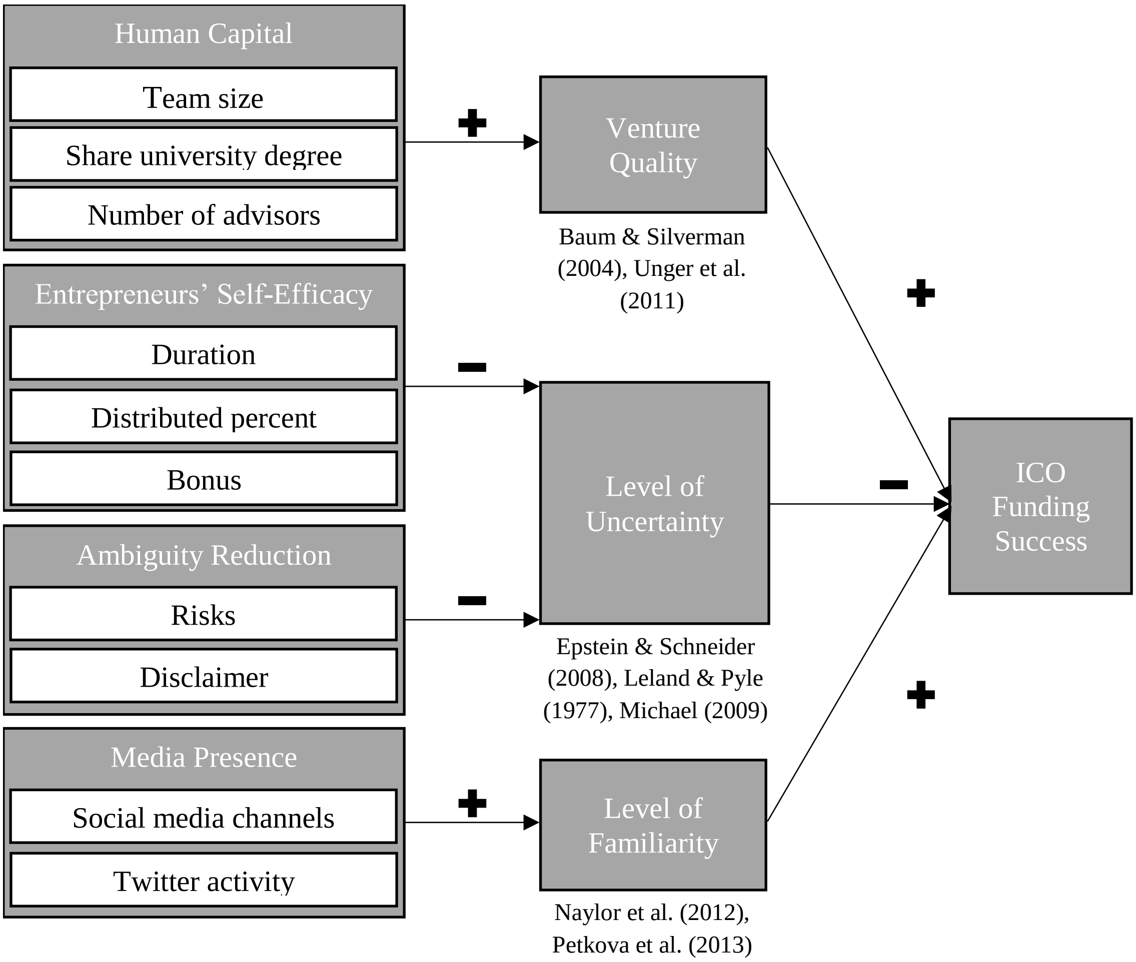
3.1 Determinants of ICO funding success

Like any other kind of investments in new ventures, ICO investments are subject to the well-documented principal agent problem (Fama and Jensen, 1983; Jensen and Meckling, 1976; Van Osnabrugge, 2000). On the one hand, investors (principals) try to select the best investment from the given options. On the other hand, entrepreneurs or project representatives (agents) aim to attract funds from investors. However, this allocation process is characterized by information asymmetries. Agents usually have more information about the true value of the project than the principals. Due to the low level of legal clarity combined with the anonymity of participants in the DLT, as described in the previous chapter, those information asymmetries are eminently high for ICOs compared to conventional start-up financing (Momtaz, 2020). Nevertheless, ICO investors, like any other types of investors, seek to reduce the likelihood to invest in 'lemons' (Akerlof, 1970). Hence, to attract funds from investors, project representatives have to decrease the information asymmetries perceived by potential investors. Therefore, according to signaling theory (Spence, 1973), project representatives need to provide information to investors to signal project quality. Though, not every type of information is an effective quality signal (Ahlers et al., 2015). In particular, effective signals are characterized as observable, i.e. investors recognize and understand them, and costly, i.e. the production of these signals entails costs (Connelly et al., 2011).

From a theoretical point of view and based on previous literature, we develop a framework on what types of information constitute effective signals that are used by entrepreneurs to convince potential investors and thus foster funding success. We argue that there are three channels that are related to funding success within the ICO context, namely (1) venture quality, (2) level of uncertainty and (3) level of familiarity. Figure II-4 shows the hypothesized model.

Figure II-4: Hypothesized model

This figure illustrates hypothesized determinants of funding success.



3.2 Human capital

Every new venture starts with a team of entrepreneurs that identifies a business opportunity and tries to exploit it (Shane and Venkataraman, 2000; Venkataraman, 1997). Therefore, it is evident that the human capital, i.e. all the knowledge, talents, skills, abilities, experience, intelligence, judgment, and wisdom of the project team (Haq, 1996), is an important factor of success for every entrepreneurial project. Unsurprisingly, a considerable amount of empirical literature has emphasized the importance of human capital for venture success (Bates, 1990; Baum and Silverman, 2004; Ray and Singh, 1980; Unger et al., 2011). Therefore, it is likely that potential investors are also aware of this relationship. Venture capitalists, for example, use firms' team characteristics as one of the most important criteria for their investment decisions (Zacharakis and Meyer, 2000).

Analogously, Ahlers et al. (2015) and Baum and Silverman (2004) demonstrated that human capital is an effective signal in conventional crowdfunding. Both, social capital, i.e. social networks and thus access to valuable information, as well as intellectual capital, i.e. employee expertise, are integral parts of human capital. Since ICO projects, similar to conventional crowdfunding projects, usually are at an early stage of the business life cycle, human capital is an important factor of project success.

Consequently, we argue that human capital is an effective signal of venture quality for potential investors and thus, positively relates to ICO funding success:

H1: Human capital positively relates to ICO funding success.

3.3 Entrepreneurs' self-efficacy

Besides the skills and knowledge of the team members, starting a new venture also requires the entrepreneurs' belief that the project will succeed. Dimov (2010) shows that opportunity confidence is one important factor of venture emergence. Opportunity confidence describes the personal belief of an entrepreneur that an opportunity at hand is feasible and that she is able to establish a venture that exploits this opportunity. Only if entrepreneurs believe that they can produce desired results by their actions, they have an incentive to start a venture. This trait is termed as 'self-efficacy' (Bandura, 2010). In this context, Baum and Locke (2004) find that entrepreneurs' self-efficacy is positively related to subsequent venture growth. The authors show that self-efficient entrepreneurs also have higher passion for the business. In the context of ICOs, entrepreneurs can show high self-efficacy by setting ICO parameters that are unambiguous, such as low bonus, short duration and low share of distributed tokens (i.e., higher share of tokens remains for the entrepreneurs). This observable and costly behavior show entrepreneurs' conviction of their own venture and might be an effective signal of venture quality that reduces the level of uncertainty from the investors' point of view. Since people prefer known risks over unknown risks (Ellsberg, 1961), we assume that entrepreneurs' self-efficacy is related with higher funding amounts. We hypothesize:

H2: Entrepreneurs' self-efficacy positively affects ICO funding success.

3.4 Ambiguity reduction

As aforementioned, known as ambiguity aversion (Ellsberg, 1961), individuals usually prefer known risks over unknown risks. In the case of investments, that implies that investors prefer an investment opportunity for which they know all underlying risks and probabilities to an equivalent investment opportunity with ambiguous information (Park and Patel, 2015). Ahlers et al. (2015) find that in the case of conventional crowdfunding, providing detailed information about risks can be an effective signal and therefore, fosters funding success. Given the low level of legal clarity in the environment of ICOs, we expect that potential investors are even more sensitive to the level of ambiguity linked to the ICO project. We argue that reducing ambiguity regarding the ICO project signals the team's awareness of potential risk factors as well as its preparedness for potential consequences. Additionally, potential investors have a better basis on which to form expectations, which is preferred by investors (Epstein and Schneider, 2008). Transferring the idea of ambiguity aversion to the ICO context, we argue that if the level of uncertainty decreases investors' likelihood to invest increases. Therefore, we hypothesize:

H3: Reducing ambiguity regarding the ICO project positively affects ICO funding success.

3.5 Level of media presence

In traditional entrepreneurial financing, such as venture capital or angel investment, personal communication is a key factor to establish social relationships between entrepreneurs and investors to decrease perceived information asymmetries and to signal project quality (Kollmann and Kuckertz, 2006; Shane and Cable, 2002). ICOs, however, like conventional crowdfunding, take place online. Consequently, most direct personal communication is mainly replaced by pseudo personal communication via social media (Drobetz et al., 2019; Moritz et al., 2015). Projects that show a higher (social) media presence are more likely to become familiar to potential investors (Heller Baird and Parasnis, 2011; Naylor et al., 2012). Moreover, active use of social media demonstrates preparedness and thus, signals venture quality (Courtney et al., 2017). Additionally, increasing social media activity can enhance the salience of an ICO and thus, possibly helps to inform investors about the upcoming investment opportunity (Solomon, 2012; Sprenger et al., 2014). Empirical literature also supports the hypothesis of media presence as an effective signal by finding that the use of media is positively related to crowdfunding success (Beier and Wagner, 2015; Courtney et al., 2017). Additionally, intense and diverse social media communication might increase the attention an ICO project receives from different types of traditional media, which is related with higher funding amounts (Petkova et al., 2013). Therefore, we hypothesize:

H4: The level of media presence positively affects funding success.

4. Data set and construction of variables

4.1 Data sources and sample construction

We obtain our data from three different sources: ICO information platforms, ICO white papers and ICO projects' social media channels. First, we collected data from seven different ICO information platforms[10] to define our sample and to derive our dependent variable, i.e. raised funds. Second, we use information from the ICO white papers to create the majority of our independent variables. Third, we investigate the presence of each ICO project on eight different social media platforms. Here, additionally, we investigate ICOs' Twitter accounts more deeply to assess the social media activity of ICOs. Please see Appendix II-A for detailed description of the data processing procedure.

Online ICO information platforms are public databases that contain information about upcoming, current and past ICOs. Therefore, those platforms are the starting point of our data collection. Typically, these platforms contain information on the name of the ICO, ICO's time schedule, details about the offering, but also links to the project's website, white paper or social media channels. After the token sale event, most platforms also list the funds raised by the ICO. However, an entry in those platforms is not mandatory. As a result, no platform contains complete information about all ICOs that have taken place. Therefore, and to get an initial sample as comprehensive as possible, we collect data from seven different ICO information platforms from July 2014 to January 2018. We manually match the data from the seven different ICO information platforms and remove duplicates. We highlight the importance of a manual merging procedure, as the names of the projects often slightly differentiate among the different platforms. Next, we remove ICO pre-sales from our sample, as we are only interested in ICO main sales. This procedure results in a sample of 1,057 different ICOs.

[10] We use the platforms icodata.io, icotracker.net, icobazaar.com, tokendata.io, icobench.com, smithandcrown.com and icodrops.com.

After defining our sample, we collected data to generate variables to proxy for ICO funding success as well as for the different signals within the ICO context. We use the data from the different ICO platforms to obtain values for our dependent variable. As mentioned above, ICO information platforms offer comprehensive data about the ICOs besides the collected funds. However, we are not able to obtain a time stamp for the data entries. To avoid potential reverse causality issues, we therefore collect our data on the explanatory variables from other sources than the ICO information platforms. ICOs' white papers are our first source of data for our explanatory variables. White papers are documents written by ICO initiators to promote and explain their products or services as well as to present the project team and the planned ICO schedule to potential investors. As those white papers offer a creation date, we can base our investigation on information that were available to investors before the actual ICO period. From those white papers, we obtain data regarding the projects' human capital, entrepreneurs' self-efficacy, ambiguity reduction as well as our control variables.

The second source of data for our explanatory variables are social media platforms. More precisely, we use the data from social media platforms to proxy for the level of media presence of each ICO project. Therefore, we first scan eight different social media platforms, namely Twitter, Facebook, Bitcointalk, Github, Reddit, Telegram, Medium and Slack, for accounts of each ICO project that have been set up before an ICO. Moreover, we assess the activity of each ICO project on Twitter before the main token sale event.

4.2 Measure of ICO funding success

In the context of entrepreneurship and early stage financing, success is not a clear defined concept. Thus, studies use diverse approaches to capture funding success (e.g., Ahlers et al., 2015; Ahlstrom and Bruton, 2006; Courtney et al., 2017). As we try to capture funding success from the ICO initiators' perspective, we use collected funds during the token sale event as our dependent variable. The more funds an ICO project collects, the more successful is the ICO.

Another possibility to assess early stage investment success would be to investigate if there was a successful exit (e.g., an initial public offering or a private placement) in the future. However, this definition rather describes success from investors' view, while we would like to capture success from the perspective of ICO initiators. Furthermore, we lack the information on whether an ICO campaign even considers an exit in the (near) future. Therefore, we focus our investigation on the ICO event itself instead of a potential exit event in the future.

A further option would be to define success in relation to the funding goal. Subsequently, some researchers measure success with a binary variable being one if the funding target was reached and zero otherwise (Courtney et al., 2017; Wang et al., 2018), or by a metric variable that captures the funds actually collected in proportion to the funding goal (Duffner et al., 2009). Most ICOs, however, do not define an explicit funding goal (Fisch, 2019). Frequently, ICOs only disclose a so-called 'soft cap' or 'hard cap'. The soft cap describes a threshold that, if it is not reached during the ICO, usually leads to a complete refund of all ICO investments. As ICO initiators try to avoid such an event, they often set the soft cap to an especially low level. The hard cap, on the other hand, defines the maximum amount of total investment approved by the algorithm of the ICO's smart contract. Hence, the level of the hard cap often does not relate to the funds needed for the accomplishment of the underlying project. Therefore, the soft cap and the hard cap commonly are not adequate benchmarks for ICO funding success.

4.3 Construction of explanatory variables

We collect data for six different categories of variables: (1) funding success, (2) human capital, (3) entrepreneurs' self-efficacy, (4) ambiguity reduction, (5) level of media presence and (6) controls. We use the following variables for our estimations:

As discussed in the previous section, we define ICO success as the funds collected by a project during the token sale event. Therefore, the dependent variable in our model is the amount raised by the project during ICO main sale in million USD (*raised mUSD).*

Our first category of explanatory variables captures human capital. Following the literature on conventional crowdfunding (Ahlers et al., 2015), we extract the size of the project team (*team size*) as well as the share of team members that hold a university degree (*share university degrees*) to proxy for human capital. Moreover, we argue that projects' advisors can offer the team valuable guidance as well as the access to a personal business network and though, improve human capital (social capital) as well. Therefore, we use the projects' number of advisors (*number advisors*) as a third proxy for human capital.

To proxy for entrepreneurs' self-efficacy, we obtain data about ICO duration (*duration*), the share of tokens distributed to the public during ICO (*distributed percent*) and potential bonuses (*bonus*) from projects' white papers. We argue that a short ICO duration (set prior to the token sale event), such as in the case of equity crowdfunding (Lukkarinen et al., 2016), signals the project team's confidence in their ability to collect the needed funds in a short period and therefore, can serve as a proxy for entrepreneurs' self-efficacy. Moreover, we argue that the higher the share of tokens that remains in the ownership of the project team the higher the team's confidence in project success (Ahlers et al., 2015). It follows that the lower the share of tokens distributed to the public, the higher the team's self-efficacy. This is in line with the literature documenting that entrepreneurs with a higher self-efficacy hold larger stakes in their venture (Cassar and Friedman, 2009). Lastly, we argue that the lower potential discounts or bonuses in an ICO that initiators offer to investors the higher the project team's confidence in the project's quality. Accordingly, setting those ICO parameters in the described manner reduces the level of uncertainty from the investors' point of view.

Our third category of explanatory variables comprises proxies for ambiguity reduction. Some white papers offer a disclaimer containing legal information about the ICO (investment). Moreover, a decent number of white papers offer a section about potential risk factors linked to a participation in the ICO. In this context, Arnold et al. (2010) as well as Park and Patel (2015) show that there is a relationship between the ambiguity of a project perceived by investors and the risk factors section in the underlying IPO prospectus. Therefore, we use a dummy variable that captures the existence of a section regarding potential risk factors (*risk factors*) in the ICO's white paper as a proxy for ambiguity reduction. Additionally, we include a second dummy variable equaling one if there is a legal disclaimer (*disclaimer*) in the corresponding white paper, and zero otherwise.

As explained above, we use data from eight different social media platforms to assess the level of media presence of a project. Therefore, we count the number of social media channels a project uses before ICO. The resulting variable (*social count*) is a proxy for the diversity of communication and our first measure of the level of media presence of a project.

Figure II-5: Usage of social media platforms by ICO projects in percent

This figure shows the percentage of ICO projects that use the respective social media platform. The first bar includes all ICO projects that use at least one social media platform.

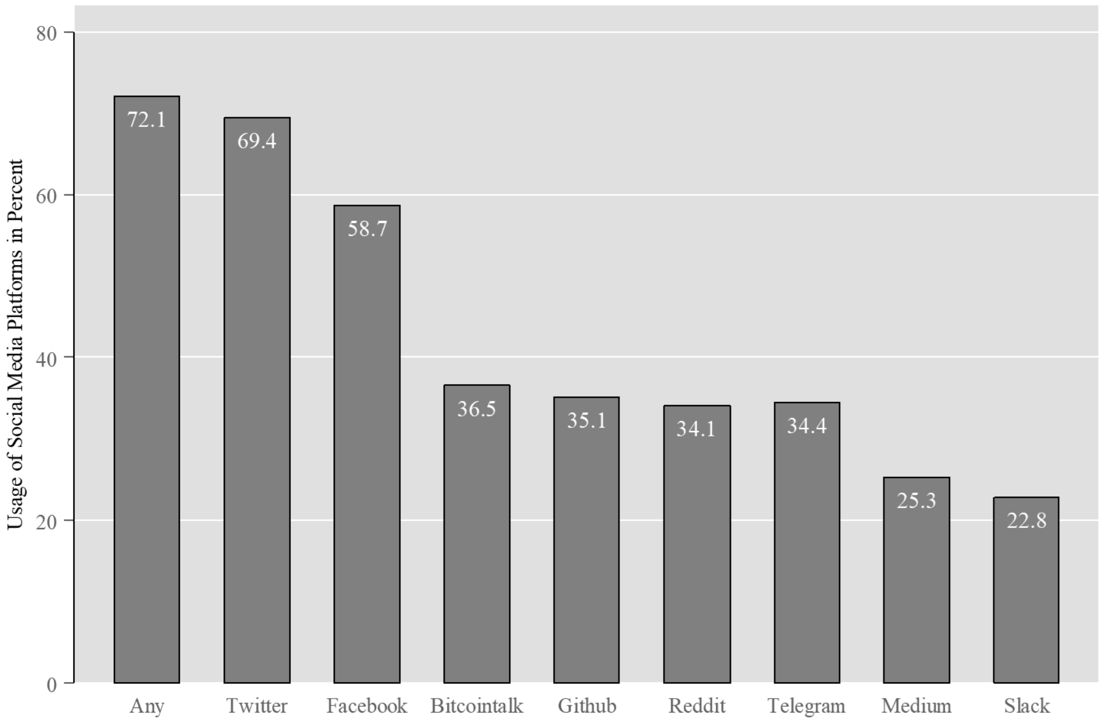


Figure II-5 gives an overview over the usage of the different social media platforms among the projects in our sample. As shown in Figure II-5, more than 72 percent of ICOs use at least one social media channel. Among the eight different channels, Twitter is the most prominent one. Considering all projects that use at least one social media channel, more than 96 percent of them use Twitter.

In addition to the number of social media platforms, the activity on those platforms is an important factor for the overall media presence of an ICO project. As Twitter is the most common social media channel among ICO projects, we identify the number of posts of each ICO project on Twitter in the last 60 days before the ICO as a second measure of media presence (*Twitter activity*).

Further controls constitute the last category of variables. Our control variables are the token price during ICO (*token price*), the projects' funding goal (*goal mUSD*) and a dummy variable for the existence of a pre-sale before ICO (*pre-sale*).

The token price is the price (i.e., amount of traditional fiat money or established cryptocurrencies) an investor has to pay for one token of the ICO in USD. Often the token price is stated in Ethereum or Bitcoin and therefore, the USD token price varies over time due to the significant fluctuations of these cryptocurrencies. In those cases, we identify the average USD token price during the ICO period.

A further control variable is the funding goal. However, as mentioned in section 3.1, project initiators often define no concrete funding goal. Mostly, only the so-called hard cap is given. Nevertheless, following the literature on conventional crowdfunding (Ahlers et al., 2015; Courtney et al., 2017), we control for the stated funding goal, soft cap, or hard cap (depending on availability) in USD (*goal mUSD*) but point out that the reliability of this control variable is relatively small.[11]

Lastly, we identify whether a project offers a pre-sale before the ICO main sale. We do so, because such projects could be more familiar to investors. The resulting variable (*pre-sale)* is a dummy variable that equals one, if there was a pre-sale before the ICO, and zero otherwise.

4.4 Summary statistics and correlations

Table II-1 provides the summary statistics for our sample. It is structured in six panels (A to F). Note that we have 1,057 observation for our variables *raised mUSD* and *social count* as we obtained the data for those variables from ICO information platforms and social media platforms, respectively. Other variables (apart from *Twitter activity*), however, are obtained from the projects' white papers. Hence, this data is only available for ICOs providing a white paper before the token sale event. As only about 82 percent

[11] Only 43 of the projects provide a specific funding goal. Moreover, 95 projects provide a soft cap. Other projects either provide a hard cap or no information about a funding goal at all. Therefore, we create the variable *goal mUSD* as follows: If the projects provides a specific funding goal the variable equals that goal. Moreover, if the project provides no funding goal, the variable equals the soft cap. If the project neither provides a funding goal nor a soft cap, the variable equals the hard cap. For reasons of robustness, we tested whether our results are affected by the construction of the variable. However, results do not change significantly. Additionally adding interaction terms for the respective goal types does not change our results. Regression results are available upon request.

of the projects in our sample provided a white paper, the maximum number of observations for those variables is 863. The variable *Twitter activity* holds only 734 observations, as it is only available for projects that had a Twitter account before ICO.

Table II-1: Summary statistics

This table contains the summary statistics of our dataset. We define the variables as follows: *raised mUSD* is the amount raised by the project during the ICO main sale in million USD. *team size* is the number of individuals in the project team. *share university degree* is the share of the team members that hold a university degree. *number advisors* is the number of advisors of the ICO project. *duration* is the duration of the ICO in days. *distributed percent* is the share of tokens that is distributed to the public during the ICO. *bonus* is the maximum bonus that is granted to investors during the ICO. *risks* is a dummy variable that equals one if there is a section in the ICO white paper that declares potential risk factors of the ICO (investment), and zero otherwise. *disclaimer* is a dummy variable that equals one if there is a (legal) disclaimer in the ICO white paper, and zero otherwise. *social count* is the number of social media platforms the ICO project uses. *Twitter activity* is the number is tweets the ICO project posted in 60 days before the start of the ICO. *token price* is the price of the token during the ICO in USD. *pre-sale* is a dummy variable that equals one if there was a pre-sale before the ICO main-sale. *goal mUSD* is the fund raising goal of the ICO project in USD.

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) p5	(6) p25	(7) p50	(8) p75	(9) p95	(10) Max
Panel A: ICO success										
raised mUSD	1,057	8.64	20.59	0	0	0	0.89	9.75	37.86	258.00
Panel B: Human capital										
team size	863	4.47	5.91	0	0	0	3	7	14	80
share university degree	863	0.14	0.27	0	0	0	0	0.17	0.88	1.00
number advisors	863	1.81	3.56	0	0	0	0	3	10	35
Panel C: Entrepreneurs' self-efficacy										
duration	830	34.48	26.48	1	1	17	31	45	91	195
distributed percent	809	0.61	0.25	0.00	0.15	0.45	0.62	0.80	1.00	1.00
bonus	863	0.23	0.56	0.00	0.00	0.00	0.11	0.30	0.79	10.00
Panel D: Ambiguity reduction										
risks	863	0.22	0.42	0	0	0	0	0	1	1
disclaimer	863	0.34	0.47	0	0	0	0	1	1	1
Panel E: Media presence										
social count	1,057	3.16	2.48	0	0	0	3	5	7	8
Twitter activity	734	64.30	106.43	0	0	0	29	81	258	1,087
Panel F: Controls										
token price	776	15.66	288.16	0.00	0.01	0.10	0.30	1.00	10.00	7,912.60
pre-sale	863	0.52	0.50	0	0	0	1	1	1	1
goal mUSD	774	29.79	46.63	0.03	0.50	4.50	15.01	35.00	100.00	500.00

As can be inferred from Panel A, ICO projects raised 8.64 million USD on average. The median, however, is only 893 thousand USD indicating a positively skewed distribution of *raised mUSD*.[12] More than 25 percent of the ICOs collected no funds at all. It follows that despite the high popularity of ICOs investors did not blindly delegate money to every project that was somehow related to distributed ledger technology. The maximum raised

[12] Plotting the residuals of the regressions from our main specification, we find no deviation from the assumption of normal distributed standard errors except for heteroscedasticity. Therefore, we estimate heteroscedasticity-consistent standard errors. Moreover, using log-transformed *raised mUSD* does not significantly change the results of our investigations. Regression results are available upon request.

funds by one project in our sample are 258 million USD by the Hdac project. While some news articles report ICOs that raised much higher sums (Kharif, 2018), the relatively low value in our sample results from the fact that we restricted our sample to ICOs that were completed until January 2018.[13]

Panel B provides data about the human capital of ICO projects. The average stated team size is four, while 14 percent of the team members declare to own a university degree. Moreover, ICO projects in our sample present two advisors on average. More than 25 percent of ICOs do not present any founders or team members in their white papers.

Panel C shows that the mean of ICO duration is about 34 days. However, there are also ICOs that take place on only one day or that take up to 195 days. The average ICO distributes 61 percent of generated tokens to the public. Therefore, founders on average retain a 39 percent of tokens. The bonus fluctuates between 0 and 1,000 percent, while being 25 percent on average. A bonus of 25 percent implies that when you buy one token and you fulfill specific criteria, you receive 1.25 tokens instead. Note that we always capture the highest possible bonus during the main sale.

As can be inferred from Panel D, only 22 percent of ICO white papers present potential risk factors, while about 34 percent provide a legal disclaimer.

Panel E provides information about the level of media presence of ICO projects before the token sale. Projects run three social media channels on average. However, while the maximum of *social count* is eight, more than 25 percent of the projects in our sample use no social media channel at all. The mean of *Twitter activity* is 64.30, implying that the average ICO twitter account posts about 64 tweets in the 60 days before ICO. However, the median of the variable is only 29, showing that the mean is driven by a few projects that write many tweets (up to 1,087) before ICO.

Controls (Panel F) show that the token price is 15.66 USD on average, while the median price of one token is 0.30 USD. As many projects state the token price in Bitcoin or Ethereum, the corresponding USD price is subject to significant fluctuations. For instance, the minimum Bitcoin price in our sample period was 572 USD, while the maximum price

[13] The EOS ICO, for example, collected more than four billion USD, however, over several sale events from June 2017 until June 2018.

was 19,479 USD. More than 50 percent of the projects offer a pre-sale before the main sale event. The mean of *goal mUSD* is more than 29 million USD, while the median is 15.01 million USD.

Table II-2 presents the correlation matrix of our variables. Our main variable of interest and proxy for ICO success (*raised mUSD*) is positively correlated to *team size* and the number of advisors, indicating a positive relationship between ICO success and human capital. However, there is no correlation between *raised mUSD* and the share of team members holding a university degree. Moreover, *raised mUSD* is statistically significantly related to all three proxies for the entrepreneurs' self-efficacy (*duration*, *distributed percent*, *bonus*) as well. As higher values of those variables indicate a lower self-efficacy, negative signs of the correlations indicate a positive relationship between entrepreneurs' self-efficacy and ICO success. There seems to be no linear relationship between our proxies for the ambiguity reduction, namely *disclaimer* and *risks* and ICO success. However, the level of (social) media presence (*social count* and *Twitter activity*) positively related to ICO success. Overall, the correlations suggest that human capital, entrepreneurs' self-efficacy and media presence are effective signals within the ICO context.

Apart from linear relationships between explanatory variables and ICO success, there are also relationships between several of our explanatory variables. Especially, *pre-sale* and *social count* are statistically significantly related to most of our other explanatory variables. To assess potential collinearity issues in our main regression models, we calculate the variance inflation factors (VIF) for all our model specifications. We find a maximum VIF of 1.61 indicating no severe collinearity issues in our regressions.

Table II-2: Correlation matrix

This table shows the Pearson correlation coefficients for *raised mUSD* and all explanatory variables used in our main regressions. p-values are given in parentheses below the coefficients. ***, **, * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
raised mUSD (1)	1													
team size (2)	0.125***	1												
	(0.000)													
share university degree (3)	0.029	0.302***	1											
	(0.335)	(0.000)												
number advisors (4)	0.109***	0.526***	0.299***	1										
	(0.000)	(0.000)	(0.000)											
duration (5)	-0.179***	-0.049	0.015	-0.021	1									
	(0.000)	(0.108)	(0.613)	(0.486)										
distributed percent (6)	-0.174***	-0.041	-0.056**	-0.088***	0.054*	1								
	(0.000)	(0.151)	(0.046)	(0.002)	(0.099)									
bonus (7)	-0.076**	-0.078***	-0.056**	-0.059**	0.055	0.003	1							
	(0.028)	(0.006)	(0.049)	(0.036)	(0.101)	(0.927)								
disclaimer (8)	0.047	0.150***	0.091***	0.190***	0.008	-0.042	-0.001	1						
	(0.166)	(0.000)	(0.007)	(0.000)	(0.824)	(0.226)	(0.982)							
risks (9)	-0.006	0.161***	0.115***	0.165***	-0.002	0.027	-0.023	0.388***	1					
	(0.858)	(0.000)	(0.001)	(0.000)	(0.948)	(0.437)	(0.522)	(0.000)						
social count (10)	0.253***	0.188***	0.068***	0.197***	-0.135***	-0.049*	-0.010	0.087**	0.059*	1				
	(0.000)	(0.000)	(0.009)	(0.000)	(0.000)	(0.087)	(0.724)	(0.011)	(0.083)					
Twitter activity (11)	0.111***	0.213***	0.114***	0.203***	-0.051	-0.040	-0.060*	0.009	0.079**	0.105***	1			
	(0.003)	(0.000)	(0.000)	(0.000)	(0.162)	(0.216)	(0.067)	(0.823)	(0.046)	(0.001)				
token price (12)	-0.008	-0.035	-0.023	-0.025	-0.033	0.041	0.011	-0.034	-0.020	0.000	0.009	1		
	(0.804)	(0.284)	(0.473)	(0.450)	(0.322)	(0.222)	(0.758)	(0.348)	(0.572)	(0.994)	(0.803)			
pre-sale (13)	0.057*	0.376***	0.277***	0.273***	-0.003	-0.057**	-0.078***	0.171***	0.056*	0.157***	0.145***	-0.035	1	
	(0.061)	(0.000)	(0.000)	(0.000)	(0.922)	(0.044)	(0.006)	(0.000)	(0.095)	(0.000)	(0.000)	(0.281)		
goal mUSD (14)	0.174***	0.009	-0.049	0.018	0.046	0.000	0.048	-0.007	-0.018	-0.012	0.008	-0.013	0.014	1
	(0.000)	(0.774)	(0.125)	(0.572)	(0.172)	(0.996)	(0.167)	(0.853)	(0.620)	(0.708)	(0.831)	(0.700)	(0.659)	

5. Results

5.1 Mean differences tests

The correlation matrix presented above provides a first glimpse on potential relationships between our explanatory variables and ICO success. To get further insights into the relationship between quality signals and ICO success, we apply a mean difference test. Therefore, we perform a median split according to *raised mUSD* and then, compare the means of our explanatory variables for the resulting subsamples. Table II-3 presents the results.

Table II-3: Median split of *raised mUSD*

This table presents a comparison of the means of our explanatory variables for the two subsamples resulting from a median split according to *raised mUSD*. We define the variables as follows: *raised mUSD* is the amount raised by the project during the ICO main sale in million USD. *team size* is the number of individuals in the project team. *share university degree* is the share of the team members that hold a university degree. *number advisors* is the number of advisors of the ICO project. *duration* is the duration of the ICO in days. *distributed percent* is the share of tokens that is distributed to the public during the ICO. *bonus* is the maximum bonus that is granted to investors during the ICO. *risks* is a dummy variable that equals one if there is a section in the ICO white paper that declares potential risk factors of the ICO (investment), and zero otherwise. *disclaimer* is a dummy variable that equals one if there is a (legal) disclaimer in the ICO white paper, and zero otherwise. *social count* is the number of social media platforms the ICO project uses. *Twitter activity* is the number is tweets the ICO project posted in 60 days before the start of the ICO. *token price* is the price of the token during the ICO in USD. *pre-sale* is a dummy variable that equals one if there was a pre-sale before the ICO main-sale. *goal mUSD* is the fund raising goal of the ICO project in USD. ***, **, * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

		raised mUSD			
VARIABLES	Number of Observations	Below Median Group (B)	Above Median Group (A)	Difference (A) – (B)	t-statistic
Human capital					
team size	664	3.720	6.295	2.575	5.366***
share university degree	664	0.145	0.165	0.0206	0.956
number advisors	664	1.241	2.810	1.569	5.564***
Self-efficacy					
duration	664	41.430	26.270	-15.160	-7.605***
distributed percent	664	0.657	0.585	-0.072	-3.977***
bonus	664	0.343	0.178	-0.165	-3.454***
Ambiguity reduction					
disclaimer	664	0.352	0.392	0.039	1.043
risks	664	0.229	0.268	0.039	1.167
Media presence					
social count	664	2.732	4.669	1.937	11.38***
Twitter activity	521	35.000	63.870	28.870	4.201***
Additional controls					
token price	664	9.640	26.290	16.650	0.688
pre-sale	664	0.584	0.557	-0.027	-0.705
goal mUSD	664	26.620	30.560	3.935	1.253

Regarding human capital, we find a significant difference between the below median group of *raised mUSD* and the above median group of *raised mUSD* for two out of three variables. While the average team size of projects in the below median group is 3.72, the average of team size of projects in the above median group is nearly twice as high at 6.30. The number of advisors is even more than twice as high for projects in the above median group (2.80) compared to projects in the below median group (1.24). The number of people involved in a project positively relates to ICO success, while the education of the team members does not. Overall, evidence from the mean difference tests suggest that human capital is an effective signal for venture quality.

In line with correlation results, the mean difference tests for our proxies for entrepreneurs' self-efficacy are statistically significant. All three tests indicate that a higher self-efficacy leads to a higher ICO funding amount. The mean difference is most striking for ICO duration. While ICOs in the below mean group on average show a duration of more than 41 days, ICOs in the above median group only show an average of about 26 days.

Again, we find no evidence for a relationship between our proxies for ambiguity reduction and *raised mUSD*. Hence, investors seem not to perceive the existence of a legal disclaimer or a passage about potential risk factors as a signal for project quality.

For the proxy variables for the level of media presence, on the other hand, we find statistically significant mean differences. On average, projects in the above median group use 1.94 social media channels more than projects in the below median group. Moreover, projects initiators in the above median group write on average 28.87 more Twitter messages in the 60 days before the token sale event than the project initiators in the below median group. Note that we lose observations for ICOs that had no twitter account before the token sale event. In sum, results indicate that media presence effectively affects funding success. Lastly, tests for our controls show no significant mean differences.

5.2 Multiple regression results

In this section, we analyze the explanatory factors of ICO success within a linear regression framework. Results are presented in Table II-4.

Table II-4: Regression of *raised mUSD* on quality signals

This table presents results from ordinary least squares linear regressions (using robust standard errors) with the absolute funding amount (*raised mUSD*) as the dependent variable. We define the variables as follows: *raised mUSD* is the amount raised by the project during the ICO main sale in million USD. *team size* is the number of individuals in the project team. *share university degree* is the share of the team members that hold a university degree. *number advisors* is the number of advisors of the ICO project. *duration* is the duration of the ICO in days. *distributed percent* is the share of tokens that is distributed to the public during the ICO. *bonus* is the maximum bonus that is granted to investors during the ICO. *risks* is a dummy variable that equals one if there is a section in the ICO white paper that declares potential risk factors of the ICO (investment), and zero otherwise. *disclaimer* is a dummy variable that equals one if there is a (legal) disclaimer in the ICO white paper, and zero otherwise. *social count* is the number of social media platforms the ICO project uses. *Twitter activity* is the number is tweets the ICO project posted in 60 days before the start of the ICO. *token price* is the price of the token during the ICO in USD. *pre-sale* is a dummy variable that equals one if there was a pre-sale before the ICO main-sale. *goal mUSD* is the fund raising goal of the ICO project in USD. We use fixed effects for time (month-year), industry, token form (utility token, security token or currency token) and the ICO project's country of origin. t statistics in parentheses: ***, **, * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

	Model 1: all observations			Model 2: subsample of ICO projects running a Twitter account before ICO		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: *raised mUSD*	Coefficient	Beta	t-statistic	Coefficient	Beta	t-statistic
Human capital						
team size	0.029	0.008	0.182	0.042	0.012	0.235
share university degree	1.025	0.013	0.441	1.080	0.013	0.348
number advisors	-0.073	-0.012	-0.325	-0.129	-0.022	-0.515
Self-efficacy						
duration	-0.134***	-0.165***	-4.557	-0.122***	-0.146***	-3.706
distributed percent	-7.770*	-0.085*	-1.696	-10.725*	-0.103*	-1.683
bonus	-2.843**	-0.081**	-2.252	-3.148*	-0.082*	-1.870
Ambiguity reduction						
Risks	-1.359	-0.027	-0.698	-2.072	-0.038	-0.793
disclaimer	2.329	0.052	1.223	3.310	0.068	1.385
Media presence						
social count	2.099***	0.232***	4.671	2.396***	0.176***	3.659
Twitter activity				0.017*	0.080*	1.916
Additional controls						
token price	-0.001	-0.016	-0.962	-0.001	-0.010	-0.432
pre-sale	-2.938*	-0.067*	-1.648	-4.367*	-0.090*	-1.713
goal mUSD	0.164***	0.306***	2.779	0.196***	0.343***	2.651
Constant	5.938		0.552	20.802*		1.949
Observations	664			521		
R-squared	0.314			0.340		
Month-Year FE	YES			YES		
Industry FE	YES			YES		
Token Form FE	YES			YES		
Country FE	YES			YES		

There are 664 observations in our first estimation. We are able to explain 31.4 percent of the variation of *raised mUSD*. Column 1 shows regression coefficients of our first model. With regard to human capital, the significant correlations between *raised mUSD* and *team size* and *number advisors*, respectively, vanish when we control for other factors that are related to *raised mUSD*. Hence, there is no significant linear relationship between *raised mUSD* and any of our three proxy variables for human capital in our multiple regressions. Therefore, we reject our hypothesis H1 that human capital positively affects ICO funding success. Our results suggest that human capital is no significant signal for project quality from an ICO investors' point of view. In this regard, ICOs seem to differ from conventional crowdfunding investment (Ahlers et al., 2015; Barbi and Mattioli, 2019; Piva and Rossi-Lamastra, 2018). We argue that human capital plays a less important role from an ICO investor's perspective as ICO projects on average are much larger than conventional crowdfunding campaigns (see above). Consequently, stronger inter-personal ties between backers and investors than in ICO campaigns characterize conventional crowdfunding campaigns.

The coefficients for *duration*, *distributed percent* and *bonus* confirm our results from the correlation analysis (see Table II-2) and the median split (see Table II-3). The coefficients are statistically significant while showing a negative sign each. For every day less that an ICO last, it collects 134 thousand USD more. This finding is in line with the literature showing that IPOs with a shorter duration are perceived as less risky (Brooks et al., 2009). Moreover, for one percentage point of tokens less distributed to the public (and thus one percentage point of tokens more reserved by the ICO founders), an ICO collects 78 thousand USD more. We argue that, a low share of tokens distributed to the public indeed signals entrepreneurs' confidence in the value of their project, lower the level of uncertainty and, following the entrepreneurial ownership retention hypothesis (Leland and Pyle, 1977), is positively related to project value and ICO success. This is in line with the literature finding that ownership retention is positively related to firm value after an IPO (Downes and Heinkel, 1982). Lastly, for a one-percentage point lower bonus, an ICO collects 28 thousand USD more. This conforms to literature stating that customers may perceive high discounts as a signal for insufficient project quality (Gwinner et al., 1998). Overall, evidence supports our hypothesis H2. Entrepreneurs' self-efficacy is an important signal for project quality from an ICO investors' point of view.

With regard to the ambiguity reduction, we detect no significant relationship between used proxies and *raised mUSD*. Coefficients for both, *risks* and *disclaimer*, do not significantly differentiate from zero. This, again, is in line with our prior results (see Table II-2 and Table II-3) and literature on crowdfunding (Ahlers et al., 2015). Thus, evidently, ICO investors do not care about the declaration of potential risks or legal information about the investment. In conclusion, we argue that, contrary to hypothesis H3, ambiguity reduction is not a signal of quality for ICO investors.

With regard to media presence, however, evidence supports our hypothesis H4. We find that media presence constitutes an important factor of ICO success. For each social media platform a project uses, it is able to collect 2.10 million USD more. Results are in line with our prior investigations (see Table II-2 and Table II-3). Moreover, investigations on conventional crowdfunding find similar relationships (Barbi and Mattioli, 2019; Lukkarinen et al., 2016). To get deeper insights into the role of social media, we later include the Twitter activity of the projects in our second regression model.

Controls reveal that the *token price* is not related to the amount of USD raised. As the token price is arbitrarily divisible, this result is not a surprise. Projects with a pre-sale, on average collect about three million USD less. A possible interpretation for this finding might be that those projects attract institutional investors during the pre-sale who then do not invest during the main sale event anymore. The literature on conventional crowdfunding, however, finds a positive relationship between the availability of a pre-sale and funding success (Barbi and Mattioli, 2019; Lukkarinen et al., 2016). We argue that ICOs differ from conventional crowdfunding in this regard as the pre-sale at conventional crowdfunding if often exclusively directed to institutional investors. The involvement of institutional investors may convey credibility in the crowdfunding project (Lukkarinen et al., 2016). In ICOs, however, the pre-sale is mostly open to the public as well. Consequently, the ICO pre-sale is a substitute to the actual main sale, possibly reducing the funding amount in the ICO main sale. Lastly, projects with a higher goal or cap attract more funds. For a one-dollar higher goal, an ICO project is able to collect additional 16.4 cents. However, this result has to be interpreted with caution because of the fact that project often state no real goal but only specify a soft cap or hard cap.

As mentioned before, we deploy a second regression model to investigate the role of the ICO projects' Twitter activity. Results can be obtained from columns 4 to 6 of Table II-4. We only include ICOs in our estimation that had a Twitter account before the token sale event. Consequently, the number of observations drops to 521. However, the R-squared increases to 34 percent. Moreover, the constant increases significantly and is now statistically significant at the ten percent level. By implication, this increase means that this sub sample contains ICOs that are more successful. This is not surprising as the prior regression shows that ICOs that use more social media channels attract more investments.

With regard to *Twitter activity*, the regression coefficient is positive and statistically significant. For each Twitter message in the 60 days before the ICO, an ICO project is able to collect 17 thousand USD more. Evidently, media presence turns out to be an effective signal that entrepreneurs may use to induce investors to invest in an ICO. This, again, supports our hypothesis H4.

The other variables do not change notably among the two regression models indicating high robustness of our results.

Overall, entrepreneurs' self-efficacy as well as level of media presence constitute effective signals from the viewpoint of (potential) ICO investors. Evidence suggests, however, that human capital and ambiguity reduction are less important for investors.

5.3 Assessing the potential bias from unobservable omitted variables

Our main regressions are able to explain up to 34 percent of the variation of *raised mUSD*. Therefore, like fore many other empirical studies, a large part of the variation of our dependent variable remains unexplained, possibly resulting in an omitted variable bias. While we are not able to fully rule out the existence of this bias, we can assess the importance of selection of unobservable variables. Following the method of Oster (2019), we calculate δ for each of our proxies for entrepreneurs' self-efficacy and level of media presence. δ specifies how large the share of variation of *raised mUSD* that unobservable variables are able to explain relative to the share of variation explained by the control variables included in our regression model needs to be, to diminish the estimated effect of our explanatory variables of interest on ICO success. Therefore, we calculate

$$\delta = \frac{\beta_{full}}{\beta_{restricted} - \beta_{full}} \cdot \frac{R_{full} - R_{restricted}}{R_{max} - R_{full}} \tag{1}$$

, where β_{full} is the coefficient of our explanatory variable of interest using the full set of controls from our regression model in the previous section, while $\beta_{restricted}$ is the coefficient of our explanatory variable of interest from the model using the explanatory variable of interest only. R_{full} and $R_{restricted}$ are the R-squareds of the particular regression model, while R_{max} is the R-squared of a hypothetical estimation that includes both, observable and unobservable variables. We follow Oster's (2019) recommendation by setting $R_{max}$to $1.3 * R_{full}$. Table II-5 presents the results.

Results imply that the explanatory power of a potential omitted variable has to be 2.3 (*duration*) to 14.1 (*bonus*) times higher than the actual explanatory power of our full regression model to vanish the effect of our explanatory variable of interest. Oster (2019) suggests that a δ of more than one is an indication that there is no significant omitted variable bias in the given regression model. Therefore, we argue that omitted variables are no serious issue in our investigation. Moreover, the beta range in Table II-5 provides a range for the coefficients of our explanatory variables of interest when adjusting our estimations for a potential unobservable omitted variable effect. As none of the beta ranges encloses zero, results suggest that estimated coefficients are still different from zero.

Table II-5: Selection on observables to assess bias from unobservables

This table presents the results of Oster's (2017) test for unobservable selection and coefficient stability. δ is the degree of selection on unobservable variables relative to observed variables that would be necessary to explain away the results given the full model specifications. It is calculated as $\frac{\beta_{full}}{\beta_{restricted}-\beta_{full}} \cdot \frac{R_{full}-R_{restricted}}{R_{max}-R_{full}}$, where β_{full} is the coefficient of our explanatory variable of interest using the full set of controls as presented in column 4 of Table II-4, while $\beta_{restricted}$ is the coefficient of our explanatory variable of interest from the model using the variable of interest as explanatory variable only. R_{full} and $R_{restricted}$ are the R-squares of the particular regression models, while R_{max} is the R-square of a hypothetical estimation that includes both, observable and unobservable variables. We follow Oster's (2017) recommendation by setting $R_{max}$to 1.3*R_{full}. The beta range is $[\beta^*, \beta_{full}]$, where β^* is the bias-adjusted treatment effect that is calculated as $\beta^* = \beta_{full} - (\beta_{restricted} - \beta_{full}) * \frac{R_{max}-R_{full}}{R_{full}-R_{restricted}}$.

Explanatory variable of interest	Full model	δ	Beta range
duration	Controls + Month-Year FE + Industry FE + Token Form FE + Country FE	2.260	[-0.100, -0.122]
distributed percent	Controls + Month-Year FE + Industry FE + Token Form FE + Country FE	2.562	[-8.759, -10.720]
bonus	Controls + Month-Year FE + Industry FE + Token Form FE + Country FE	14.110	[-3.177, -3.148]
social count	Controls + Month-Year FE + Industry FE + Token Form FE + Country FE	4.374	[2.329, 2.396]
Twitter activity	Controls + Month-Year FE + Industry FE + Token Form FE + Country FE	2.786	[0.013, 0.017]

6. Limitations

Although our empirical results are based on a comprehensive database of mainly manually collected information and seem to be robust to omitted variable bias as suggested by the findings of the Oster (2019) test, our study may suffer from some limitations and remaining questions, respectively. In the following, we describe the most relevant limitations of our study.

First, the principles of ICOs and thus, the anonymity of ICO investors, prevent us from gaining deeper insights on the individuals investing in ICO projects. Therefore, we are not able to analyze investor characteristics, such as demographic factors or professional expertise, and their impact on investment behavior. Moreover, we are unable to identify investors' underlying motives with certainty. While we assume that investors are motivated by monetary rewards in the first place, non-monetary rewards, such as the access to future products and services as granted by utility tokens, might be more important for some investors. We try to take this into account by including fixed effects for the type of issued tokens in our main specifications.

Second, many of our variables only serve as proxies for broader concepts, e.g., we measure human capital signals by the size of project team and the share of team members that hold a university degree. Due to limited information that are provided by ICO initiators, however, we assume that potential investors do not have more relevant information as well. Therefore, our used proxies might be a close approximation of the respective quality signals. Moreover, our proxies are used in other relevant studies (e.g Ahlers et al., 2015).

Third, our data might be subject to a selection bias. Since we derive our initial ICO sample from online ICO information platforms, our statistical population is restricted to ICOs that are included on those platforms. We try to minimize the potential selection bias by incorporating data on ICOs from seven different ICO information platforms. Nevertheless, some ICOs might take place without being represented on any of those information platforms.

Fourth, it seems promising to obtain an in-depth understanding of signaling in ICOs by investigating various subgroups of ICOs. The effects we find in our major specifications might differ for different ICOs' countries of origin, token form or company age.

Unfortunately, our final sample consists of only 662 ICOs and because we include 14 explanatory variables as well as fixed effects for month-year, industry, token form and country, our sample is too restricted to perform further subgroup analyses.

Fifth, as presented in the previous chapter, results from the test proposed by Oster (2019) reveal that it is very unlikely that our regression models suffer from a omitted variables bias. However, we cannot certainly rule out that other ICO characteristics, which we do not consider or are unobservable, might be effective signals in the ICO context as well.

7. Conclusion

This paper examines which project signals that are provided by ICO initiators encourage investors to invest financial resources in an ICO context. Using data directly obtained from ICO white papers that are timestamped and therefore guaranteed to have been released before the ICO event, we highlight the importance of media presence and entrepreneurs' self-efficacy for ICO success. However, surprisingly, human capital as well as factors that reduce ambiguity (providing a disclaimer and information on potential risks) do not seem to determine ICO success. Thus, especially familiarity with an ICO project among potential investors and entrepreneurs' self-efficacy are drivers of funding success in the ICO context.

The implications of our main findings are manifold. From an entrepreneur's point of view, who attempts to conduct an ICO, our findings provide some kind of guidance. It seems promising to engage actively in social media activities before the token sale. Moreover, a short ICO duration, high share of tokens retained by the ICO initiator, as well as a low bonus provided to investors are effective signals that can increase funding success.

Furthermore, with respect to policy implications, our findings emphasize investors' differentiated assessment of potential investment opportunities in the ICO context. However, we highlight investors' indifference regarding the provision of legal disclaimers and potential risk factors. Against the background of the high number of scams in the ICO market, it is necessary to sensitize potential investors to underlying risks.

8. References

Adhami, S., Giudici, G., & Martinazzi, S. (2018). Why do businesses go crypto? An empirical analysis of initial coin offerings. *Journal of Economics and Business*, *100*(2010), 64–75.

Ahlers, G. K. C., Cumming, D., Günther, C., & Schweizer, D. (2015). Signaling in Equity Crowdfunding. *Entrepreneurship Theory and Practice*, *39*(4), 955–980.

Ahlstrom, D., & Bruton, G. D. (2006). Venture Capital in Emerging Economies: Networks and Institutional Change. *Entrepreneurship Theory and Practice*, *30*(2), 299–320.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500.

Ante, L., Sandner, P., & Fiedler, I. (2018). Blockchain-Based ICOs: Pure Hype or the Dawn of a New Era of Startup Financing? *Journal of Risk and Financial Management*, *11*(4), 80–98.

Arnold, T., Fishe, R. P. H., & North, D. (2010). The Effects of Ambiguous Information on Initial and Subsequent IPO Returns. *Financial Management*, *39*(4), 1497–1519.

Bandura, A. (2010). Self-Efficacy. In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc.

Barbi, A. M., & Mattioli, S. (2019). Human capital, investor trust, and equity crowdfunding. *Research in International Business and Finance*, *49*, 1–12.

Barsan, I. M. (2017). Legal Challenges of Initial Coin Offerings (ICO). *Revue Trimestrielle de Droit Financier (RTDF)*, *3*, 54–65.

Bates, T. (1990). Human Capital Inputs and Small Business Longevity. *The Review of Economics and Statistics*, *72*(4), 551–559.

Baum, J. A. C., & Silverman, B. S. (2004). Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*, *19*(3), 411–436.

Baum, J. R., & Locke, E. A. (2004). The Relationship of Entrepreneurial Traits, Skill, and Motivation to Subsequent Venture Growth. *Journal of Applied Psychology*, *89*(4), 587–598.

Beier, M., & Wagner, K. (2015). Crowdfunding Success: A Perspective from Social Media and E-Commerce. *International Conference of Information Systems 2015*, 1–16.

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, *29*(5), 585–609.

Bolotaeva, O. S., Stepanova, A. A., & Alekseeva, S. S. (2019). The Legal Nature of Cryptocurrency. *IOP Conference Series: Earth and Environmental Science*, *272*(3), 1–5.

Brooks, R., Fry, T. R. L., Dimovski, W., & Mihajilo, S. (2009). A duration analysis of the time from prospectus to listing for Australian initial public offerings. *Applied Financial Economics*, *19*(3), 183–190.

Cassar, G., & Friedman, H. (2009). Does self-efficacy affect entrepreneurial investment? *Strategic Entrepreneurship Journal*, *3*(3), 241–260.

Chen, K. (2019). Information asymmetry in initial coin offerings (ICOs): Investigating the effects of multiple channel signals. *Electronic Commerce Research and Applications*, *36*(May), 1–11.

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling Theory: A Review and Assessment. *Journal of Management*, *37*(1), 39–67.

Cosh, A., Cumming, D., & Hughes, A. (2009). Outside Enterpreneurial Capital. *The Economic Journal*, *119*(540), 1494–1533.

Courtney, C., Dutta, S., & Li, Y. (2017). Resolving Information Asymmetry: Signaling, Endorsement, and Crowdfunding Success. *Entrepreneurship Theory and Practice*, *41*(2), 265–290.

Debler, J. (2018). Foreign initial coin offering issuers beware: The securities and exchange commission is watching. *Cornell International Law Journal*, *51*(1), 245–272.

Deng, H., Huang, R. H., & Wu, Q. (2018). The Regulation of Initial Coin Offerings in China: Problems, Prognoses and Prospects. *European Business Organization Law Review*, *19*(3), 465–502.

Dimov, D. (2010). Nascent Entrepreneurs and Venture Emergence: Opportunity Confidence, Human Capital, and Early Planning. *Journal of Management Studies*, *47*(6), 1123–1153.

Dobrauz-Saldapenna, G., & Klebeck, U. (2019). Initial Coin Offering— Legal and Regulatory Challenges of Crossing the Borders. *The Journal of Alternative Investments*, *21*(4), 81–94.

Dowlat, S. (2018). *Cryptoasset Market Coverage Initiation: Network Creation*. https://research.bloomberg.com/pub/res/d28giW28tf6G7T_Wr77aU0gDgFQ

Downes, D. H., & Heinkel, R. (1982). Signaling and the Valuation of Unseasoned New Issues. *The Journal of Finance*, *37*(1), 1–10.

Drobetz, W., Momtaz, P. P., & Schröder, H. (2019). Investor Sentiment and Initial Coin Offerings. *The Journal of Alternative Investments*, *21*(4), 41–55.

Duffner, S., Schmid, M. M., & Zimmermann, H. (2009). Trust and Success in Venture Capital Financing-an Empirical Analysis with German Survey Data. *Kyklos*, *62*(1), 15–43.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.

Enyi, J., & Le, N. (2017). *The Legal Nature of Cryptocurrencies in the US and the Applicable Rules*.

Epstein, L. G., & Schneider, M. (2008). Ambiguity, Information Quality, and Asset Pricing. *The Journal of Finance*, *63*(1), 197–228.

EY. (2017). *EY Research: Initial Coin Offerings (ICOs)*. https://www.ey.com/Publication/vwLUAssets/ey-research-initial-coin-offerings-icos/$File/ey-research-initial-coin-offerings-icos.pdf

EY. (2018). *Initial Coin Offering (ICOs) - The Class of 2017 - one year later*. https://www.ey.com/Publication/vwLUAssets/ey-study-ico-research/$FILE/ey-study-ico-research.pdf

Fama, E. F., & Jensen, M. C. (1983). Separation of Ownership and Control. *The Journal of Law and Economics*, *26*(2), 301–325.

FINMA. (2018). *Guidelines for Enquiries Regarding the Regulatory Framework for Initial Coin Offerings (ICOs)*. https://www.finma.ch/en/~/media/finma/dokumente/dokumentencenter/myfinma/1bewilligung/fintech/wegleitung-ico.pdf

Fisch, C. (2019). Initial coin offerings (ICOs) to finance new ventures. *Journal of Business Venturing*, *34*(1), 1–22.

Giudici, G., Guerini, M., & Rossi-Lamastra, C. (2018). Reward-based crowdfunding of entrepreneurial projects: the effect of local altruism and localized social capital on proponents' success. *Small Business Economics*, *50*(2), 307–324.

Gwinner, K. P., Gremler, D. D., & Bitner, M. J. (1998). Relational Benefits in Services Industries: The Customer's Perspective. *Journal of the Academy of Marketing Science*, *26*(2), 101–114.

Hacker, P., & Thomale, C. (2018). Crypto-Securities Regulation: ICOs, Token Sales and Cryptocurrencies under EU Financial Law. *European Company and Financial Law Review*, *15*(4), 645–696.

Halevi, S., & Krawczyk, H. (2006). Strengthening Digital Signatures Via Randomized Hashing. In C. Dwork (Ed.), *Advances in Cryptology - CRYPTO 2006* (Issue 4117, pp. 41–59). Springer.

Haq, M. ul. (1996). *Reflection on Human Development*. Oxford University Press.

Heller Baird, C., & Parasnis, G. (2011). From social media to social customer relationship management. *Strategy & Leadership*, *39*(5), 30–37.

Howell, S. T., Niessner, M., & Yermack, D. (2019). Initial Coin Offerings: Financing Growth with Cryptocurrency Token Sales. *The Review of Financial Studies (forthcoming)*.

Huang, W., Meoli, M., & Vismara, S. (2019). The geography of initial coin offerings. *Small Business Economics (forthcoming)*.

IRS. (2014). *Notice 2014-21*. https://www.irs.gov/pub/irs-drop/n-14-21.pdf

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360.

Kharif, O. (2018). *How's that ICO working Out?* https://www.bloomberg.com/news/articles/2018-12-14/crypto-s-15-biggest-icos-by-the-numbers

Kollmann, T., & Kuckertz, A. (2006). Investor relations for start-ups: an analysis of venture capital investors' communicative needs. *International Journal of Technology Management*, *34*(1/2), 47–62.

Kromidha, E., & Li, M. C. (2019). Determinants of Leadership in Online Social Trading: A Signaling Theory Perspective. *Journal of Business Research*, *97*, 184–197.

Leland, H. E., & Pyle, D. H. (1977). Informational Asymmetries, Financial Structure, and Financial Intermediation. *The Journal of Finance*, *32*(2), 371–387.

Li, J. (2016). Equity Crowdfunding in China: Current Practice and Important Legal Issues. *The Asian Business Lawyer*, *18*(Fall), 59–131.

Lukkarinen, A., Teich, J. E., Wallenius, H., & Wallenius, J. (2016). Success drivers of online equity crowdfunding campaigns. *Decision Support Systems*, *87*, 26–38.

MAS. (2017a). *Consumer Advisory on Investment Schemes Involving Digital Tokens (Including Virtual Currencies)*. http://www.mas.gov.sg/News-and-Publications/Media-Releases/2017/Consumer-Advisory-on-Investment-Schemes-Involving-Digital-Tokens.aspx

MAS. (2017b). *MAS Clarifies Regulatory Position on the Offer of Digital Tokens in Singapore*.

Masiak, C., Block, J. H., Masiak, T., Neuenkirch, M., & Pielen, K. N. (2019). Initial coin offerings (ICOs): market cycles and relationship with bitcoin and ether. *Small Business Economics (forthcoming)*.

Maume, P., & Fromberger, M. (2019). Regulation of Initial Coin Offerings: Reconciling US and EU Securities Laws. *Chicago Journal of International Law Chicago Journal of International Law*, *19*(2), 548–585.

Michael, S. C. (2009). Entrepreneurial signaling to attract resources: the case of franchising. *Managerial and Decision Economics*, *30*(6), 405–422.

MinFin. (2018). *On Digital Financial Assets*. https://www.minfin.ru/ru/document/?id_4=121810&page_id=2104&popup=Y&area_id=4

Momtaz, P. P. (2020). Entrepreneurial Finance and Moral Hazard: Evidence from Token Offerings. *Journal of Business Venturing (forthcoming)*.

Moritz, A., Block, J., & Lutz, E. (2015). Investor communication in equity-based crowdfunding: a qualitative-empirical study. *Qualitative Research in Financial Markets*, *7*(3), 309–342.

Murphy, F. (1946). *U.S. Reports: S.E.C. v. Howey Co., 328 U.S. 293*. https://www.loc.gov/item/usrep328293/

Natarajan, H., Krause, S., & Gradstein, H. (2017). Distributed Ledger Technology and Blockchain. In *FinTech note* (Vol. 1). World Bank.

Naylor, R. W., Lamberton, C. P., & West, P. M. (2012). Beyond the “Like” Button: The Impact of Mere Virtual Presence on Brand Evaluations and Purchase Intentions in Social Media Settings. *Journal of Marketing*, *76*(6), 105–120.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204.

Park, H. D., & Patel, P. C. (2015). How Does Ambiguity Influence IPO Underpricing? The Role of the Signalling Environment. *Journal of Management Studies*, *52*(6), 796–818.

Partz, H. (2020). *Russia Postpones Its Crypto Law Again, Now Blaming Coronavirus*. https://cointelegraph.com/news/russia-postpones-its-crypto-law-again-now-blaming-coronavirus

Paschen, J. (2017). Choose wisely: Crowdfunding through the stages of the startup life cycle. *Business Horizons*, *60*(2), 179–188.

PBC. (2017). *Announcement of the China Insurance Regulatory Commission on Preventing Financing Risks of Token Issuance*. http://www.pbc.gov.cn/goutongjiaoliu/113456/113469/3374222/index.html

Petkova, A. P., Rindova, V. P., & Gupta, A. K. (2013). No News Is Bad News: Sensegiving Activities, Media Attention, and Venture Capital Funding of New Technology Organizations. *Organization Science*, *24*(3), 865–888.

Piva, E., & Rossi-Lamastra, C. (2018). Human capital signals and entrepreneurs' success in equity crowdfunding. *Small Business Economics*, *51*(3), 667–686.

Podolny, J. M. (1993). A Status-Based Model of Market Competition. *American Journal of Sociology*, *98*(4), 829–872.

Prasad, D., Bruton, G. D., & Vozikis, G. (2000). Signaling value to businessangels: The proportion of the entrepreneur's net worth invested in a new venture as a decision signal. *Venture Capital*, *2*(3), 167–182.

Ray, J. J., & Singh, S. (1980). Effects of Individual Differences on Productivity among Farmers in India. *The Journal of Social Psychology*, *112*(1), 11–17.

Roosenboom, P., van der Kolk, T., & de Jong, A. (2020). What determines success in initial coin offerings? *Venture Capital (forthcoming)*.

SEC. (2017). *Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934: The DAO*. https://www.sec.gov/litigation/investreport/34-81207.pdf

Shane, S., & Cable, D. (2002). Network Ties, Reputation, and the Financing of New Ventures. *Management Science*, *48*(3), 364–381.

Shane, S., & Venkataraman, S. (2000). The Promise of Enterpreneurship as a Field of Research. *The Academy of Management Review*, *25*(1), 217–226.

Solomon, M. R. (2012). *Consumer Behaviour: Buying, Having, Being*. Pearson Higher Education AU.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374.

Sprenger, T. O., Sandner, P. G., Tumasjan, A., & Welpe, I. M. (2014). News or Noise? Using Twitter to Identify and Understand Company-specific News Flow. *Journal of Business Finance & Accounting*, *41*(7–8), 791–830.

Statista. (2019a). *Global Comparison - Transaction Value in million US$*. https://www.statista.com/outlook/335/100/crowdfunding/worldwide#market-globalRevenue

Statista. (2019b). *Global Comparison - Transaction Value in million US$*. https://www.statista.com/outlook/377/100/crowdinvesting/worldwide#market-globalRevenue

Statista. (2019c). *Number of Campaigns in thousand*. https://www.statista.com/outlook/335/100/crowdfunding/worldwide#market-users

Statista. (2019d). *Number of Campaigns in thousand*. https://www.statista.com/outlook/377/100/crowdinvesting/worldwide#market-users

Tiwari, M., Gepp, A., & Kumar, K. (2020). The future of raising finance - a new opportunity to commit fraud: a review of initial coin offering (ICOs) scams. *Crime, Law and Social Change*, *73*(4), 417–441.

Trump, B. D., Florin, M.-V., Matthews, H. S., Sicker, D., & Linkov, I. (2018). Governing the Use of Blockchain and Distributed Ledger Technologies: Not One-Size-Fits-All. *IEEE Engineering Management Review*, *46*(3), 56–62.

Unger, J. M., Rauch, A., Frese, M., & Rosenbusch, N. (2011). Human capital and entrepreneurial success: A meta-analytical review. *Journal of Business Venturing*, *26*(3), 341–358.

Van Osnabrugge, M. (2000). A comparison of business angel and venture capitalist investment procedures: An agency theory-based analysis. *Venture Capital*, *2*(2), 91–109.

Venkataraman, S. (1997). The Distinctive Domain of Entrepreneurship Research: An Editor's Perspective. *Advances in Entrepreneurship, Firm Emergence, and Growth*, *3*, 119–138.

Wang, N., Li, Q., Liang, H., Ye, T., & Ge, S. (2018). Understanding the importance of interaction between creators and backers in crowdfunding success. *Electronic Commerce Research and Applications*, *27*, 106–117.

Yu, F. R., Liu, J., He, Y., Si, P., & Zhang, Y. (2018). Virtualization for Distributed Ledger Technology (vDLT). *IEEE Access*, *6*, 25019–25028.

Zacharakis, A. L., & Meyer, G. D. (2000). The potential of actuarial decision models. *Journal of Business Venturing*, *15*(4), 323–346.

9. Appendix

Appendix II-A: Descriptions of variables

Table II-A.1 contains descriptions and construction details of all variables used in this paper.

Table II-A.1: Descriptions of variables

	Variable	Unit	Explanation
ICO Success	*raised mUSD*	USD	Amount raised by the project during the ICO main sale period in million USD.
Human capital	*team size*	#	Number of members in the project team.
	share university degree	%	Share of the team members that hold a university degree.
	number advisors	#	Number of advisors of the ICO project.
Entrepreneurs' self-efficacy	*duration*	Days	Duration of the ICO main sale in days.
	distributed percent	%	Share of tokens that is distributed to the public during the ICO main sale period.
	bonus	%	Maximum bonus that is granted to investors during the ICO main sale period.
Ambiguity reduction	*risks*	1/0	Dummy variable that equals one if there is a section in the ICO white paper that declares potential risk factors of the ICO (investment), and zero otherwise.
	disclaimer	1/0	Dummy variable that equals one if there is a (legal) disclaimer in the ICO white paper, and zero otherwise.
Media presence	*social count*	#	Number of social media platforms the ICO project uses.
	Twitter activity	#	Number of tweets the ICO project posted in 60 days before the start of the ICO main sale.
Controls	*token price*	USD	Price of the issued tokens during the ICO main sale period in USD.
	pre-sale	1/0	Dummy variable that equals one if there was a pre-sale before the ICO main sale, and zero otherwise.
	goal mUSD	USD	Fund raising goal of the ICO project in USD.

Appendix II-B: Data processing procedure

We started the empirical analysis by obtaining an initial sample from the ICO information platform icotracker.net. Our investigation covers the period from July 1, 2014 to January 31, 2018. Based on this initial list of ICO projects, we searched the remaining considered ICO information platforms, namely smithandcrown.com, icodata.io, icobazaar.com, tokendata.io, icobench.com and icodrops.com, for further ICO projects and added them to our initial sample. This process – after manually removing duplicates and ICO pre-sales - resulted in a sample of 1,679 ICO projects. For these ICO projects, we obtain data on raised funds as well as data on ICO projects' country of origin from all aforementioned ICO information platforms. Since not all ICO information platforms provide data for every project in our sample and, in addition, the data points partially overlap, we decided to prioritize the platforms according to their data coverage. Thus, we derived data according to following order: icodata.io, tokendata.io, icobench.com, smithandcrown.com, icodrops.com, icotracker.net and icobazaar.com. Here, the former platform covers 621 ICOs whereas the latter platform provide data on 122 ICOs. Deriving data from seven different ICO information platforms enabled us to retrieve as many ICOs as possible and thus reduce a potential selection bias. As we are interested in the ICOs' funding success, we removed all observations without any information on raised funds. Overall, this procedure resulted in a sample of 1,057 ICO projects.

Based on this sample, we searched for ICO white papers on the ICO projects' websites. If a project's website was not available or did not provide a white paper (or only a white paper that dated from after the ICO period), we searched all aforementioned ICO information platforms for these white papers. Overall, we obtained white papers for 863 ICO projects. We used these white papers to manually derive all of our explanatory variables, except for the variables on media presence. Two independent coders reviewed all white papers and a common consensus was reached in case of differences between the two coders.

Lastly, we investigated the presence of each ICO project on eight different social media platforms, namely Twitter, Facebook, Bitcointalk, Github, Reddit, Telegram, Medium and Slack. Here, for each social media platform, we use a dummy variable that equals one, if an ICO project was represented on the respective platform, and zero otherwise. Additionally, we investigate ICOs' Twitter accounts more deeply to assess the social media activity of ICOs. In particular, we web scraped the Twitter timelines of each ICO project represented on Twitter using R. Overall, we obtained Twitter data containing the content as well as the metadata on each Tweet until January 2018 for 774 ICO projects.

Please see Table II-B.1 for an overview of used data sources and the corresponding derived variables and Figure II-B.2 for the flowchart of the data collection and cleaning process.

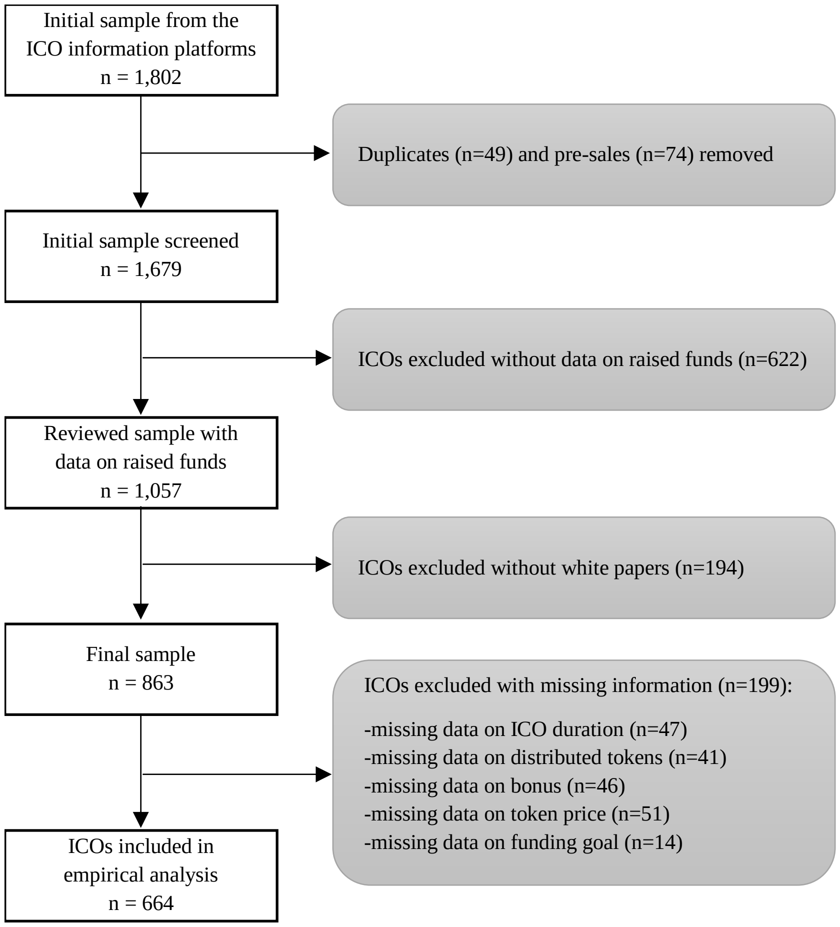
Table II-B.1: Overview of data sources

This table shows our sources of data and derived variables from these sources.

Source	Derived variables
ICO information platforms i.e. icodata.io, tokendata.io, icobench.com, smithandcrown.com, icodrops.com, icotracker.net, icobazaar.com	*raised mUSD* Also: information on ICO projects' country of origin
White Papers	*team size, share university degree, number advisors, duration, distributed percent, bonus, risks, disclaimer, token price, pre-sale, goal mUSD* Also: information on token form and industry
Social media websites i.e. Twitter, Facebook, Bitcointalk, Github, Reddit, Telegram, Medium, Slack	*social count*
Twitter	*Twitter activity*

Figure II-B.2: Flowchart

This figure shows the flowchart of data collection and cleaning process.



III. Among peers: the impact of homophily in online investment

Co-authors: Philipp Ritter, Oscar Stolper
Own share: 40%

Among peers: the impact of homophily in online investment

Daniel Czaja[a] Philipp Ritter[b] Oscar Stolper[c]

Abstract - We investigate if homophily—peoples' affinity for similar others—is an issue in online investment, too. Drawing on nearly 14,000 loan applications and 65,000 investments obtained from one of the leading online peer-to-peer lending platforms in Europe, we document strong evidence in support of a homophily effect on investors' financial decision making. Controlling for a host of alternative determinants, being in the same age group as a given investor increases a loan applicant's odds of being funded by as much as 14%, while same-sex dyads are associated with 6% higher odds of investment. Moreover, any additional demographic similarity increases the average investment amount by nearly 10%. At this, the impact of homophily on investors' funding propensity proves substantially larger for female investors. Finally, we document a –19bp. difference in risk-adjusted interest rates of loans associated with investor-borrower dyads exhibiting the highest versus lowest number of homophilous ties. This evidence is hard to square with the notion that investors' affinity for similar borrowers in online peer-to-peer lending follows economic rationale.

Keywords: Homophily, online investment, peer-to-peer lending, crowdfunding, funding probability, decision heuristics

JEL-Codes: D12, D14, G20, G41

---

[a] Institute of Financial Services, University of Giessen, Daniel.L.Czaja@wirtschaft.uni-giessen.de.

[b] Institute of Accounting and Finance, University of Marburg, Philipp.Ritter@wiwi.uni-marburg.de.

[c] Institute of Accounting and Finance, University of Marburg, Oscar.Stolper@wiwi.uni-marburg.de.

1. Introduction

In this study, we investigate if the impact of homophily—a sociological principle which has been shown to affect agents' decision making on offline financial markets (e.g., Hegde and Tumlinson, 2014; Stolper and Walter, 2019)—spills over to online investment. Homophily denotes the phenomenon that people are more likely to form relational ties with others who are similar to themselves than with those who are not (Lazarsfeld and Merton, 1954).[1] This affinity for similar others is a fundamental principle underlying human relationships. On the one hand, it explains why people's personal networks are surprisingly homogeneous across sociodemographic characteristics and behavioral traits. On the other hand, homophily confines people's individual social environments and affects the information they obtain, the mindset they build, and the way they perceive interactions (e.g., Burt, 1982; Friedkin, 1993; Lawrence, 2006)

Specifically, homophily fosters mutual understanding in peoples' interactions with others in private life but also in business relations. Thus, in our context of online peer-to-peer lending, homophily implies that when an investor identifies a given borrower as being similar to her, a shift in her normative expectations should lead to a benevolent attitude toward the borrower such that the investor sees her behavior in a favorable light (Mills and Clark, 1982; Uzzi, 1999; Silver, 1990). Hence, all else equal, we thus expect investors to display a higher propensity to fund borrowers with whom they share more homophilous ties.

What is yet unanswered, however, is whether homophily requires in-person interaction to be able to affect individuals' financial decision-making. If so, online peer-to-peer lending would carry the potential to mitigate the impact of homophily on individuals' behavior. By addressing this question empirically, this study seeks to improve our understanding of the dynamics and outcomes pertaining to one of the fastest-growing financial innovations for retail investors in recent years. In fact, ever since the launch of the first platforms in 2005, peer-to-peer loans have quickly evolved into an important alternative finance vehicle and represent a worldwide market volume of nearly 280 billion USD as of 2018

[1] See McPherson et al. (2001) for a review of the extensive body of literature on homophily and Lawrence and Shah (2020) for a taxonomy of the most commonly used empirical measures of homophily based on a survey of research in homophily from 1954 through 2018.

(Ziegler et al., 2020). More generally, with an increasing number of retail financial services now provided virtually, the question if and to what extent behavioral patterns observed in offline financial markets spill over to the online sphere is highly relevant to help assess the potential of online solutions when it comes to enhancing market efficiency.

In order to examine a potential impact of homophily in online investment, we draw on detailed data from one of the leading online peer-to-peer lending platforms in Europe, which includes a total of 13,721 loan applications and 64,730 investments. Our identification strategy is straightforward. Following extant research applying the concept of homophily in business economics, we capture potential homophilous ties via demographic commonalities shared by clients and advisors, i.e. unambiguously and systematically available data. Moreover, our dataset allows us to apply the potential-dyads method in order to measure investors' funding choices, which does not focus on analyzing choice outcomes but instead explicitly incorporates all investment alternatives available to agents at the time they make their funding decision.

We document strong evidence in support of a homophily effect in online peer-to-peer lending. Even after controlling for a battery of alternative determinants, one additional demographic similarity in a potential investor-borrower dyad increases the investor's odds of funding the corresponding loan application by more than 11% while the average investment amount grows by nearly 10%. Moreover, we observe a significant gender gap in the strength of the documented homophily effect on investors' likelihood of investment: regardless of the homophily dimension under review, the impact of investor-borrower sameness on the propensity to fund a loan application turns out to be substantially larger for female investors. Finally, we investigate if investor homophily in online peer-to-peer lending reflects economically rational investor decisions or instead shows a behavior, which suggests systematic deviations from economic optimality. In the corresponding analysis, we observe a statistically and economically meaningful difference of -19 basis points (bp) in risk-adjusted interest rates when we compare loans of investor-borrower dyads exhibiting the highest versus lowest number of homophilous ties. This finding is hard to square with the notion that investors' higher likelihood of funding similar borrowers in online peer-to-peer lending is primarily based on economic rationale.

Our study contributes to at least three different strands of literature. First, we contribute to research investigating the market efficiency of online peer-to-peer lending which has received increasing attention in recent years. In fact, several contributions point to significant inefficiencies in peer-to-peer lending and highlight the relevance of investment determinants unrelated to economic fundamentals. Pope and Sydnor (2011) document a lower average funding probability for African-American borrowers on peer-to-peer lending platforms while Ravina (2008) finds that white borrowers and those perceived to be more physically attractive are more likely to receive funding. Moreover, Pope and Sydnor (2011) and Duarte et al. (2012) find higher funding chances for women on US online peer-to-peer lending platforms; however, this evidence of gender discrimination could not be confirmed for the German market (Barasinska and Schäfer, 2014). Similarly, Herzenstein et al. (2011) report a higher relevance of uninspected information such as loan descriptions as compared to verified hard facts such as the loan's interest rate. In a related study, Dorfleitner et al. (2016) analyze the length of the text describing a given loan application as well as the number of typos in it and find that, while this soft information is unrelated to the average probability of default, it still affects the probability of successful funding. Finally, Lin and Viswanathan (2016) recently document that locally biased investments—a robust phenomenon on offline financial markets (cf., e.g., Baltzer et al., 2013; Ivković and Weisbrenner, 2005)—turn out to be an issue in online peer-to-peer lending as well. However, the authors confine their analysis to the potential impact of geographic proximity, while we are the first to take a holistic approach towards investigating the impact of homophily on individuals' online investment decisions.

Second, we add to recent research documenting an impact of homophily on agents' economic choices on offline markets. Hegde and Tumlinson (2014) show that U.S. venture capitalists are more likely to invest in start-ups with coethnic executives. Jaspersen and Limbach (2018) find that, even when controlling for educational background and regional proximity, mutual fund managers are found to overweight firms led by CEOs who resemble them in terms of age, ethnicity and gender. Only recently, Stolper and Walter (2019) find that homophily has a significantly positive impact on the likelihood of following financial advice. Retail investors' increased likelihood of following stems from homophily on gender and age for male clients and from sameness on marital and parental status for female advisees. We extend this literature by investigating if sociodemographic similarities still govern peoples' financial decision making when they do not manifest saliently in physical interactions.

More broadly, our results extend evidence of homophily on virtual peer-to-peer platforms in other areas of life. In early exploratory studies of the similarity of dyads on social network sites, Thelwall (2009) and Baym and Ledbetter (2009) find that relationships on Myspace and Last.fm, respectively, are strongly characterized by homophily. Wang et al. (2008) document homophily grounded credibility perceptions of online health information, which drive the persuasive process. This positive impact of homophily on source credibility has been confirmed by Wright (2000) for online support providers outside the health domain. Centola and van de Rijt (2015) study how people select their peer contacts in online health communities and show that participants disregard information about others' fitness levels, exercise preferences, and workout experiences but instead select partners almost entirely on the basis of similarities on gender and age, i.e. common sources of homophily in offline relationships. Consistent with these findings, we document that homophily affects individuals' *financial* decisions on virtual peer-to-peer platforms, too.

2. Data and key variables

2.1 Platform details

We obtain data from one of the leading online peer-to-peer lending platforms in Europe, which operates in Germany. To register to this online market for unsecured personal loans, borrowers and investors first share their personal data, the accuracy of which is verified by the platform based on the applicants' official ID. As part of the registration process, applicant borrowers also authorize the platform to retrieve their individual credit rating from Germany's FICO-equivalent, i.e. the credit rating agency Schufa.[2] Applicant

[2] Schufa is Germany's quasi-monopolist in the provision of consumer credit ratings and maintains scores for approximately 70 million Germans, i.e. 85% of the country's total population (Schufa Holding AG, 2019). Schufa analyzes consumers' financial behavior in order to assess their default risk and computes a score which takes one of 15 different values ranging from A ('excellent') to P ('very poor'). The corresponding algorithm is a trade secret of Schufa.

borrowers featuring a Schufa score below K (corresponding to a probability of default in excess of 27.01%) are denied access to the platform.

Upon successful registration, the borrower can submit a loan application by determining the amount, duration, category as well as a free-text header shown in the listing. Moreover, she can choose to add an individual loan description and a picture of (i) herself, (ii) the project or object to be funded or, alternatively, (iii) a random illustration displayed along with the loan application. If no edits are made, the platform automatically assigns a loan description based on the loan category and adds a random pictogram. Importantly, the platform also sets the loan's interest rate, which predominantly derives from the borrower's Schufa credit rating. Borrowers have no means of adjusting the interest rate on their own.[3]

Investors can browse all published loan applications, with the most recently accepted loans showing up first. As is common on online peer-to-peer lending platforms, loans need not be funded entirely by a single investor. Instead investors place a bid of at least 250 € per loan for as many loans as they wish to hold a stake in.[4] Given the comparably high minimum bid amount (on Prosper.com, e.g., minimum bids are as low as $50) and a mean bid amount of 436 €, we are confident that investors do not simply use 'play money' but instead use this online lending marketplace as a serious investment vehicle. The platform allows a 14-day period to have investors fund the loan. In case of full funding within 14 days or less, the loan amount is immediately paid out to the borrower. If more

[3] Note that this means of interest rate determination differs from auction-like mechanisms on other large online peer-to-peer lending platforms. On Prosper.com, e.g., investors specify the minimum interest rate at which they are willing to lend. Funds from different investors are then pooled to determine the lowest possible interest rate the borrower will pay. At 0.7895, *Borrower credit score* is highly correlated with *Loan interest rate*, but does not explain it entirely. Interest rate calculation is not fully reproducible since the platform does not disclose the relevant algorithm. Competitors state that borrower information such as user data collected online and earlier repayments feed into the calculation of interest rates. Such additional data is unavailable for the platform under review.

[4] Note that the platform under review offers two types of investment: the investor may either choose for herself which loan to invest in (self-directed investment). Alternatively, the investor determines the amount of money to be invested, minimum borrower credit scores and a target return on the investment. Based on these parameters, the platform automatically assembles a loan basket (delegated investment). Since we are interested in how homophily potentially affects individuals' investment behavior, we omit delegated investments.

than 75% of the loan is fund via investors, the platform bridges the funding gap and co-finances the loan. Otherwise, the loan is delisted.

The platform provides investors with a wide range of information on borrowers and their loan applications. Specifically, each loan application includes the loan amount and respective interest rate, its duration, current financing status along with the number of submitted bids as well as an assignment to one of the preset loan categories and, if applicable, a description of the loan and the borrower's profile picture.[5] Moreover, each loan application discloses the borrower's age, gender and state of residence as well as her Schufa credit score to provide investors with an indication of her default risk. Borrowers cannot bypass the publication of this information. The platform does not offer a private chat function, which ensures that all interactions between investors and borrowers are captured in our dataset. Moreover, borrowers' names or residential addresses are not disclosed either. Thus, investors have no means to enforce debt redemption by, e.g., paying a private visit to borrowers and therefore, geographic proximity to the borrower should not be a rational input parameter to the investor's decision on which loan to fund (c.f. Lin and Viswanathan, 2016). Instead, any subsequent repayment issues or defaults are handled directly by the platform and an associated debt collection agency.

2.2 Sample

We draw on detailed records of loan applications and corresponding investor bids for the twelve-year period from March 2007 until October 2018. Our dataset includes a total of 13,721 loan applications and 64,730 bids.

Panel A of Table III-1 reports summary statistics of these loans and associated bids.[6] The median loan amounts to 5,000 € and carries an interest rate of 6.5% and a duration of 60 months. Loan amounts vary considerably from 500 up to 50,000 € and interest rates also spread widely from 2.0% up to 18.0%. The mean funding status of 97.7% 14 days

[5] Loans may be assigned to any of a total of 22 different loan categories. In the vein of Lin and Viswanathan (2016), we group these categories in seven broader clusters (i) debt consolidation, (ii) higher education, (iii) car purchase, (iv) home improvement, (v) startup capital, (vi) leisure, and (vii) other. We compare either of these subject clusters to the remaining pool of unassigned loans.

[6] Table III-A.1 in Appendix III-A provides descriptions of all variables used in the analysis.

after loan publication shows a high investment interest and implies that the platform hardly ever tops off investor money to fill a funding gap.

Panel B of Table III-1 provides descriptive statistics of investors and borrowers in our sample. In October 2018, the median borrower is aged 49 years and applies for a single loan amounting to 5,250 €. The median investor is slightly younger (47 years) and submits six bids summing up to a total investment of 2,000 €. At 25% female borrowers and 7% female investors, gender proportions are comparable to what has been documented in prior research on peer-to-peer lending platforms (Barasinska et al., 2011; Dorfleitner et al., 2016; Duarte et al., 2012).[7] Moreover, Barasinska et al. (2011) show for Germany that users of online peer-to-peer lending platforms have become increasingly similar to borrowers on the regular market for consumer credit with respect to gender, age and geographic dispersion of their places of residence. We square our data with demographics obtained from the latest wave of Germany's major household survey, the Socio-Economic Panel (SOEP), and corroborate the findings of Barasinska et al., (2011).

Table III-1: Summary statistics

This table presents the descriptive statistics of our sample. See Table III-A.1 in Appendix III-A for detailed variable descriptions.

	N	Mean	SD	Min	p25	Median	p75	Max
Panel A: Loans and bids								
Loan amount	13,721	8,116	8,887	500.0	1,500	5,000	10,500	50,000
Loan interest rate	13,721	7.300	3.039	2.000	5.000	6.500	9.000	18.00
Loan duration	13,721	51.25	15.65	36.00	36.00	60.00	60.00	84.00
Loan financing status	13,721	97.74	12.13	0.000	100.0	100.0	100.0	100.0
Bid amount	64,730	435.7	429.4	250.0	250.0	250.0	500.0	19,000
Bids per loan	64,730	17.74	18.91	0.000	5.000	11.00	24.00	133.0
Panel B: Borrowers and investors								
Borrower age	11,705	49.26	12.66	20.00	40.00	49.00	57.00	100.0
Borrower credit score	13,721	3.437	2.348	1.000	1.000	3.000	5.000	10.00
Number loans per borrower	11,705	1.150	0.529	1.000	1.000	1.000	1.000	19.00
Total bid amount per borrower	11,705	9,756	11,719	500.0	1,500	5,250	13,500	111,000
Investor age	7,386	48.08	11.79	19.00	39.00	47.00	56.00	93.00
Number bids per investor	7,386	10.75	17.65	1.000	2.000	6.000	14.00	452.0
Total funding per investor	7,386	4,594	8,537	250.0	750.0	2,000	5,250	180,500

[7] See Table III-B.1 in Appendix III-B for a detailed breakdown of borrowers and investors by gender.

2.3 Measuring the impact of homophily

Following related research in business economics investigating the impact of homophily (c.f., e.g., Hegde and Tumlinson, 2014; Hwang and Kim, 2009; Stolper and Walter, 2019), we capture potential homophilous ties by means of demographic similarities between investors and borrowers. As outlined in section 1, demographic attributes are necessary for homophily to evolve since they provide people with salient features allow them to detect elements of similarity in others. Peoples' race and ethnicity, gender and age as well as their religion, education, occupation and social status have been identified as the most important dimensions of homophily (McPherson et al., 2001).

Germany is one of the most ethnically homogenous countries in the world (Alesina et al., 2003), religious homophily has decreased in relevance over time (Kalmijn, 1998), and evidence suggests that homophilous ties between confidants are less likely to evolve based on education and occupation as compared to other demographics (Louch, 2000). Consequently, we focus on the most salient homophily dimensions age and gender in our analysis. Moreover, we include the major cause of homophily, i.e. regional proximity between investor and borrower.[8] The evidence in Huberman (2001), Lai and Teo (2008) and Strong and Xu (2003) suggests that geographical propinquity—regardless of whether there is actual economic benefits to it—promote trust and a disproportionately benevolent disposition toward local transaction partners. Corroborating the (behavioral rather than economically rational) relevance of geography for investors in peer-to-peer lending, Lin and Viswanathan (2016) show that homophily at least partly drives the home bias in funding choices on Prosper.com.

We construct the indicator variable *Same gender,* which assumes a value of 1 if both investor and borrower are female or male, respectively. Consistent with the relatively low proportion of female investors and borrowers, the fraction of same-sex dyads in our data amounts to 76.3%. Next, following Stolper and Walter (2019), we capture age similarity with the binary variable *Same age*, which equals 1 if the absolute age difference between investor and borrower is no larger than 5 years, which applies for 28.9% of the sampled bids. Finally, the variable *Same state* indicates whether investor and borrower live in the

[8] See McPherson et al. (2001) for an excellent review of the various demographic characteristics and their role in breeding homophily.

same federal state. Germany is divided into 16 federal states and, as described in section 2.2, the distribution of borrower and investors in our dataset is largely representative of the population distribution in Germany. 13.6% of the sampled bids are submitted by investors who live in the same state as the borrower funded by the bid.

To quantify the intensity of homophily between investor and borrower, we combine the three sociodemographic dimensions in one measure. In the vein of Girard et al. (2015) and Stolper and Walter (2019), we construct the variable *Number of homophilous ties* counting the number of similarities which investor and borrower share in terms of age, gender, and state of residence. Straightforwardly, a value of zero indicates nonexistent homophilous ties, whereas values of 1 to 3 describe the intensity of homophilous ties between investor and borrower, where a higher value corresponds to a stronger link.

Figure III-1: Distribution of *Number of homophilous ties*

This figure plots the distribution of our key variable measuring the number of homophilous ties between borrower and investor. Homophilous ties are *Same age*, *Same state* and *Same gender*. Relative frequencies pertain to the total number of submitted bids (N=64,730).

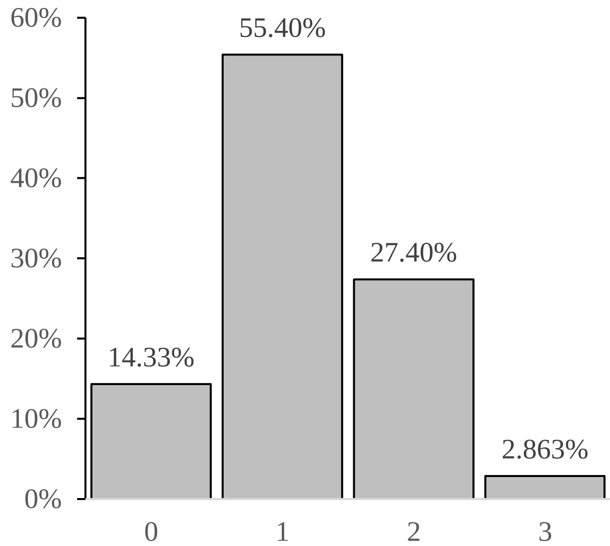


Figure III-1 plots the distribution of *Number of homophilous ties*. In more than 85% of dyads under review, investors fund borrowers with whom they share at least one commonality. 27% (3%) of bids feature investor-borrower dyads with two (the maximum of three) homophilous ties.

2.4 A dyadic analysis of investors' funding choices

To examine the impact of homophilous ties on investors' propensity to fund a given loan, we follow Sorenson and Stuart (2001) and Lin and Viswanathan (2016) and apply the potential-dyads approach, which is not confined to studying choice outcomes but instead explicitly models the alternatives available to agents at the time of their choice. Specifically, we juxtapose each empirically observable bid in our sample with all loan applications that were available for investment at the time a given bid was submitted. To illustrate how we proceed, assume we identify *n* potential loans available to a given investor at the time she submits her bid. In a first step, we construct a binary variable *Dyad_match_d*, which, straightforwardly, equals one if an investment was made and zero for the remaining *n-1* potential loan applications the investor could have funded but chose not to. During our period under review, we identify a total of 676,683 investable loan applications – i.e. possible credit relationships between investors and borrowers – which lead to 64,730 actual investor bids. Next, we then relate our key explanatory variable *Number of homophilous ties* to the likelihood of each potential investor-borrower dyad being realized.

2.5 Descriptive evidence

Are individual investors on online peer-to-peer lending platforms more likely to invest in loan applications by borrowers with whom they share more homophilous ties?

Table III-2: Investor-borrower dyads by *Number of homophilous ties*

This table presents the distribution of potential dyads and matching dyads organized by the number of homophilous ties. Percentages presented in the rightmost column indicate the proportion of matching dyads to potential dyads by *Number of homophilous ties*. The bottom line shows the difference between the proportion of matching dyads in case investors and borrowers feature three homophilous ties and zero homophilous ties, respectively. *** indicates statistical significance at the 1% level.

Number of homophilous ties	N potential dyads	*Dyad_match_d*	
		N	%
0	114,646	9,278	8.093
1	376,553	35,862	9.550
2	168,876	17,737	10.50
3	16,608	1,853	11.16
All	676,683	64,730	9.566
Diff. 3 – 0			3.065 pp.***
(χ^2)			(175.5)

We begin our analysis with a discussion of descriptive evidence reported in Table III-2. While the 114,646 potential dyads in which investors and borrowers do not share a single common homophily trait result in only 8.1% actual investments, this proportion climbs to 11.2% for the subgroup of dyads featuring the highest possible number of sociodemographic similarities between investors and borrowers. Thus, the unconditional difference in *Dyad_match_d* amounts to a highly significant 3.1 pp. when we compare investors and borrowers displaying the most versus least intense homophilous ties. Moreover, we document that, unconditionally, investors' propensity to fund a loan increases monotonically with every additional demographic commonality. In sum, the results presented in Table III-2 provide preliminary evidence in support of the hypothesis that the number of homophilous ties between investors and borrowers is positively associated with the likelihood of funding a given loan application.

3. Regression results

3.1 Model

Next, we investigate if the positive relationship between the number of homophilous ties present in a given investor-borrower dyad and the investor's propensity of funding the associated loan persists once we control for a battery of loan attributes and borrower characteristics previously shown to determine the likelihood of funding on peer-to-peer lending platforms. More specifically, we estimate the following logistic regression model

$$Dyad\_match\_d_{i,j,k} = \alpha_i + \beta Number\ of\ homophilous\ ties_{j,k} + \delta' m_{i,k} \quad (1)$$

$$+\gamma' c_{i,j} + \theta' v_i + \varepsilon_{i,j,k}$$

where $Dyad\_match\_d_{i,j,k}$ denotes our binary variable indicating whether an investor j invests in the loan application i of borrower k. *Number of homophilous ties*$_{j,k}$ counts the number of commonalities along the three homophily dimensions between investor j and borrower k. To gauge the impact of either of the three homophily dimensions separately, we estimate a related specification of the model in which we substitute *Number of homophilous ties* by (i) *Same age* indicating whether the investor and the borrower are in the same age group, (ii) *Same state* indicating whether both parties reside in the same federal state, and (iii) *Same gender* indicating same-sex dyads, respectively.

We draw on prior research on online peer-to-peer investment to identify the relevant control variables and include vectors of borrower characteristics ($m_{i,k}$) and dyadic-level information (v_i) available for each loan application i. Specifically, controls include the borrower's job and credit score, the amount of money the borrower is applying for, the loan's interest rate and duration, the loan category assigned by the borrower, and whether she has specified a loan description and picture. Moreover, we control for investor experience as proxied by the number of bids she made before considering loan application i ($c_{i,j}$). In supplementary specifications, we additionally include investor fixed-effects to control for time-invariant unobserved heterogeneity across the investors under review. Finally, we account for the potential investment universe at the time the investor submits her bid as well as the listing position of loan application *i* to control for a potential placement effect as proposed in Jacobs and Hillert (2016).[9]

3.2 Homophily and investment propensity

Table III-3 reports results obtained from various specifications of the logistic regression model formalized in equation (1). We find that the descriptive evidence of homophily affecting investment propensity as reported in section 2.5 persists once we allow for the impact of several other explanations of investment propensity. The first column of Table III-3 shows that, controlling for a host of alternative determinants, one additional homophilous tie in a potential investor-borrower dyad increases the odds that the investor submits a bid on the respective loan application by more than 11% (i.e., $e^{0.1063}$-1), which is not only highly significant in statistical terms but economically meaningful, too. At peak, the odds of an investor submitting a bid to a loan application of a borrower with whom she shares commonalities along all three homophily dimensions are about 38% ($e^{3*0.1063}$-1) higher than the odds to invest in a loan of an applicant borrower with whom she shares no homophilous ties.

[9] We estimate the model using heteroscedasticity-consistent robust standard errors and add month-year fixed effects to capture potential seasonalities.

Table III-3: Homophily and investment propensity

This table reports coefficient estimates obtained from various specifications of a logistic regression of *Dyad_match_d* on *Number of homophilous ties* and *Same gender*, *Same state*, and *Same age*, respectively, as well as a comprehensive set of control variables. See Table III-A.1 in Appendix III-A for detailed variable descriptions and section 3.1 for a detailed description of the regression model. Heteroscedasticity-consistent robust standard errors are reported in parentheses. *** indicates statistical significance at the 1% level.

	(1)	(2)	(3)	(4)
Number of homophilous ties	0.1063***		0.1117***	
	(0.006434)		(0.006814)	
Same gender		0.05987***		0.05852***
		(0.01073)		(0.01141)
Same state		0.1348***		0.1447***
		(0.01331)		(0.01402)
Same age		0.1305***		0.1386***
		(0.009993)		(0.01064)
Borrower credit score	-0.7905***	-0.7906***	-0.8226***	-0.8229***
	(0.005643)	(0.005644)	(0.005621)	(0.005623)
Loan amount log	-0.06701***	-0.06630***	-0.07532***	-0.07464***
	(0.007787)	(0.0078)	(0.007921)	(0.007943)
Loan interest rate log	5.746***	5.746***	5.991***	5.991***
	(0.04271)	(0.04272)	(0.04351)	(0.04351)
Loan duration	-0.02240***	-0.02244***	-0.02351***	-0.02354***
	(0.0004854)	(0.0004856)	(0.0005001)	(0.0005003)
Loan description automatic	-0.2460***	-0.2465***	-0.2459***	-0.2466***
	(0.01412)	(0.01412)	(0.01440)	(0.01443)
Loan photo manual	0.1003***	0.1008***	0.1055***	0.1059***
	(0.01580)	(0.01581)	(0.01590)	(0.01591)
Loan rank	-0.1494***	-0.1494***	-0.1449***	-0.1449***
	(0.002031)	(0.002031)	(0.001411)	(0.001412)
Number previous bids	-0.002509***	-0.002518***	-0.003542***	-0.003544***
	(0.0004559)	(0.0004559)	(0.0008104)	(0.0008104)
Simultaneously available loans	-0.07778***	-0.07782***	-0.07824***	-0.07830***
	(0.001354)	(0.001354)	(0.001413)	(0.001424)
Loan startup capital	-0.05450***	-0.05310***	-0.05901***	-0.05733***
	(0.01749)	(0.01749)	(0.01803)	(0.01802)
Loan leisure	-0.06655***	-0.07081***	-0.06220***	-0.06661***
	(0.02333)	(0.02335)	(0.02383)	(0.02384)
Loan debt consolidation	-0.1475***	-0.1471***	-0.1579***	-0.1575***
	(0.01579)	(0.01579)	(0.01634)	(0.01634)
*Loan home improvemen*t	-0.01049	-0.01013	-0.01993	-0.01962
	(0.01753)	(0.01753)	(0.01801)	(0.01801)
Loan car purchase	-0.3006***	-0.2988***	-0.3087***	-0.3065***
	(0.02211)	(0.02212)	(0.02233)	(0.02240)
Loan higher education	-0.02987	-0.03436	-0.03270	-0.03752
	(0.03220)	(0.03222)	(0.03270)	(0.03273)
Loan other	0.2907***	0.2896***	0.2842***	0.2832***
	(0.05960)	(0.05952)	(0.06064)	(0.06062)
Borrower job official	0.08723***	0.08706***	0.09280***	0.09290***
	(0.02588)	(0.02589)	(0.02672)	(0.02671)
Borrower job self-employed	0.1309***	0.1329***	0.1290***	0.1315***
	(0.01184)	(0.01186)	(0.01222)	(0.01224)
Borrower job other	0.03976	0.03906	0.03080	0.02901
	(0.1126)	(0.1124)	(0.1138)	(0.1138)
Constant	18.67***	18.67***		
	(0.3009)	(0.3008)		
Time FE	YES	YES	YES	YES
Investor FE	NO	NO	YES	YES
Observations	676,683	676,683	676,549	676,549
Number of Investors	7,386	7,386	7,298	7,298
Pseudo R-Squared	0.2199	0.2200	0.1970	0.1972
Wald chi2	52,783	52,800	75,094	75,127
Prob.	0.0000	0.0000	0.0000	0.0000

Moreover, coefficient estimates pertaining to the individual homophily dimensions show a statistically and economically significant impact of the respective ties at the individual level. The effect size of *Same state* corroborates the results of Lin and Viswanathan (2016) for investor-borrower dyads in their study of home bias on the US peer-to-peer lending platform Prosper.com. While they show that living in the same federal state as the investor improves a given borrower's odds of being funded by approximately 21%, we report an increase of 14.4%. Beyond homophily induced by geographic proximity, however, we provide novel evidence for homophily effects with respect to the investor's age group and gender. Specifically, being in the same age group as a given investor increases a loan applicant's odds of funding by as much as 14.0%, while same-sex dyads are associated with 6.0% higher odds of investment.

Direction and effect size of the various control variables included in the model largely echo prior research. Intuitively, we document that a higher interest rate and a better risk score, respectively, increase the likelihood of investment, which ties in with, e.g., Duarte et al. (2012), Lin et al. (2013) and Lin and Viswanathan (2016). Moreover, consistent with the findings in Dorfleitner et al. (2016) and Herzenstein et al. (2011), requested loan amount and loan duration negatively impact funding probability. Our data also allows us to investigate the potential impact of soft information, i.e. whether a borrower chose to provide an individual loan description and photo along with the loan application. Confirming prior evidence, we find that investors prefer customized borrower input over automatically generated content (e.g., Iyer et al., 2016) and that loan applications listed at the top are significantly more likely to be selected by investors (e.g., Jacobs and Hillert, 2016). Moreover, investors on average seem to prefer loans requested by officials and self-employed persons to those by employees. Somewhat surprisingly, assigning a loan to any of the subject categories provided by the platform, if anything, lowers the likelihood of being funded.[10] Finally, results remain virtually unchanged after controlling for time-invariant unobserved heterogeneity across investors by estimating both specifications of the logistic regression model including investor fixed effects.

[10] *Loan miscellaneous* pools unassigned loan applications and represents the baseline category in the regression model. Note that we group the 22 preset loan categories in seven subject clusters as described in section 2.1 and footnote 36.

3.3 Gender differences in the impact of homophily on investment propensity

While previous research points to considerable gender differences in the strength of homophily, the evidence is largely inconclusive. On the one hand, early research by Levinson et al. (1984) shows that, unlike female patients, males prefer their doctors to be men, regardless of the health problem and those health problems associated with a greater degree of physical intimacy result in an even stronger homophily among male patients. Similarly, evidence from the organizational behavior literature suggests that men are more likely to form homophilous ties and to have stronger homophilous ties than women (e.g., Ibarra, 1992). At the same time, both men and women prefer to consult with men when acquiring information related to more distant domains (Aldrich, 1989; Bernard et al., 1988). With respect to our setting, this implies that gender homophily, in particular, could be limited to male investors. Moreover, Brashears (2008) document a greater relative importance of age homophily among men, a finding, which is corroborated in recent research by Stolper and Walter (2019) who show stronger gender homophily and age homophily for male clients of financial advisors and a stronger homophily on family status for female advisees.

By contrast, however, Popielarz (1999) finds that women are more likely than men to belong to sex-segregated groups and, in addition, women's networks are predominantly composed of members of the same age, education, as well as marital and work status, which indicates substantial gender-specific homophily. Moreover, this finding appears to persist even in the absence of physical interaction: Volkovich et al. (2014) investigate gender differences in the dayds formed in an online social network and find that gender homophily is much more pronounced among female users.

Hence, we examine whether or not a gender gap in the strength of homophily can be observed in an online *financial* market, too. To this end, we interact each of the three homophily dimensions with investor gender. Formally, we specify the following modification of the generic model introduced in section 3.1:

$$\begin{aligned} Dyad\ Match\_d_{i,j,k} = {} & \alpha_i + \beta_1[Homophily\ dimension]_{j,k} + \beta_2 Investor\ gender_j \quad (2) \\ & +\beta_3[Homophily\ dimension]_{j,k} \times Investor\ gender_j \\ & +\delta' m_{i,k} + \gamma' c_{i,j} + \theta' v_i + \varepsilon_{i,j,k} \end{aligned}$$

where *Homophily dimension*$_{j,k}$ denotes (i) *Same age* indicating whether the investor and the borrower are in the same age group, (ii) *Same state* indicating whether both parties reside in the same federal state, or (iii) *Same gender* indicating same-sex dyads. The binary variable *Investor gender*$_j$ equals one for male investors and zero for female investors. We add the same control variables and fixed effects as in the baseline model.

Table III-4: Homophily and investment propensity – gender differences

This table reports coefficient estimates obtained from a modification of the generic model introduced in section 3.1, which interacts each of the three homophily, dimensions with investor gender. The second and third column show the impact of each homophily dimension on *Dyad_match_d* by investor gender. The fourth column reports the differences between the coefficient estimates. To obtain the statistical significance of the impact of any of the three homophilous ties on the propensity to invest in a given loan application for male investors $(\beta_1 + \beta_3)$, we rerun the logistic regressions with rescaled values of the indicator variable *Investor gender*$_j$. Heteroscedasticity-consistent robust standard errors are reported in parentheses. *** indicates statistical significance at the 1% level.

	Dependent Variable *Dyad_match_d*				
Homophily dimension	Female investor (β_1)	Male investor $(\beta_1 + \beta_3)$	Diff. (β_3)	N	R^2
Same gender	0.3594***	0.04681***	-0.3126***	676,683	0.2194
	(0.04892)	(0.01143)	(0.05041)		
Same state	0.2984***	0.1270***	-0.1714***	676,683	0.2195
	(0.06120)	(0.01364)	(0.06270)		
Same age	0.2690***	0.1244***	-0.1446***	676,683	0.2197
	(0.04691)	(0.01023)	(0.04803)		

Table III-4 reports the corresponding results.[11] We observe significant gender differences in the strength of the documented homophily effect on investors' likelihood of investment. Specifically, the impact of investor-borrower sameness on the propensity to fund a loan application turns out to be substantially larger for female investors regardless of the homophily dimension under review. Sharing the same gender with the borrower increases the odds to invest by as much as 43% $(e^{0.3594} - 1)$. Similarly, the estimated effect sizes for *Same state* and *Same age* jump to 35% and 31%, respectively. By contrast, coefficient estimates pertaining to male investors essentially replicate our evidence obtained for the overall sample: effect sizes for *Same age* and *Same state* amount to roughly 13% each, while *Same gender* is associated with an increase in the odds of funding of about 5%.

[11] Note that, in order to capture the statistical significance of the coefficient estimates pertaining to $(\beta_1+\beta_3)$, we rerun each regression using rescaled values of the indicator variable *Investor gender*$_j$.

Taken together, our results suggest that the main result of Volkovich et al. (2014), i.e. that gender homophily in online social networks is much more pronounced among women, applies to the context of online peer-to-peer lending, as well.

3.4 Homophily and investment amount

Next, we examine whether homophily, besides altering investors' likelihood of funding a given loan application, affects the amount of money they are willing to invest in the respective loan. To this end, we replace the binary dependent variable *Dyad_match_d* by the euro amount of money, which investor *j* places in a bid pertaining to a loan application *i* of borrower *k*, denoted *Dyad_match_€*. For any potential dyad not resulting in a bid, we set the bid amount to zero. Thus, because our new dependent variable *Dyad_match_€* is censored at zero, we run various specifications of a Tobit regression model including the same explanatory variables as in equation (1).[12]

Table III-5 reports the corresponding results. As can be obtained from specification (1), we find that one additional homophilous tie between investor and borrower increases the average bid amount by nearly 10% (i.e., [(435.70 + 41.18)/435.70] €), which proves an economically relevant effect size. Again, we rerun the regressions for each homophily dimension individually. Consistent with the evidence pertaining to investors' funding propensity reported in section 3.2, the results of specification (2) show that being in the same age group and living in the same state as a given borrower increases investors' average bid amounts by roughly 50 € or 11.5%, while same-gender dyads are associated with a 5.5% increase in investment. Interestingly, the coefficient estimate on *Same state* suggests a more pronounced impact of geography-induced homophily as compared to Lin and Viswanathan (2016), who document an maximum increase of about $16.

[12] We select a Tobit model rather than a linear specification since (i) linear models can produce negative fitted values leading to implausible predictions of bid amount and (ii) the distribution of *Dyad_match_€* is not (approximately) normally distributed and inference from a linear model might thus be impaired (c.f., e.g., Wooldridge, 2010).

Table III-5: Homophily and investment amount

This table presents the results from Tobit regressions of *Dyad_match_ €* on *Number of homophilous ties* and *Same gender*, *Same state*, and *Same age*, respectively, as well as a comprehensive set of control variables. See Table III-A.1 in Appendix III-A for detailed variable descriptions. Heteroscedasticity-consistent robust standard errors are reported in parentheses. *** indicates statistical significance at the 1% level.

	(1)	(2)	(3)	(4)
Number of homophilous ties	41.18***		43.82***	
	(2.689)		(2.702)	
Same gender		23.93***		23.66***
		(4.450)		(4.490)
Same state		52.48***		55.38***
		(5.547)		(5.551)
Same age		50.12***		55.04***
		(4.208)		(4.222)
Borrower credit score	-299.8***	-299.9***	-316.0***	-316.1***
	(4.000)	(4.001)	(4.018)	(4.019)
Loan amount log	-7.379**	-7.114**	-12.72***	-12.44***
	(3.133)	(3.134)	(3.051)	(3.052)
Loan interest rate log	2,184***	2,184***	2,306***	2,306***
	(29.51)	(29.51)	(29.54)	(29.55)
Loan duration	-8.662***	-8.680***	-9.357***	-9.379***
	(0.2173)	(0.2174)	(0.2135)	(0.2137)
Loan description automatic	-96.54***	-96.72***	-96.68***	-96.92***
	(5.936)	(5.937)	(5.756)	(5.758)
Loan photo manual	48.12***	48.41***	48.15***	48.46***
	(6.677)	(6.679)	(6.474)	(6.476)
Loan rank	-46.01***	-45.99***	-45.09***	-45.07***
	(0.8368)	(0.8368)	(0.7930)	(0.7930)
Number previous bids	-1.056***	-1.060***	-1.742***	-1.742***
	(0.1852)	(0.1852)	(0.3347)	(0.3347)
Simultaneously available loans	-27.30***	-27.33***	-27.28***	-27.30***
	(0.6184)	(0.6185)	(0.6611)	(0.6613)
Loan startup capital	-18.41**	-17.91**	-19.98***	-19.40***
	(7.304)	(7.304)	(7.101)	(7.100)
Loan leisure	-25.37***	-26.82***	-24.21***	-25.81***
	(9.609)	(9.618)	(9.343)	(9.350)
Loan debt consolidation	-60.84***	-60.66***	-63.63***	-63.41***
	(6.564)	(6.564)	(6.390)	(6.389)
*Loan home improvemen*t	-5.779	-5.630	-8.262	-8.095
	(7.266)	(7.265)	(7.081)	(7.080)
Loan car purchase	-105.3***	-104.6***	-109.1***	-108.3***
	(9.144)	(9.146)	(8.903)	(8.906)
Loan higher education	-13.30	-14.97	-11.33	-13.11
	(13.23)	(13.23)	(12.83)	(12.83)
Loan other	128.1***	127.5***	124.8***	124.3***
	(24.57)	(24.56)	(23.83)	(23.81)
Borrower job official	22.94**	23.02**	25.30**	25.40**
	(10.41)	(10.42)	(10.15)	(10.15)
Borrower job self-employed	53.18***	53.92***	53.67***	54.52***
	(4.980)	(4.986)	(4.841)	(4.848)
Borrower job other	63.91	63.44	57.50	56.86
	(49.07)	(48.98)	(47.44)	(47.36)
Constant	6,852***	6,853***		
	(150.6)	(150.7)		
Time FE	YES	YES	YES	YES
Investor FE	NO	NO	YES	YES
Observations	676,683	676,683	676,683	676,683
Number of Investors	7,386	7,386	7,386	7,386
Pseudo R-Squared	0.05981	0.05981	0.06423	0.06423
F	53.00	52.40		
Prob.	0.000	0.000		

3.5 Is homophily detrimental to investors?

In this section, we investigate if homophily in online peer-to-peer lending reflects investor rationality in terms of *homo economicus* or, on the contrary, points to heuristic and biased decision making leading to outcomes which deviate from economic optimality. Given the nature of the decision problem—investment under risk—financial theory would impose that agents make funding choices determined by the economic payoffs associated with the available loan applications. It is conceivable that investors disproportionately fund similar borrowers simply because their loans provide better average risk-adjusted returns and not because of a general affinity for similar others unrelated to economic fundamentals. In fact, the investor behavior observed in our study might be rational if the number of homophilous ties were systematically indicative of the borrower's creditworthiness above and beyond her credit rating. However, even if investor-borrower similarity (or else the borrower's homophily dimensions which feed into that similarity) explained components of creditworthiness that are not captured in the borrower's individual credit score, the investor could not learn about this causality, anyway: the lending platform does not disclose the algorithm that underlies the calculation of interest rates. Hence, we assess a given loan's expected risk-return profile from the investor perspective and focus on the information, which is published on the platform, i.e. the borrower's individual credit rating and the interest rate assigned to the loan. We argue that, given the investor's information set at the time of her funding decision, she might rationally overweight borrowers who are similar to her if they feature systematically higher interest rates as compared to less similar borrowers with an identical credit score.[13]

To analyze whether or not homophily carries economic benefits in the context of online peer-to-peer lending, we set the observable interest rate as the outcome variable and investigate if—all else equal—interest rates exhibit systematic variation for investor-borrower dyads featuring different numbers of homophilous ties. At this, we first apply a nearest-neighbor matching procedure to match bids based on the number of homophilous ties between the corresponding investor-borrower dyads.[14] Specifically, we perform an

[13] Ideally, we would have analyzed the ultimate outcome of the sampled loans in order to assess their economic benefit, i.e. whether they were repaid or defaulted. Unfortunately, however, we lack the respective data.

[14] Note that we focus on actually submitted bids and exclude potential dyads in our economic analysis in section 3.5.

exact matching on loan risk as well as on the loan category and apply the nearest-neighbor approach to match the remaining covariates investor age and lending experience proxied by the number of previous bids, loan amount, loan duration, bid amount and month as well as the number of simultaneously available loans.[15] Next, we estimate the average difference in interest rates of the resulting pairs of nearest-neighbor loans within the various loan categories, whose borrowers—importantly—feature the exact same credit risk score.[16]

Table III-6: Loan interest rates by *Number of homophilous ties*

This table presents differences in loan interest rates by *Number of homophilous ties*. The middle column reports the difference between the average loan interest rate charged for the groups specified in the first and second column, respectively. Differences are estimated using the dyadic bid data and applying a nearest-neighbor matching procedure as described in section 3.5. Heteroscedasticity-consistent robust standard errors are corrected for a large-sample bias. N indicates the number of individual investors included in either comparison. *** and ** indicate statistical significance at the 1% and 5% level, respectively.

Group of bids with	Compared to group of bids with	Change in interest rate (pp.)	Robust SE	p-value	N
1 Tie	0 Ties	-0.04658***	0.0001537	0.002439	44,217
2 Ties	0 Ties	-0.05649***	0.0001815	0.001853	26,125
3 Ties	0 Ties	-0.1943***	0.0004940	0.0008388	9,510
2 Ties	1 Tie	-0.008866	0.0001089	0.4158	52,986
3 Ties	1 Tie	-0.08269**	0.0003802	0.02964	34,824
3 Ties	2 Ties	-0.04206	0.0003597	0.2423	18,385

Table III-6 reports the corresponding results. To illustrate, the first row, e.g., reports the average change in interest rates of the group of bids of investor-borrower dyads featuring one homophilous tie compared to the group of bids with zero ties between the contracting parties. The middle column reports the average differences in interest rates between the groups specified in the first and second column, respectively. Interestingly, average differences reveal a *negative* relationship between the intensity of homophily between investors and borrowers and the interest rate investors are willing to accept when funding the corresponding loan. Reported differences in four of the six possible group comparisons are statistically significantly negative. Moreover, the negative differences in interest rates are increasing in the number of homophilous ties observed for the respective

[15] Note that we dichotomize *Borrower job* and hence perform an exact matching on the resulting indicator variables *Self-employed_d, Official_d, Employee_d* and *Other_d*. Naturally, investor gender is also matched exactly.

[16] We estimate robust standard errors and, additionally, correct for a potential large-sample bias which may result from matching on several continuous covariates (c.f., e.g., Abadie and Imbens, 2006).

nearest-neighbor investor-borrower dyads. We observe an economically significant difference of -19 basis points (bp) in loan interest rates when we compare investor-borrower dyads exhibiting the highest versus lowest number of homophilous ties.

These findings are hard to square with the notion that investors predominantly maximize expected risk-adjusted returns of the available loan applications when arriving at their funding decisions. By contrast, sociodemographic similarity seems to make investors accept lower *ex ante* risk-adjusted returns. Taken together, the evidence is consistent with homophily being a behavioral bias rather than an economic rationale in the context of online peer-to-peer lending. Specifically, our results suggest that, in an online setting, too, homophily fosters interpersonal trust formation independent of fundamentals.

4. Discussion and concluding remarks

The question if and to what extent behavioral patterns observed with regards to investors' decision making in offline financial markets spill over to the online sphere is highly relevant to help assess the potential of online solutions to improve market efficiency. In this study, we analyze whether or not homophily—a sociological phenomenon shown to govern agents' decision making on offline financial markets—continues to be an issue when it comes to investors' funding choices on an online peer-to-peer lending platform. Homophily, individuals' propensity to associate with others like them implies that when an investor identifies a given borrower as being similar to her, she tends to see the borrower's behavior in a favorable light, and, all else equal, should display a higher propensity to fund borrowers with whom she shares more commonalities. However, it is an open question if homophily requires in-person interaction in order to evolve. If so, online peer-to-peer lending, on the contrary, might mitigate the impact of homophily on peoples' financial decision-making.

In this study, we test these competing hypotheses and document strong evidence in support of a homophily effect in online peer-to-peer lending. Even after controlling for a host of extant determinants of the likelihood of investment, any additional demographic commonality among a potential investor-borrower dyad increases the investor's odds of funding the corresponding loan application by more than 11% while the average euro amount invested grows by almost 10%. Moreover, we observe a substantial gender gap regarding the effect size of homophily on investors' propensity to invest: regardless of the homophily dimension under review, the impact of investor-borrower sameness on the likelihood of funding a given loan application proves significantly higher for female investors. Finally, we investigate if the observed impact of homophily on investors' funding choices in online peer-to-peer lending reflects economically rational investor decisions or instead points to systematic deviations from economically optimal behavior. In this supplementary analysis, we document a statistically and economically significant difference of -19 basis points (bp) in risk-adjusted interest rates among loans of investor-borrower dyads exhibiting the highest versus lowest number of homophilous ties. This result is inconsistent with the conjecture that investors' propensity to fund similar borrowers in online peer-to-peer lending stems from economic rationale and instead suggests that homophily in peer-to-peer lending is a behavioral bias.

Hence, sociodemographic similarities still govern peoples' financial decision making even when they do not manifest saliently in physical interactions. At this, our findings contribute to research investigating the market efficiency of online peer-to-peer lending which has received increasing attention in recent years. We corroborate prior evidence of significant inefficiencies in peer-to-peer lending and the relevance of investment determinants unrelated to economic fundamentals. However, while prior research limits to single heuristics and biases with respect to investors' decision behavior, we are the first to take a holistic approach towards investigating the impact of homophily on individuals' online investment decisions.

Given that loans granted via peer-to-peer lending have quickly become one of the major sources of alternative financing for private households in recent years, future research into the dynamics and market outcomes on the corresponding platforms seems worthwhile in order to improve our understanding of both investors' and borrowers' online behavior as well as to inform service providers and policy makers about the potential to advance allocation efficiency via peer-to-peer lending.

5. References

Abadie, A., & Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica, 74*(1), 235–267.

Aldrich, H. (1989). Networking among women entrepreneurs. In O. Hagan, C. Rivchun, & D. Sexton (Eds.), Women-owned businesses (pp. 103–132). Praeger.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization.

Baltzer, M., Stolper, O., & Walter, A. (2013). Is local bias a cross-border phenomenon? Evidence from individual investors' international asset allocation. *Journal of Banking & Finance, 37*(8), 2823–2835.

Barasinska, N., Jentzsch, N., & Schäfer, D. (2011). Internet-Kreditplattformen ziehen immer mehr traditionelle Kreditnehmer an. *DIW Wochenbericht*, 78(36), 2–9.

Barasinska, N., & Schäfer, D. (2014). Is Crowdfunding Different? Evidence on the Relation between Gender and Funding Success from a German Peer-to-Peer Lending Platform. *German Economic Review, 15*(4), 436–452.

Baym, N. K., & Ledbetter, A. (2009). Tunes that bind? *Information, Communication & Society*, *12*(3), 408–427.

Bernard, H. R., Killworth, P. D., Evans, M. J., McCarty, C., & Shelley, G. A. (1988). Studying Social Relations Cross-Culturally. *Ethnology, 27*(2), 155–179.

Brashears, M. E. (2008). Gender and homophily: Differences in male and female association in Blau space. *Social Science Research, 37*(2), 400–415.

Burt, R. S. (1982). Toward a structural theory of action: Network models of social structure, perception, and action. Academic Press.

Centola, D., & van de Rijt, A. (2015). Choosing your network: Social preferences in an online health community. *Social Science & Medicine, 125*, 19–31.

Destatis. (2020). Fläche und Bevölkerung nach Ländern. Statistische Ämter Des Bundes Und Der Länder - Gemeinsames Statistikportal. http://www.statistikportal.de/de/bevoelkerung/flaeche-und-bevoelkerung

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking & Finance, 64*, 169–187.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *Review of Financial Studies, 25*(8), 2455–2484.

Friedkin, N. E. (1993). Structural Bases of Interpersonal Influence in Groups: A Longitudinal Case Study. *American Sociological Review, 58*(6), 861–872.

Girard, Y., Hett, F., & Schunk, D. (2015). How individual characteristics shape the structure of social networks. *Journal of Economic Behavior & Organization, 115*, 197–216.

Hegde, D., & Tumlinson, J. (2014). Does Social Proximity Enhance Business Partnerships? Theory and Evidence from Ethnicity's Role in U.S. Venture Capital. *Management Science, 60*(9), 2355–2380.

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell Me a Good Story and I May Lend you Money: The Role of Narratives in Peer-to-Peer Lending Decisions. *Journal of Marketing Research, 48*(SPL), 138–149.

Huberman, G. (2001). Familiarity Breeds Investment. *Review of Financial Studies, 14*(3), 659–680.

Hwang, B.-H., & Kim, S. (2009). It pays to have friends. *Journal of Financial Economics, 93*(1), 138–158.

Ibarra, H. (1992). Homophily and Differential Returns: Sex Differences in Network Structure and Access in an Advertising Firm. *Administrative Science Quarterly, 37*(3), 422–447.

Ivković, Z., & Weisbrenner, S. (2005). Local Does as Local Is: Information Content of the Geography of Individual Investors' Common Stock Investments. *Journal of Finance, 60*(1), 267–306.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening Peers Softly: Inferring the Quality of Small Borrowers. *Management Science, 62*(6), 1554–1577.

Jacobs, H., & Hillert, A. (2016). Alphabetic Bias, Investor Recognition, and Trading Behavior. *Review of Finance, 20*(2), 693–723.

Jaspersen, S., & Limbach, P. K. F. (2018). Knowing Me, Knowing You? Similarity to the CEO and Fund Managers' Investment Decisions.

Kalmijn, M. (1998). Intermarriage and Homogamy: Causes, Patterns, Trends. *Annual Review of Sociology, 24*(1), 395–421.

Lai, S., & Teo, M. (2008). Home-Biased Analysts in Emerging Markets. *Journal of Financial and Quantitative Analysis, 43*(3), 685–716.

Lawrence, B. S. (2006). Organizational Reference Groups: A Missing Perspective on Social Context. *Organization Science, 17*(1), 80–100.

Lawrence, B., & Shah, N. (2020). Homophily: Measures and Meaning. Academy of Management Annals.

Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a Social Process: A Substantive and Methodological Analysis. In M. Berger, T. Abel, & C. H. Page (Eds.), Freedom and Control in Modern Society (pp. 18–66). Van Nostrand.

Levinson, R. M., McCollum, K. T., & Kutner, N. G. (1984). Gender homophily in preferences for physicians. *Sex Roles, 10*(5–6), 315–325.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. *Management Science, 59*(1), 17–35.

Lin, M., & Viswanathan, S. (2016). Home Bias in Online Investments: An Empirical Study of an Online Crowdfunding Market. *Management Science, 62*(5), 1393–1414.

Louch, H. (2000). Personal network integration: transitivity and homophily in strong-tie relations. *Social Networks, 22*(1), 45–64.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*(1), 415–444.

Mills, J., & Clark, M. S. (1982). Communal and exchange relationships. *Review of Personality and Social Psychology, 3*, 121–144.

Pope, D. G., & Sydnor, J. R. (2011). What's in a Picture?: Evidence of Discrimination from Prosper.com. *Journal of Human Resources, 46*(1), 53–92.

Popielarz, P. A. (1999). (IN) VOLUNTARY ASSOCIATION: A Multilevel Analysis of Gender Segregation in Voluntary Organizations. *Gender & Society, 13*(2), 234–250.

Ravina, E. (2008). Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets. SSRN Electronic Journal.

Schufa Holding AG. (2019). Summary for Policymakers. In Intergovernmental Panel on Climate Change (Ed.), Schufa Kredit Kompass 2018 (Vol. 53, Issue 9). Cambridge University Press.

Silver, A. (1990). Friendship in Commercial Society: Eighteenth-Century Social Theory and Modern Sociology. *American Journal of Sociology, 95*(6), 1474–1504.

Sorenson, O., & Stuart, T. E. (2001). Syndication Networks and the Spatial Distribution of Venture Capital Investments. *American Journal of Sociology, 106*(6), 1546–1588.

Stolper, O., & Walter, A. (2019). Birds of a Feather: The Impact of Homophily on the Propensity to Follow Financial Advice. *Review of Financial Studies, 32*(2), 524–563.

Strong, N., & Xu, X. (2003). Understanding the Equity Home Bias: Evidence from Survey Data. *Review of Economics and Statistics, 85*(2), 307–312.

Thelwall, M. (2009). Homophily in MySpace. *Journal of the American Society for Information Science and Technology, 60*(2), 219–231.

Uzzi, B. (1999). Embeddedness in the Making of Financial Capital: How Social Relations and Networks Benefit Firms Seeking Financing. *American Sociological Review, 64*(4), 481–505.

Volkovich, Y., Laniado, D., Kappler, K. E., & Kaltenbrunner, A. (2014). Gender Patterns in a Large Online Social Network (pp. 139–150).

Wang, Z., Walther, J. B., Pingree, S., & Hawkins, R. P. (2008). Health Information, Credibility, Homophily, and Influence via the Internet: Web Sites Versus Discussion Groups. *Health Communication, 23*(4), 358–368.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). MIT Press.

Wright, K. (2000). Perceptions of on-line support providers: An examination of perceived homophily, source credibility, communication and social support within on-line support groups. *Communication Quarterly, 48*(1), 44–59.

Ziegler, T., Shneor, R., Wenzlaff, K., Wang, B. W., Kim, J., Odorovic, A., Paes, F. F. de C., Suresh, K., Zhang, B. Z., Johanson, D., Lopez, C., Mammadova, L., Adams, N., & Luo, D. (2020). The Global Alternative Finance Market Benchmarking Report. https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/alternative-finance/downloads/2020-04-22-ccaf-global-alternative-finance-market-benchmarking-report.pdf

6. Appendix

Appendix III-A: Definitions of variables

Table III-A.1: Definition of variables

Variable	Measurement unit	Explanation
Loan amount	Euro	Amount of money the borrower is applying for.
Loan interest rate	%	Interest rate, which is calculated by the P2P platform.
Borrower credit score	#	Numeric Schufa (Germany's biggest scoring agency) score, ranging from 1 (Risk-Quota of 0.8%) to 15 (Risk-Quota of 96.08%). P2P platform allows maximum Risk-Quota of 41.77% (Schufa Score 10). Original score values are given in letters and transferred to numbers. 1 corresponds to letter A, 10 corresponds to letter K.
Loan duration	Months	Credit period.
Loan financing status	%	Percentage of loan's amount that was financed at the time the loan was taken offline.
Loan miscellaneous	1/0	Variable indicating whether the borrower has specified a particular purpose of the loan as a loan category (0) or not (1).
Loan startup capital	1/0	Variable indicating whether loan category chosen by the borrower corresponds to 'startup capital' category (1) or not (0).
Loan leisure	1/0	Variable indicating whether loan category chosen by the borrower corresponds to 'leisure' category (1) or not (0).
Loan debt consolidation	1/0	Variable indicating whether loan category chosen by the borrower corresponds to 'debt consolidation' category (1) or not (0).
*Loan home improvemen*t	1/0	Variable indicating whether loan category chosen by the borrower corresponds to 'home improvement' category (1) or not (0).
Loan car purchase	1/0	Variable indicating whether loan category chosen by the borrower corresponds to 'car purchase' category (1) or not (0).
Loan higher education	1/0	Variable indicating whether loan category chosen by the borrower corresponds to 'higher education' category (1) or not (0).
Loan other	1/0	Variable indicating whether loan category chosen by the borrower corresponds to any aforementioned specific category (0) or not (1).
Loan description automatic	1/0	Variable indicating whether borrower chose an individual loan description (0) or whether loan description was automatically filled by P2P platform (1).
Loan photo manual	1/0	Variable indicating whether borrower chose an individual photo for the loan application (1) or whether photo was instead automatically chosen by P2P platform (0).
Loan rank	#	Variable indicating the position of a loan application as presented on the P2P platform website. Loan applications are presented in chronological order, with the youngest application ranked first place.
Bid amount	Euro	Amount of money that investor places in one bid. Only one bid per investor per loan possible.
Bids per loan	#	Number of bids that has been submitted to one loan.
Borrower age	Years	Borrower's age on October 15th, 2018.
Number loans per borrower	#	Number of loans that one single borrower applied for during the period from 2007 to 2018. Only one loan application possible at the same time.
Total bid amount per borrower	€	Total amount that that one single borrower applied for during the period from 2007 to 2018
Borrower job employed	1/0	Variable indicating whether a borrower is an employee (1) or not (0).
Borrower job official	1/0	Variable indicating whether a borrower is a public servant (1) or not (0).

Borrower job self-employed	1/0	Variable indicating whether a borrower is self-employed (1) or not (0).
Borrower job other	1/0	Variable indicating whether borrower's job corresponds to any aforementioned specific category (0) or not (1).
Borrower gender	1/0	Variable indicating whether borrower's gender is male (1) or female (0).
Investor age	Years	Investor's age on October 15th, 2018.
Number bids per investor	#	Number of bids that were submitted by one single investor during the period from 2007 to 2018. Only one bid per loan per investor can be submitted.
Total funding per investor	€	Total amount that was invested by one single investor during the period from 2007 to 2018.
Number previous bids	#	Number of bids submitted by one investor before placing the observed bid.
Investor gender	1/0	Variable indicating whether investor's gender is male (1) or female (0).
Same age	1/0	Variable indicating whether borrower and investor are in the same age group (+/- 5 years difference between investor and borrower) (1 if same age group).
Same state	1/0	Variable indicating whether borrower and investor are living in the same federal state (1 if same state).
Same gender	1/0	Variable indicating whether borrower and investor have the same gender (1 if same gender).
Number of homophilous ties	#	Sum of binary specification of *Same age*, *Same state* and *Same gender*.
Dyad_match_d	1/0	Variable indicating whether a loan – which was available at the time a bid was submitted – was matched (1) by this certain bid or not (0).
Dyad_match_€	Euro	Euro amount of money, which an investor places in a bid pertaining to a loan application of a borrower.
Simultaneously available loans	#	Number of dyads that are available at the time the bid is submitted.

Appendix III-B: Detailed information on investors and borrowers

Table III-B.1: Gender proportions of investors and borrowers

This table shows the distribution of borrowers and investors by gender.

	Borrower	Investor
Male	8,911 (76.13%)	6,901 (93.43%)
Female	2,794 (23.87%)	485 (6.566%)
All	11,705 (100.0%)	7,386 (100.0%)

Table III-B.2: Distribution of borrowers and investors by place of residence

This table presents the distribution of sampled investors and borrowers by place of residence. For comparison, the rightmost columns present the distribution of the German population by place of residence (Source: Destatis)

	Borrowers		Investors		German population	
	N	%	N	%	N	%
Baden-Württemberg	1,486	12.69	1,085	14.69	11,069,533	13.47
Bavaria	1,822	15.67	1,348	18.25	13,076,721	15.91
Berlin	832	7.108	471	6.377	3,644,826	4.435
Brandenburg	448	3.827	175	2.369	2,511,917	3.057
Bremen	96	0.8202	68	0.9207	682,986	0.8311
Hamburg	327	2.794	249	3.371	1,841,179	2.241
Hesse	927	7.920	657	8.895	6,265,809	7.625
Lower Saxony	1,063	9.082	589	7.975	7,982,448	9.714
Mecklenburg-Vorpommern	207	1.768	90	1.219	1,609,675	1.959
North Rhine-Westphalia	2,313	19.76	1,503	20.35	17,932,651	21.82
Rhineland-Palatinate	525	4.485	332	4.595	4,084,844	4.971
Saarland	99	0.8458	57	0.7717	990,509	1.205
Saxony	577	4.930	281	3.804	4,077,937	4.962
Saxony-Anhalt	277	2.367	103	1.395	2,208,321	2.687
Schleswig-Holstein	402	3.434	209	2.830	2,896,712	3.525
Thuringia	299	2.554	129	1.747	2,143,145	2.608
N/A	5	0.04272	40	0.5416	-	-
Total	11,705	100.0	7,386	100.0	82,175,684	100.0

IV. Occupational self-selection among bankers and financial regulators: evidence from content analysis

Co-authors: Peter Tillmann, Andreas Walter
Own share: 70%

Occupational self-selection among bankers and financial regulators: evidence from content analysis

Daniel Czaja[a] Peter Tillmann[b] Andreas Walter[c]

Abstract - Using a unique dataset of 532 face-to-face interviews from a German business newspaper, we examine the psychological characteristics of bankers and financial regulators. Applying computerized content analysis techniques, we find that linguistic styles differ significantly among bankers and regulators even when controlling for topics derived from structural topic modeling suggesting differences in well-known psychological characteristics. In particular, we find that bankers' linguistic style marks them as more selfish and overconfident. Regulators, in contrast, show more linguistic markers of cognitive complexity. Results support existing theoretical considerations that suggest self-selection into different occupations by bankers and regulators due to different psychological characteristics.

Keywords: Psychological characteristics; Occupation; Self-selection; Personality; Content analysis; Topic Modelling

JEL-Codes: G21, G28, J45

[a] Institute of Financial Services, University of Giessen, Daniel.L.Czaja@wirtschaft.uni-giessen.de.

[b] Department of Economics, University of Giessen, Peter.Tillmann@wirtschaft.uni-giessen.de.

[c] Institute of Financial Services, University of Giessen, Andreas.Walter@wirtschaft.uni-giessen.de.

1. Introduction

"Language is the dress of thought."

(Johnson, 1779)

Although being highly regulated (Bond and Glode, 2014), the financial services industry is littered by problematic financial practices culminating in scandals and fraudulent practices that cause economic damages as well as a lack of stability and reputation of the financial system (e.g., Abrantes-Metz et al., 2012; Barth et al., 2012; Cohn et al., 2014). Those detrimental practices are often attributed to the business culture in the financial industry and personality features of the key actors (i.e. bankers) within the industry, respectively (e.g., Carney, 2014; Clarke, 1988; Cohn et al., 2014; Lagarde, 2014). On the other hand, acting as a counterpart to the actors within the financial industry, regulators have to prevent detrimental practices by assuring that financial institutions comply with laws (Shive and Forster, 2017). Put differently, the objectives of regulators are aligned with those of society. Accordingly, regulators are characterized as altruistic individuals that are motivated by a broader mission, i.e. promotion of justice (Besley and Ghatak, 2005). Following the argumentation of Besley and Ghatak (2005), people motivated by a higher mission and altruism, respectively, self-select into occupations that typically offer lower compensation while selfish people strive for monetary incentives (Gill et al., 2015; Kosfeld and von Siemens, 2011). Since, normally, regulators earn less than bankers (Henderson and Tung, 2012; Philippon and Reshef, 2012), this rationale suggests differences in psychological characteristics among bankers and financial regulators.

Based on aforementioned considerations regarding differences between bankers and financial regulators, Bond and Glode (2014) propose a labor market model in which individuals choose to work as regulators or bankers due to heterogeneous psychological characteristics. This model builds on a basic theory stating that individuals' seek occupations that fit with their values and traits (Judge and Bretz, 1992; Lyons et al., 2006).

However, so far, there is no empirical investigation of the occupational self-selection in the context of the labor market for bankers and regulators. To this end, our study addresses this research gap and provides evidence for the hypothesis of occupational self-selection among bankers and regulators by investigating whether financial regulators differ from bankers with respect to the features of bankers identified in the previous literature.

Specifically, these characteristics comprise selfishness (Gill et al., 2015), cognitive complexity (Bond and Glode, 2014), dishonesty (Cohn et al., 2014) and overconfidence (Kaustia and Perttula, 2012; Suntheim, 2016). For this purpose, as research on personality usually relies on self-reported questionnaires, which is not applicable to analyze professional (business) leaders, we make use of bankers' and regulators' language use, respectively. As exemplified by the initial quote from the famous English writer Samuel Johnson, spoken words transmit more than just facts. Known as self-disclosure, speakers, even if unintended or suppressed, "make themselves the subject of their message" (Pearce and Sharp, 1973, p. 409). To investigate differences in bankers' and regulators' psychological characteristics, we take advantage of the systematic relation between language use and psychological characteristics (Gottschalk et al., 1997; Kim, 2013; Lee et al., 2007; Pennebaker and Stone, 2003; Tausczik and Pennebaker, 2010). The key assumption behind our analysis is that psychological characteristics expressed in language are deeply rooted in a person's personality and remain stable even if circumstances change. Hence, personality traits remain unchanged over the career, the evolution of the financial system, changes in regulation etc..

For our research, we draw on face-to-face interview data from the *Börsen-Zeitung*, a leading German daily business newspaper. Our final data covers 532 interviews (637,187 words in total) with representatives from companies in the financial sector[1], on the one hand, and representatives from the financial regulatory authority and central banks[2], on the other hand. It covers a 15-year period from 2003 to 2017. As language can vary according to the situations in which their users find themselves (Benson and Greaves, 1973), the uniform interview situation in our dataset allows us to examine bankers' and regulators' language use without a systematic bias resulting from context specifics. In particular, we start the analysis of the written contents using the *Linguistic Inquiry and Word Count* (*LIWC2015*) (Pennebaker et al., 2015a). We use LIWC2015, as a large body of literature from different fields has provided substantial evidence and validity for the relationship between psychological characteristics and language use

[1] This group comprises two industries, namely the banking services industry and the investment banking and investment services industry. From here, we refer to this group as *bankers*.

[2] From here, we refer to this group as *regulators.*

applying LIWC word lists (see Tausczik and Pennebaker (2010) for a comprehensive overview). Based on word counts derived from LIWC2015, we follow the literature in constructing summary variables for psychological characteristics. More specifically, we derive measures of psychological characteristics that have received the most attention in the previous literature on bankers' traits, namely selfishness, cognitive complexity, dishonesty and overconfidence. Using a linear regression framework with time- and interviewer-fixed effects, we investigate differences in those summary variables between bankers and regulators.

Results suggest that regulators are less selfish and less overconfident than bankers. Additionally, regulators’ show higher cognitive complexity. As an interview’s topics might have an impact on the manifestation of linguistic markers of psychological characteristics (Mehl et al., 2012), we repeat our analyses by additionally controlling for the topic proportions of the respective interview that we derived from structural topic modelling (Roberts et al., 2014). However, results qualitatively remain the same. Therefore, the analysis of features of bankers and regulators reveals that they differ significantly with regard to psychological characteristics and thus, our findings support the hypothesis of self-selection into different occupations among bankers and regulators.

Our research sheds new light on bankers’ and regulators’ psychological characteristics. Contrary to what is often claimed (e.g., Bond and Glode, 2014), we cannot confirm that regulators are not as smart (cognitive complexity) as bankers. Additionally, our results suggest that regulators actually are less prone to selfishness as well as biased self-attribution and overconfidence, respectively. On the one hand, it can be assumed that, based on virtually the same educational background as regulators, bankers self-select into occupations with higher monetary incentives due to their specific psychological characteristics. Those characteristics, which include detrimental features such as selfishness and overconfidence, trigger problematic financial practices and a lack of trust in the financial industry (Bond and Glode, 2014; Gill et al., 2015). On the other hand, regulators’ self-selection can be explained by an underlying broader mission and a social purpose and thus should lead per se to benefits for society at large (Besley and Ghatak, 2005). Policy-makers should take this self-selection of bankers and regulators into account when debating about regulation, especially against the background of a well-documented relationship between psychological characteristics and job performance (e.g., Barrick et al., 2002; Cohn et al., 2014; O'Boyle et al., 2011). In this context, a discussion on a further restriction of the so-called revolving door between bankers and regulators (Shive and Forster, 2017), i.e. the migration of employees between regulatory authorities and private companies, is indispensable.

2. Data

2.1 Sample selection

For our research, we use face-to-face interview data from the *Börsen-Zeitung*, a leading German daily business newspaper. The *Börsen-Zeitung* focuses on financial markets, corporate reports and the banking industry and is mainly read by members of top management teams (Börsen-Zeitung, 2019).

We collected interview data from Lexis-Nexis and the website of the *Börsen-Zeitung* for the period of 2003 to 2017. Therefore, we searched for all articles tagged as an interview and reviewed them manually to keep actual interviews only. Besides textual data, we obtained data on the interviewers and interviewees, i.e. their name and gender, as well as further metadata, such as the publication date and the company or institution an interviewee represents. Based on this information, we screened our initial sample for interviews with representatives from companies in the financial sector[3], on the one hand, and representatives from the financial regulatory authority and central banks, on the other hand. This data collection and cleaning procedure resulted in a sample of 794 interviews. Next, we made two adjustments. First, we excluded all interviews with non-German-speaking interviewees to rule out biases resulting from translations[4]. Second, taking the revolving door effect (Shive and Forster, 2017) into account, we exclude all interviews with interviewees being classified as both, banker and regulator in our sample. Our final data covers 532 interviews, which we separated into interviewers' and interviewees'

[3] We used the industry classification provided by Thomson Reuters Eikon to identify companies from the financial services industry.

[4] Therefore, we identified the locations of companies' headquarters by using Thomson Reuters Eikon database and kept only those with a headquarter in Germany, Austria or Switzerland. For public institutions, classification was conducted manually.

speech parts.[5] In total, we analyze 44,446 sentences or 637,188 words spoken by the interviewees.

Additionally, to control for interviewees' heterogeneity we obtain further data on interviewee characteristics, i.e. their age and their position in the corresponding company or institution. Therefore, we reviewed the info boxes about the interviewees that are typically provided by the *Börsen-Zeitung* in addition to the actual interviews. Those info boxes have a length of 150-200 words on average and provide further information on interviewees. Additionally, we obtained data from company and institution websites if some information on an interviewee was lacking.

The uniform interview situation in our dataset allows us to examine bankers' and regulators' language use without a context-specific systematic bias that can be found when speakers find themselves in different situations (Benson and Greaves, 1973).

2.2 Summary statistics

Table IV-1 reports the summary statistics. Overall, our final sample covers 532 interviews, with 474 bankers accounting for a total of 556,562 words (39,259 sentences) and 58 regulators accounting for a total of 80,626 words (5,187 sentences). We present the descriptive statistics separated according to group membership. In particular, columns 1 to 5 report the descriptive statistics for bankers whereas columns 6 to 10 provide information on regulators.

[5] Please note that we have also performed our analyses with the larger sample of 783 interviews (only excluding interviews with interviewees being classified as both, banker and regulator in our sample). Qualitatively, however, the results remain the same.

Table IV-1: Summary statistics

This table presents the descriptive statistics of our sample. *Age of interviewee* is the interviewee's age at the time of the interview in years. *Sex of interviewee* indicates interviewee's gender. *Sex of interviewer* indicates interviewer's gender. *Leading Position* is an indicator variable equaling one if the interviewee is in a leading position and zero otherwise. *Member board of supervisors* is an indicator variable equaling one if the interviewee is a member of the board of supervisors in the corresponding company and zero otherwise. *CEO* is an indicator variable equaling one if the interviewee is the CEO of the board of supervisors in the corresponding company and zero otherwise. *Board member (other)* is an indicator variable Indicator variable equaling one if the interviewee is a member of the executive board (not CEO) in the corresponding company and zero otherwise. *Division manager* is an indicator variable equaling one if the interviewee is a division manager in the corresponding company and zero otherwise. *Operational employee* is an indicator variable equaling one if the interviewee is neither a member of the executive board nor a member of the board of supervisors or a division manager in the corresponding company and zero otherwise. *Member of executive board* is an indicator variable equaling one if the interviewee is a member of the executive board in the corresponding institution and zero otherwise. *Member of presiding board* is an indicator variable equaling one if the interviewee is a member of the presiding board in the corresponding institution and zero otherwise. *Other position* is an indicator variable equaling one if the interviewee is neither a member of the executive board nor a member of the presiding board in the corresponding institution and zero otherwise. *Financial regulation* is an indicator variable equaling one if the interviewee works for a financial regulation authority and zero otherwise. *Central bank* is an indicator variable equaling one if the interviewee works for a central bank and zero otherwise. *Wordcount* is the total count of words an interviewee uses in an interview. *Interviewee's speech share* is the proportion of interviewee's total word count in relation to the length of the interview. Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables.

	Bankers					**Regulators**				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	N	Mean (SD)	Min	p50	Max	N	Mean (SD)	Min	p50	Max
A. Demographics										
Age of interviewee [years]	436	51.83 (± 7.74)	28	51	85	57	55.37 (± 6.56)	42	57	68
Sex of interviewee [0/1]	474	0.03 (± 0.18)	0	0	1	58	0.21 (± 0.41)	0	0	1
Sex of interviewer [0/1]	474	0.21 (± 0.41)	0	0	1	58	0.14 (± 0.35)	0	0	1
Leading Position [0/1]	474	0.88 (± 0.32)	0	1	1	58	0.90 (± 0.31)	0	1	1
Member board of supervisors [0/1]	474	0.03 (± 0.17)	0	0	1					
CEO [0/1]	474	0.61 (± 0.49)	0	1	1					
Board member (other) [0/1	474	0.24 (± 0.43)	0	0	1					
Division manager [0/1	474	0.08 (± 0.28)	0	0	1					
Operational employee [0/1]	474	0.00 (± 0.06)	0	0	1					
Member of executive board [0/1]						58	0.47 (± 0.50)	0	0	1
Member of presiding board [0/1]						58	0.43 (± 0.50)	0	0	1
Other position [0/1]						58	0.10 (± 0.31)	0	1	1
Financial regulation [0/1]						58	0.34 (± 0.48)	0	0	1
Central bank [0/1]						58	0.66 (± 0.48)	0	1	1
B. Interviews										
Wordcount [#]	474	1174.18 (± 622.71)	254	1,052	2,431	58	1390.10 (± 621.57)	341	1,549	2,387
Interviewee's speech share [%]	474	80.35 (± 5.49)	57.13	80.63	93.49	58	77.99 (± 6.89)	58.71	78.22	91.68

On average, regulators are about 55 years old and thus slightly older than bankers (about 52 years) in our sample. Furthermore, the share of women is evidently higher among regulators (21%) than among bankers (3%) supporting the finding of a greater gender inequality in the private sector (Andreeva and Bertaud, 2013; Collischon, 2019). As can be also inferred from the descriptive statistics, the share of female interviewers is higher for bankers; however, the difference is not statistically significant. Additionally, the share of interviewees in a leading position is slightly higher for regulators, but also not statistically significant. More specifically, with regard to interviewee's position, statistics show that about 85% (*CEO* and *Board member (other)*) of considered bankers are board members whereas about 90% of considered regulators are members of an executive or presiding board. Statistics regarding interviews reveal that, on average, regulators use a higher number of words (1,390) than bankers do (1,174), however, the speech share of regulators (78%) is lower than for bankers (80%). Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables used in this study.

3. Hypotheses and key variables

3.1 Linguistic manifestations of occupational classes

Occupations are an important feature of societies as they are a common feature of social organization. In general, individuals self-select into different occupations due to their psychological characteristics and social preferences (e.g., Besley and Ghatak, 2005; M. Lagarde and Blaauw, 2014; Serra et al., 2011). For example, Kosfeld and von Siemens (2011) investigate a competitive labor market with heterogeneous workers. In the provided setting, selfish workers self-select into occupations with high monetary incentives. On the other hand, altruistic individuals that are motivated by a broader mission self-select into occupations offering moderate monetary incentives (Besley and Ghatak, 2005). Demiralp (2011) identifies individuals' effort decisions as another determinant of occupational self-selection. In addition, there is some evidence for occupational self-selection in the non-profit sector. Serra et al. (2011) find that philanthropic health professionals have an increased likelihood to be employed in the non-profit sector earning lower wages. Carpenter and Myers (2010) find a positive correlation between individuals' altruism and their decision to volunteer in a fire department. Furnham et al. (2014) investigate differences between the public and the private sector with regard to dysfunctional traits. Private sector employees, especially those in the financial industry, exhibit higher overconfidence than public sector employees. Furthermore, Carpenter and Gong (2013) provide evidence that some individuals are attracted by their organization's mission.

As outlined above, occupational self-selection bases on individuals' psychological characteristics. The term *psychological characteristics* includes both, psychological traits (personality) and psychological states. In our understanding, following personality models, such as the six-factor model (Ashton et al., 2004, 2000) this umbrella term also includes virtues, such as honesty or its counterpart, dishonesty. Usually, research on personality relies on self-reported questionnaires to analyze psychological characteristics. As this is not an option for business leaders and high-level officials, employing a language-based method, as conducted in our research, provides new opportunities in this domain.

The use of language provides comprehensive information regarding speakers' psychological characteristics. By using a characteristic language, individuals reveal their individual personality. Psychological research finds a systematic and validated relation between language use and different aspects of psychological characteristics (Gottschalk et al., 1997; Kim, 2013; Lee et al., 2007; Pennebaker and Stone, 2003; Tausczik and Pennebaker, 2010).

In the vein of previous literature, we analyze bankers' and regulators' psychological characteristics using the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015b) software representing the most established software to conduct a computer-aided content analysis. Using LIWC word lists that classify words according to their (psychological) meaning, a large body of literature from different fields provides substantial evidence for the relationship between psychological characteristics and language use (see Tausczik and Pennebaker (2010) for a comprehensive overview). In detail, LIWC implements a word count procedure searching for almost 6,400 pre-classified words within a given text (Pennebaker et al., 2015b). In total, LIWC provides 73 word categories, i.e. linguistic dimensions, words indicating psychological constructs, words regarding personal concerns and informal language markers. In the search for pre-classified words within any given text, LIWC counts the number of words for each of these categories and returns the shares of words – relative to total words – as output.

Based on these percentages and according to previous studies, we construct linguistic measures for selfishness, cognitive complexity, dishonesty and overconfidence by combining percentages of different word categories. We describe those literature-based computation procedures to derive each particular measure in detail in section 3.2 and provide summary statistics for those measures in Table IV-B.1 in Appendix IV-B.

Based on the above findings from studies in personality research that find a significant link between specific word usage and personality traits, researchers have investigated psychological characteristics of, among others, politicians (e.g., Pennebaker et al., 2005; Slatcher et al., 2007), students (Newman et al., 2003), users of instant messengers (Hancock et al., 2007) and inpatients (Creswell et al., 2007; Pennebaker and King, 1999). In the business context, psychological characteristics of CEOs (e.g., Green et al., 2019; Scheuerlein and Chládková, 2019) and entrepreneurs (e.g., Block et al., 2019; Obschonka et al., 2017; Obschonka and Fisch, 2018) have been researched using LIWC.

3.2 Measuring speaker's psychological characteristics

3.2.1 Selfishness

Selfishness is defined as the excessive concern for oneself or for one's own personal profit or pleasure (Dubois et al., 2015) and thus, represents the opposite of altruism. In economics and finance, literature has emphasized the importance of selfishness and self-interest, respectively. For instance, Weitzel et al. (2010) show that entrepreneurially talented people are more selfish. Haynes et al. (2015) find that managers' self-interest leads to a short-term orientation with regard to decision-making behavior and performance evaluation. Furthermore, results from other studies demonstrate that economic students are more selfish than other students (Bauman and Rose, 2011; Frey and Meier, 2003; Rubinstein, 2006). Bauman and Rose (2011) and Frey and Meier (2003) provide evidence that the selfishness of economists can be explained by a selection effect. Among economic students, students preferring to work in the financial industry are most selfish (Gill et al., 2015). According to Gill et al. (2015), selfish individuals are largely attracted by the high monetary incentives in the financial industry.

In contrast to this, altruistic individuals motivated by a broader mission self-select into occupations with lower monetary incentives (Besley and Ghatak, 2005; Kosfeld and von

Siemens, 2011). Pro-social values, thus, as people's intrinsic motivation factor are crucial for labor supply (Lagarde and Blaauw, 2014).

Overall, different social preferences of individuals influence the match between individuals and their occupation. In this context, individuals' occupational self-selection is mainly driven by the varying monetary incentives in different occupations (Gill et al., 2015; Kosfeld and von Siemens, 2011). Based on differences between the public and the private sector and as monetary incentives are evidently lower for regulators than for bankers (Henderson and Tung, 2012; Philippon and Reshef, 2012), we hypothesize that bankers are more selfish than regulators.

To measure selfishness, we make use of the relationship between pronoun use and speaker's self-focus. For example, Chopik et al. (2014) show the relevance of pronouns for measuring self-focus versus other-focus. Methodologically, following Czaja and Röder (2020) and Kim (2013), we assess interviewees' excessive self-referencing by summing the shares of first person singular personal pronouns (*I*) and first person plural personal pronouns (*we*) and subtracting the share of third person personal pronouns (*other*): $SelfRef = I + we - other$. Relative frequencies pertain to the interviewees' total word count.

Additionally, we investigate speakers' narcissism. Narcissism can be defined as excessive selfishness (Kernberg, 1998; Naidu et al., 2019) and thus, represents a more extreme feature. Moreover, selfishness is an integral part of narcissists' psychological characteristics (Campbell et al., 2005). Contrary to what many studies assumed in the past, Holtzman et al. (2019) showed that narcissism is unrelated to using first-person singular pronouns. Instead, the authors find that narcissists use higher levels of achievement words (*achiev*), inhibition words (*inhib*), optimism and energy words (*optim*), second person personal pronouns (*youtotal*), references to sports (*sports*) and death (*death*) as well as lower levels of third person personal pronouns (*other*), anxiety words (*anx*), negative emotion words (*negemo*), insight words (*insight*), tentative words (*tent*), perception words (*perc*), feeling words (*feel*) and references to home (*home*) and music (*music*). Therefore, in accordance with the psycholinguistic literature (Newman et al., 2003; Slatcher et al., 2007) we compute our linguistic measure of narcissism by summing and subtracting the z-scores (across speakers)[6] for the aforementioned categories: $Narc = \underline{z}achiev + \underline{z}inhib + \underline{z}optim + \underline{z}youtotal + \underline{z}sports + \underline{z}death - \underline{z}other - \underline{z}anx - \underline{z}negemo - \underline{z}insight - \underline{z}tent - \underline{z}perc - \underline{z}feel - \underline{z}home - \underline{z}music.$

3.2.2 Cognitive complexity

Cognitive complexity constitutes a psychological characteristic and is defined as individuals' abilities to process information affecting them (Suedfeld, 2009). High cognitive complexity indicates that an individual carefully evaluates all relevant information on a topic and then integrates it into a coherent position. Low complexity, on the other hand, indicates that an individual's powers of comprehension are limited. Cognitive complexity positively correlates with intelligence (Hansell et al., 2015). Moreover, it is positively associated with academic performance (Zhang et al., 2012).

Proposing a labor market model for bankers and regulators, Bond and Glode (2014) argue that bankers and regulators have heterogeneous abilities. More specifically, the authors postulate that regulators are not as smart as bankers. To this end, based on the proposed labor market model and on the positive correlation between cognitive complexity and intelligence, we hypothesize that bankers show higher cognitive complexity than regulators.

We use three different measures for cognitive complexity. First, following Slatcher et al. (2007) we construct *CogScore*. Previous research finds that cognitive complexity is associated with the use of more exclusive words (*excl*), tentative words (*tent*), negations (*negate*), discrepancies (*disc*) and less inclusive words (*incl*). Thus, we derive our linguistic measure of cognitive complexity by summing and subtracting the z-scores for the aforementioned categories: $CogScore = \underline{z}excl + \underline{z}tent + \underline{z}negate + \underline{z}disc - \underline{z}incl.$

Second, as complexity measures are seen as linguistic markers of cognitive complexity (Pennebaker and Stone, 2003), we use the Flesch-Reading-Ease score (*FRE*) (Flesch, 1948) as an additional proxy for cognitive complexity. The Flesch-Reading-Ease score is an established readability measure that is used extensively in the business context (e.g.,

[6] We convert the LIWC output, i.e. percentages of total words used in an analyzed text, for each relevant LIWC category to z scores across speakers according to the standard formula as follows: $z_i = \frac{x_i - mean(x)}{std.dev(x)}$. This means that data is normalized in relation to all data (and not speaker specific).

Clatworthy and Jones, 2001; Courtis, 2004). For German texts, the score is defined as follows (Groeben, 1982): $FRE = 180 - \frac{total\ words}{total\ sentences} - \left(58.5 * \frac{total\ syllabes}{total\ words}\right)$. It is important to note that the interpretation of this measure is contrary to the interpretation of the other cognitive complexity measures. A higher *FRE* value corresponds to a lower level of cognitive complexity.

Third, cognitive complexity is also associated with a higher level of cognitive process words (*CogPro*) (Pennebaker and Stone, 2003), i.e. words from the following word categories: insight (e.g., “think”, “know”), causation (e.g., “because”, “effect”), discrepancy (e.g., “should”, ”would”), tentative (e.g., “maybe”, “perhaps”), certainty (e.g., “always”, “never”) and differentiation (e.g., “but”, “else”). Additionally, the use of cognitive process words correlates with intellectual achievements (Klein and Boals, 2001). Therefore, we apply the share of cognitive process words as a third measure of cognitive complexity.

3.2.3 Dishonesty

By definition, telling untruths requires the description of non-existent events and attitudes (Newman et al., 2003). The lack of honesty leads to imagined experiences being told. According to linguistic literature, individuals describe imaginary events differently to real events, which is reflected in individuals’ language use (Newman et al., 2003; Slatcher et al., 2007; Vrij et al., 2004, 2000).

Previously published studies on dishonesty in the public and private sector have been inconsistent and contradictory (Barfort et al., 2019; Hanna and Wang, 2017; Posner and Schmidt, 1982). While Barfort et al. (2019) and Posner and Schmidt (1982) document that dishonest individuals are more likely to self-select into private sector jobs in occidental countries, Hanna and Wang (2017) provide contradictory findings for India. Thus, in this context, cultural and country differences seem to have an influence on occupational self-selection. Moreover, Barfort et al. (2019) argue that higher monetary incentives in the private sector would attract more dishonest individuals.

Although unethical business culture in the financial industry constitutes a politically relevant issue (e.g., Carney, 2014; Clarke, 1988; Cohn et al., 2014; Lagarde, 2014), much of the research up to now has been descriptive in nature. As an exception, Cohn et al. (2014) provide empirical evidence that bankers are more dishonest than both, employees from other industries and students. Based on this finding, we hypothesize that bankers are more dishonest than regulators.

Following literature (Newman et al., 2003; Slatcher et al., 2007), we include five word categories to derive our dishonesty measure. Newman et al. (2003) find that dishonest individuals use higher levels of negative emotion words (*negemo*) and motion words (*motion*). Additionally, they use lower levels of first-person singular pronouns (*I*), references to others (*other*) and exclusive words (*excl*). Therefore, following Newman et al. (2003) and Slatcher et al. (2007) our measure of dishonesty (*DScore*) is defined as follows: $DScore = \underline{z}negemo + \underline{z}motion - \underline{z}I - \underline{z}other - \underline{z}excl$.

3.2.4 Overconfidence

Overconfidence describes the tendency of individuals to overestimate their own abilities. In general, the literature suggests that overconfidence significantly influences people's behavior (McCannon et al., 2016). For example, overconfidence affects trading behavior (Barber and Odean, 2001, 2000; Chen et al., 2007; Glaser and Weber, 2007; Goetzmann and Kumar, 2008; Merkle, 2013; Odean, 1998) and managers' behavior in the context of mergers and acquisitions (Billett and Qian, 2008; Doukas and Petmezas, 2007) and forecasting (Libby and Rennekamp, 2012). In addition to the extensive literature on CEO overconfidence in general, there is also research on the overconfidence of finance professionals (Kaustia and Perttula, 2012; Suntheim, 2016). Suntheim (2016) provides empirical evidence that banks managed by an overconfident CEO are subject to higher risks. Kaustia and Perttula (2012) document overconfidence in terms of better-than-average thinking and unfounded confidence among financial advisors and bank branch managers.

A small literature investigates the problem of overconfidence among central bankers. Claussen et al. (2012) present a model of monetary policy decisions and show that overconfidence of policymakers can help to replicate several properties of the decision making process in central banks. Bennani (2020) introduces an indicator of overconfidence of the chair of the Federal Reserve. His evidence suggests that an overconfident chairman contributes to higher investor sentiment.[7] However, the

[7] The recent superstar status of central bankers („Super Mario" Draghi, „Maestro" Alan Greenspan) suggests that the public perception of regulators or central bankers can be consistent with overconfidence.

investigated period is characterized by a varying degree of overconfidence among examined chairmen with positive and negative values, centered around zero.

Given the empirical evidence of overconfidence among finance professionals combined with the scarce findings regarding financial regulators, we hypothesize that bankers show a higher level of overconfidence than regulators.

Methodologically, we follow Kim (2013) who provides evidence that overconfidence is reflected in communication patterns of CEOs. Based on theoretical foundations (Gervais and Odean, 2001) postulating that the self-attribution bias triggers overconfidence, the author argues that their language-oriented measure of the self-attribution bias approximates CEOs' overconfidence. In detail, Kim (2013) puts CEOs' self-referencing, on the one hand, and referencing to others, on the other hand, into relation. According to the definition of the self-attribution bias, i.e. individuals tendency to credit oneself and one's own abilities with past success but to blame others or external factors for failures, CEOs are identified as biased if they show abnormal self-referencing after positive events and abnormal referencing to others after negative events. Following this rationale, we derive our first measure of overconfidence as follows:

$$SAB = \begin{cases} |SelfRef|, if \ (Tone > 0 \ \wedge SelfRef > 0) \vee (Tone < 0 \ \wedge SelfRef < 0) \\ 0, \qquad otherwise \end{cases} \quad (1)$$

First, we assess interviewees' excessive self-referencing (*SelfRef*) as described in section 3.2.1. Next, we determine an interview's overall tone (*Tone*) by subtracting the share of negative connoted words from the share of positive connoted words as classified by the business-specific word lists of Bannier et al. (2019). Accordingly, we define an interview as positive if *Tone* is positive and as negative if *Tone* is negative. Then, we construct the *SAB* variable. *SAB* equals *SelfRef* in an overall positive interview and positive *SelfRef*. Analogously, *SAB* equals the absolute value of *SelfRef* in an overall negative interview and negative *SelfRef*. Otherwise, *SAB* is zero.

Additionally, for robustness reasons, we use the summary variable *Clout* provided by the LIWC software as another overconfidence measure that is used in different research areas (Duncan et al., 2019; Oliver et al., 2020; Smith-Keiling and Hyun, 2019). A higher *Clout* score indicates that the speaker is confident (Pennebaker et al., 2015b) and thus, reflects a weaker form of overconfidence.

3.3 Univariate evidence

To gain first insights into the differences between bankers and regulators regarding psychological characteristics, we begin our analysis with a discussion of descriptive evidence reported in Table IV-2. Statistical significance is tested by Student's t-test.

Table IV-2: Differences in linguistic measures between bankers and regulators

This table presents the values of used linguistic measures for both, bankers and regulators. In addition, column six shows the differences in those linguistic measures between bankers and regulators including the associated statistical significance (Student’s t-test). Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables. In addition, we standardized our linguistic measures across the entire sample by subtracting the mean and dividing by the standard deviation. The standardized values for each respective measure does not add up to zero due to unequal sample sizes of bankers and regulators. Standardized values are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	VARIABLES	Bankers	Regulators	N	Diff.	t-statistic
Selfishness	*SelfRef*	2.978	1.119	532	1.859	6.518***
		(0.095)	(-0.778)		(0.873)	
	Narc	0.151	-1.233	532	1.384	2.445**
		(0.037)	(-0.302)		(0.339)	
Cognitive complexity	*CogScore*	-0.127	1.042	532	-1.169	-2.516**
		(-0.038)	(0.310)		(-0.348)	
	FRE	54.376	49.394	532	4.982	5.460***
		(0.081)	(-0.659)		(0.740)	
	CogPro	14.864	17.079	532	-2.215	-7.079***
		(-0.103)	(0.839)		(-0.942)	
Dishonesty	*DScore*	-0.004	0.034	532	-0.038	-0.114
		(-0.002)	(0.014)		(-0.016)	
Overconfidence	*SAB*	2.077	0.463	532	1.614	5.520***
		(0.082)	(-0.666)		(0.747)	
	Clout	73.099	61.171	532	11.927	7.776***
		(0.112)	(-0.914)		(1.026)	

We document that, unconditionally, bankers show significantly higher selfishness than regulators. More specifically, bankers have a 1.859 (1.384) higher *SelfRef* (*Narc*) score, which corresponds to 0.873 (0.339) standard deviations of the variable. Regarding cognitive complexity, all three measures indicates a higher value for regulators. In particular, differences between regulators and bankers amount to 1.169, 4.982 and 2.215, respectively, corresponding to 0.348, 0.740 and 0.942 standard deviations of the respective variable. The two measures of overconfidence also show statistically significant differences. We document a 0.747 (1.026) standard deviations higher *SAB* (*Clout*) score for bankers compared to regulators. For dishonesty, we cannot prove a statistically significant effect.

In sum, the results presented in Table IV-2 provide preliminary evidence for bankers being more selfish and overconfident and regulators showing higher cognitive complexity.

4. Regression results

4.1 Model

To examine the effect of the membership in a group, i.e. being regulator or banker, on linguistic markers of psychological characteristics while controlling for interviewee and interview characteristics as well as interviewer and year fixed effects, we estimate the following model

$$LinguisticScore_{i,j,k,t} = \alpha_{i,j} + \beta Regulator_j + \delta' m_j + \varepsilon_{i,j,k,t} \quad (2)$$

where $LinguisticScore_{i,j,k,t}$ denotes the corresponding linguistic variable i for interviewee j in an interview conducted by interviewer k in year t. $Regulator_j$ denotes an indicator variable equaling one if the interviewee is a regulator and zero otherwise. Additionally, we draw on prior research on psycholinguistics to identify the relevant control variables included in the vector m_j. In particular, we control for information about interviewee j including her age (Pennebaker and Stone, 2003) and gender (Newman et al., 2008) as well as an indication whether she is in a leading position (Leonardi and Rodriguez-Lluesma, 2013). Moreover, we control for interviewee's speech share and the total count of words spoken by an interviewee. Furthermore, we add year fixed effects to address potential time effects as well as interviewer fixed effects to control for a potential interviewer effect (Davis and Silver, 2003; Davis et al., 2010). We estimate the model using standard errors clustered by interviewee and year.

4.2 Main results

Table IV-3 reports results obtained from various specifications of the linear regression model formalized in equation (2).[8] The conducted multiple regressions can explain up to 36.3% (*FRE)* of variation in our dependent variable. The regression results generally

[8] For interpretation purposes, we standardized our linguistic measures across the entire sample by subtracting the mean and dividing by the standard deviation. We go into these standardized values in more detail in the discussion section. However, we report non-standardized results in the main regression results tables. See Table IV-C.1 and Table IV-C.2 in Appendix IV-C for standardized results.

support univariate evidence presented in section 3.3.[9] In detail, regulators show a 1.949 lower self-referencing score, which is about 70% of the mean score indicating considerable differences in terms of selfishness. The estimation results for *Narc*, which can be interpreted as excessive selfishness (Kernberg, 1998; Naidu et al., 2019) show the same tendency. However, the coefficient for the regulator variable is not statistically significant. Results for our cognitive complexity proxies reveal that regulators show higher cognitive complexity. In particular, they use language that is more complex and includes more linguistic markers indicating intellectual achievements. Compared with mean values, regulators speaks 11% more complex (*FRE*) and use 13% more words indicating cognitive processing (*CogPro*). Again, we find no evidence for differences in dishonesty between bankers and regulators. Estimation results for our overconfidence proxies fully support univariate evidence. Accordingly, regulators' *SAB* score is 1.721 lower than bankers' *SAB* score, just as the *Clout* score that is 12.260 lower for regulators.

Results for included control variables reveal that only interviewee's age partly affects our investigated linguistic markers. For example, results suggest that interviewee's age positively relates to manifested cognitive complexity. Furthermore, our results partly support previous literature showing a decline in confidence among people over 70 (Orth et al., 2018). In contrast, neither interviewees' sex nor their position seems to affect the linguistic markers.

[9] Since age data for 39 interviewees is lacking, only 493 observations are included in our regressions.

Table IV-3: Regression results

This table presents our main regression results. We estimate the following model $LinguisticScore_{i,j,k,t} = \alpha_{i,j} + \beta Regulator_j + \delta' m_j + \varepsilon_{i,j,k,t}$ using ordinary least square regression analysis, where $LinguisticScore_{i,j,k,t}$ denotes the corresponding linguistic variable *i* for interviewee *j* in an interview conducted by interviewer *k* in year *t*. We add year fixed effects to address potential time effects as well as interviewer fixed effects to control for a potential interviewer effect. Our main explanatory variable is $Regulator_j$ that denotes an indicator variable equaling one if the interviewee is a regulator and zero otherwise. Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables. Double-clustered standard errors for interviewee and year are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Selfishness		**Cognitive complexity**			**Dishonesty**	**Overconfidence**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	*SelfRef*	*Narc*	*CogScore*	*FRE*	*CogPro*	*DScore*	*SAB*	*Clout*
Regulator	-1.949***	-1.160	1.201*	-5.959***	1.986***	-1.134***	-1.721***	-12.260***
	(-5.371)	(-1.308)	(1.827)	(-5.487)	(4.599)	(-2.674)	(-5.381)	(-6.724)
Leading Position	0.404	-1.072	0.067	-1.357	0.152	-0.659	-0.318	-1.402
	(1.262)	(-1.261)	(0.115)	(-1.270)	(0.404)	(-1.459)	(-0.949)	(-0.802)
Sex of interviewee	0.439	0.892	-0.849	-0.046	-0.229	-0.229	0.313	2.680
	(0.955)	(0.848)	(-1.063)	(-0.035)	(-0.411)	(-0.428)	(0.764)	(1.375)
Interviewee's speech share	-0.052***	0.039	-0.004	-0.102	-0.027	-0.023	-0.059***	-0.158
	(-2.710)	(0.939)	(-0.120)	(-1.536)	(-1.293)	(-1.085)	(-2.915)	(-1.488)
Log(Wordcount)	0.041	-0.422	1.288***	3.097***	0.929***	0.768***	0.312	-1.258
	(0.214)	(-0.905)	(3.952)	(4.526)	(4.557)	(3.351)	(1.491)	(-1.066)
Age group = 1, 50-64	0.141	-0.392	0.067	-1.007	0.097	0.235	-0.115	-1.392
	(0.656)	(-0.847)	(0.211)	(-1.607)	(0.458)	(0.994)	(-0.497)	(-1.296)
Age group = 2, >64	-0.751*	0.007	1.880***	0.189	1.244***	-0.361	-0.691	-11.878***
	(-1.748)	(0.009)	(2.782)	(0.124)	(2.991)	(-0.742)	(-1.445)	(-4.448)
Constant	4.149**	-1.234	-7.503***	41.423***	10.699***	-4.806**	3.516*	82.689***
	(2.288)	(-0.303)	(-2.669)	(7.029)	(5.534)	(-2.318)	(1.789)	(8.161)
Observations	493	493	493	493	493	493	493	493
Adjusted R^2	0.309	0.054	0.249	0.363	0.348	0.221	0.230	0.316
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Interviewer FE	YES	YES	YES	YES	YES	YES	YES	YES

4.3 Impact of interview's topic

Conceptual arguments suggest that the relationship between psychological characteristics and their linguistic markers might be context dependent (Mehl et al., 2012). According to Gibson (1979), each context takes on a psychological function for a given trait. This function defines the extent to which a trait can be revealed. Empirically, inconsistent findings regarding the relationships between word use and personality traits support those claims. For example, Pennebaker and King (1999) find that women use more first person personal pronouns in personal essays whereas Mulac and Lundell (1994) find lower use of first person personal pronouns in picture descriptions among women. Additionally, Mehl et al. (2012) explicitly investigated the contextual dependence of the relationship between word use and traits. Comparing stream-of-consciousness essays and daily conversations with other, the authors find that psychological correlates of word use can be context dependent. Therefore, we repeat our main regression analyses by additionally controlling for the topic proportions of the respective interview that we derived from structural topic modelling (STM) (Roberts et al., 2014). STM is an unsupervised machine learning algorithm that extends the well-known latent Dirichlet allocation (LDA) model (Blei et al., 2003). The LDA model represents a generative probabilistic model that can be used to identify underlying topics in a set of documents based on the correlations of words within the documents. The basic assumption of the LDA model is that words within a document are independently drawn from different bags of words that contain a set of words each. Each bag of words represents a topic, i.e. a mixture over words where each word has a probability to belong to a topic and the sum of word probabilities for each topic is one. In other words, a document is a mixture of different topics where the shares of all topics in a document add up to one. Every document has an unique mixture of topic proportions.[10] The STM algorithm additionally allows to take document metadata into account, i.e. both, the topic proportions within each document (topical prevalence) and the word distributions within each topic (topical content) can be a function of document metadata.[11]

[10] Please see Blei et al. (2003) for more details on the LDA model.

[11] Please see Roberts et al. (2014) for more details on the STM algorithm.

Applying the STM algorithm to our interview data, we include the group membership of an interviewee, i.e. banker or regulator as well as the year of an interview as topical prevalence covariates. Put differently, when deriving interview topics we incorporate the fact that bankers and regulators have different thematic focuses and that these focuses change over time. Moreover, in addition to single words we also included bigrams and trigrams in the topic generating process, i.e. coherent combinations of two or three words. Stop words without a deeper meaning were excluded.[12]

After estimating topic proportions for each interview, we create 30 variables capturing topic proportions for each interview and include those variables (except one to avoid perfect multicollinearity) in our baseline model described in section 4.1. Table IV-4 reports the regression results.

Table IV-4: Regression results with inclusion of interviews' topic proportions

This table presents a rerun of our main regression extended by the 29 variables indicating the proportions of the 30 identified topic within an interview as derived from structural topic modelling (we exclude one variable to avoid perfect multicollinearity). Our main explanatory variable is $Regulator_j$ that denotes an indicator variable equaling one if the interviewee is a regulator and zero otherwise. Control variables are the same as presented in Table IV-3. Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables. Double-clustered standard errors for interviewee and year are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Selfishness		**Cognitive complexity**			**Dishonesty**	**Overconfidence**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	*SelfRef*	*Narc*	*CogScore*	*FRE*	*CogPro*	*DScore*	*SAB*	*Clout*
Regulator	-1.024*	-4.170***	1.645**	-4.955***	1.381**	0.172	-0.894**	-6.353**
	(-1.669)	(-3.238)	(2.019)	(-2.714)	(2.479)	(0.313)	(-2.143)	(-2.084)
Observations	493	493	493	493	493	493	493	493
Adjusted R^2	0.403	0.191	0.303	0.399	0.438	0.224	0.318	0.429
Further Controls	YES	YES	YES	YES	YES	YES	YES	YES
Topic Proportions	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Interviewer FE	YES	YES	YES	YES	YES	YES	YES	YES

As can be seen from Table IV-4, incorporating the topic proportions allows us to significantly increase the explanatory power of our regression models explaining up to 43.8% (*CogPro)* of variation in our dependent variables. Estimated coefficients, however,

[12] Based on the model diagnostics reported in Figure IV-D.1 and Figure IV-D.2 in Appendix IV-D, we set the number of topics to 30. The ten most prevalent derived topics are presented as word clouds in Figure IV-D.3 in Appendix IV-D, while the distribution of top ten topics (by prevalence) by interviewee group is illustrated in Figure IV-D.4.

remain qualitatively the same. In addition, we find a statistically significant effect according the narcissism proxy variable amounting to -4.170, which corresponds to 1.020 standard deviations, indicating excessive selfishness among bankers compared to regulators.

5. Discussion and concluding remarks

Analyses of bankers' and regulators' language use in a uniform interview setting indicate that their linguistic styles differ significantly suggesting different psychological characteristics among them. Even after controlling for a battery of extant determinants of the characteristics of linguistic markers as well as interviews' topic proportions, bankers display a higher selfishness amounting to 0.481 (1.020) as measured by the standardized score for *SelfRef* (*Narc*). Thus, we can confirm our first hypothesis. Analogously, we document bankers being more prone to overconfidence by reporting a 0.414 and 0.546 higher standardized score for *SAB* and *Clout*, respectively. Therefore, again, we can confirm our hypothesis. On the other hand, regulators are associated with a higher cognitive complexity, the value of which is 0.490, 0.736 and 0.587 higher than that of bankers, depending on the standardized measure applied. This finding contradicts our hypothesis of smarter bankers. In fact, our finding support early exploratory research by Posner and Schmidt (1982) as well as empirical evidence provided by Crewson (1995) documenting that public administrators are more capable than business administrators in terms of arithmetic reasoning, mathematical knowledge and verbal expression. In the same vein, Lyons et al. (2006) show that public administrators are doing more intellectually stimulating work, i.e. performing ability challenging projects and tasks. On the contrary, especially in the corporate context, emotional intelligence, which we cannot measure here, plays a special role (Côté et al., 2010; Sadri, 2012). With regard to dishonesty, we can neither reject nor confirm our hypothesis. Although Cohn et al. (2014) document strong evidence of dishonesty in the banking industry, measuring dishonesty by means of linguistic measures show no significant differences between bankers and regulators. Our results, thus, support the main result of Rahwan et al. (2019) who find no evidence for dishonesty in the banking industry by replicating the analysis of Cohn et al. (2014) with participants from different populations, i.e. bankers from different-sized banks from the Asia Pacific region, Middle East and Europe. It follows that bankers' psychological characteristics might vary across jurisdictions.

The empirical strategy used in this study to assess group differences between bankers and regulators is itself novel, enabling us to systematically investigate actors' psychological characteristics. The advantage of this empirical strategy arises mainly from the fact that classical personality tests are not an option in this context. Nevertheless, our procedure is subject to possible limitations.

First, unfortunately, our final sample of regulators consists of only 58 interviews. One way to address this issue would be to include speeches, debates and other interview formats in the investigation. However, this would lead to a possible speech source effect as documented for example in Slatcher et al. (2007), i.e. differences in linguistic measures due to varying communication partners, networks and locations.

Second, ideally, we would have analyzed personality data of economics and business administrations students who subsequently decided to work either in the financial industry or as financial regulators. This approach would allow us to isolate the relationship between psychological characteristics and occupational self-selection. Unfortunately, however, this respective data is lacking. Nevertheless, in the absence of such data, our approach allows us to gain new insights into the psychological characteristics of bankers and regulators, which would not be possible without this approach.

Based on the strong evidence for the relationship between psychological characteristics and job performance including problematic practices (e.g., Barrick et al., 2002; Cohn et al., 2014; O'Boyle et al., 2011) as well as against the background that financial regulators frequently make subjective decisions about considered banks (Rosen, 2003), it might be expedient to alter the incentives to work in the considered occupations. Here, it must be policy makers' objective to reduce fraudulent practices that cause economic damages as well as a lack of stability and reputation of the financial system (e.g., Abrantes-Metz et al., 2012; Barth et al., 2012; Cohn et al., 2014). At the same time, a discussion on a further restriction of the revolving door between regulation and industry is indispensable.[13] However, to make informed decisions and especially against the background that our

[13] According to Lucca et al. (2014), the evidence on the revolving door is consistent with the view that the flow of workers between regulators and banks predominantly contributes to an exchange of information, not a „quid-pro-quo".

investigation can only provide a glimpse into the psychological characteristics of bankers and regulators, future studies on the current topic are required.

6. References

Abrantes-Metz, R. M., Kraten, M., Metz, A. D., & Seow, G. S. (2012). Libor manipulation? *Journal of Banking & Finance*, *36*(1), 136–150.

Andreeva, M., & Bertaud, N. (2013). Women and men in leadership positions in the European Union, 2013. In *Publications Office European Commission*.

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A Six-Factor Structure of Personality-Descriptive Adjectives: Solutions From Psycholexical Studies in Seven Languages. *Journal of Personality and Social Psychology*, *86*(2), 356–366.

Ashton, M. C., Lee, K., & Son, C. (2000). Honesty as the sixth factor of personality: correlations with machiavellianism, primary psychopathy, and social adroitness. *European Journal of Personality*, *14*(4), 359–368.

Bannier, C., Pauls, T., & Walter, A. (2019). Content analysis of business communication : introducing a German dictionary. *Journal of Business Economics*, *89*(1), 79–123.

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, *55*(2), 773–806.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261–292.

Barfort, S., Harmon, N. A., Hjorth, F., & Olsen, A. L. (2019). Sustaining Honesty in Public Service: The Role of Selection. *American Economic Journal: Economic Policy*, *11*(4), 96–123.

Barrick, M. R., Stewart, G. L., & Piotrowski, M. (2002). Personality and job performance: Test of the mediating effects of motivation among sales representatives. *Journal of Applied Psychology*, *87*(1), 43–51.

Barth, J. R., Levine, R., & Jr., G. C. (2012). *Guardians of finance: Making regulators work for us*. MIT Press.

Bauman, Y., & Rose, E. (2011). Selection or indoctrination: Why do economics students donate less than the rest? *Journal of Economic Behavior & Organization*, *79*(3), 318–327.

Bennani, H. (2020). Central Bank Communication in the Media and Investor Sentiment. *Journal of Economic Behavior & Organization (forthcoming)*.

Benson, J. D., & Greaves, W. S. (1973). *The language people really use*. Ontario Book Society of Canada.

Besley, T., & Ghatak, M. (2005). Competition and Incentives with Motivated Agents. *American Economic Review*, *95*(3), 616–636.

Billett, M. T., & Qian, Y. (2008). Are overconfident CEOs born or made? Evidence of self-attribution bias from frequent acquirers. *Management Science*, *54*(6), 1037–1051.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Block, J. H., Fisch, C. O., Obschonka, M., & Sandner, P. G. (2019). A personality perspective on business angel syndication. *Journal of Banking and Finance*, *100*, 306–327.

Bond, P., & Glode, V. (2014). The Labor Market for Bankers and Regulators. *Review of Financial Studies*, *27*(9), 2539–2579.

Börsen-Zeitung. (2019). *Strukturanalyse der Leserschaft*. https://www.boersen-zeitung.de/nutzung/Leserstrukturanalyse.pdf

Campbell, W. K., Bush, C. P., Brunell, A. B., & Shelton, J. (2005). Understanding the Social Costs of Narcissism: The Case of the Tragedy of the Commons. *Personality and Social Psychology Bulletin*, *31*(10), 1358–1368.

Carney, M. (2014). Inclusive Capitalism: Creating a Sense of the Systemic. In *Conference on Inclusive Capitalism*. BIS central bankers' speeches. https://www.bis.org/review/r140528b.pdf

Carpenter, J., & Gong, E. (2013). *Motivating Agents : How Much Does the Mission Matter?* (Institute for the Study of Labor - Discussion Paper Series).

Carpenter, J., & Myers, C. K. (2010). Why volunteer? Evidence on the role of altruism, image, and incentives. *Journal of Public Economics*, *94*(11–12), 911–920.

Chen, G., Kim, K. A., Nofsinger, J. R., & Rui, O. M. (2007). Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. *Journal of Behavioral Decision Making*, *20*(4), 425–451.

Chopik, W. J., Joshi, D. H., & Konrath, S. H. (2014). Historical changes in American self-interest: State of the Union addresses 1790 to 2012. *Personality and Individual Differences*, *66*, 128–133.

Clarke, R. L. (1988). *Bank Failure: An Evaluation of the Factors Contributing to the Failure of National Banks*. Office of the Comptroller of the Currency.

Clatworthy, M., & Jones, M. J. (2001). The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal*, *14*(3), 311–326.

Claussen, C. A., Matsen, E., Røisland, Ø., & Torvik, R. (2012). Overconfidence, monetary policy committees and chairman dominance. *Journal of Economic Behavior & Organization*, *81*(2), 699–711.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, *516*(7529), 86–89.

Collischon, M. (2019). Is There a Glass Ceiling over Germany? *German Economic Review*, *20*(4), e329–e359.

Côté, S., Lopes, P. N., Salovey, P., & Miners, C. T. H. (2010). Emotional intelligence and leadership emergence in small groups. *The Leadership Quarterly*, *21*(3), 496–508.

Courtis, J. K. (2004). Corporate report obfuscation: artefact or phenomenon? *The British Accounting Review*, *36*(3), 291–312.

Creswell, J. D., Lam, S., Stanton, A. L., Taylor, S. E., Bower, J. E., & Sherman, D. K. (2007). Does Self-Affirmation, Cognitive Processing, or Discovery of Meaning Explain Cancer-Related Health Benefits of Expressive Writing? *Personality and Social Psychology Bulletin*, *33*(2), 238–250.

Crewson, P. E. (1995). A Comparative Analysis of Public and Private Sector Entrant Quality. *American Journal of Political Science*, *39*(3), 628–639.

Czaja, D., & Röder, F. (2020). Self-attribution bias and overconfidence among nonprofessional traders. *The Quarterly Review of Economics and Finance (forthcoming)*.

Davis, D. W., & Silver, B. D. (2003). Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge. *American Journal of Political Science*, *47*(1), 33–45.

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., & Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research*, *25*(1), 14–26.

Demiralp, B. (2011). Occupational self-selection in a labor market with moral hazard. *European Economic Review*, *55*(4), 497–519.

Doukas, J. A., & Petmezas, D. (2007). Acquisitions, overconfident managers and self-attribution bias. *European Financial Management*, *13*(3), 531–577.

Dubois, D., Rucker, D. D., & Galinsky, A. D. (2015). Social class, power, and selfishness: When and why upper and lower class individuals behave unethically. *Journal of Personality and Social Psychology*, *108*(3), 436–449.

Duncan, S. Y., Chohan, R., & Ferreira, J. J. (2019). What makes the difference? Employee social media brand engagement. *Journal of Business & Industrial Marketing*, *34*(7), 1459–1467.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Frey, B. S., & Meier, S. (2003). Are Political Economists Selfish and Indoctrinated? Evidence from a Natural Experiment. *Economic Inquiry*, *41*(3), 448–462.

Furnham, A., Hyde, G., & Trickey, G. (2014). Do your Dark Side Traits Fit? Dysfunctional Personalities in Different Work Sectors. *Applied Psychology*, *63*(4), 589–606.

Gervais, S., & Odean, T. (2001). Learning to be overconfident. *The Review of Financial Studies*, *14*(1), 1–27.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Gill, A., Heinz, M., & Schumacher, H. (2015). *Social Preferences and Selection into the Financial Industry*.

Glaser, M., & Weber, M. (2007). Overconfidence and trading volume. *GENEVA Risk and Insurance Review*, *32*(1), 1–36.

Goetzmann, W. N., & Kumar, A. (2008). Equity portfolio diversification. *Review of Finance*, *12*(3), 433–463.

Gottschalk, L. A., Stein, M. K., & Shapiro, D. H. (1997). The application of computerized content analysis of speech to the diagnostic process in a psychiatric outpatient clinic. *Journal of Clinical Psychology*, *53*(5), 427–441.

Green, T. C., Jame, R., & Lock, B. (2019). Executive Extraversion: Career and Firm Outcomes. *The Accounting Review*, *94*(3), 177-204.

Groeben, N. (1982). *Lesepsychologie: Textverständnis-Textverständlichkeit*. Aschendorff.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, *45*(1), 1–23.

Hanna, R., & Wang, S.-Y. (2017). Dishonesty and Selection into Public Service: Evidence from India. *American Economic Journal: Economic Policy*, *9*(3), 262–290.

Hansell, N. K., Halford, G. S., Andrews, G., Shum, D. H. K., Harris, S. E., Davies, G., Franic, S., Christoforou, A., Zietsch, B., Painter, J., Medland, S. E., Ehli, E. A., Davies, G. E., Steen, V. M., Lundervold, A. J., Reinvang, I., Montgomery, G. W., Espeseth, T., Hulshoff Pol, H. E., … Wright, M. J. (2015). Genetic Basis of a Cognitive Complexity Metric. *PLOS ONE*, *10*(4), e0123886.

Haynes, K. T., Josefy, M., & Hitt, M. A. (2015). Tipping Point: Managers' Self-Interest, Greed, and Altruism. *Journal of Leadership and Organizational Studies*, *22*(3), 265–279.

Henderson, B. M. T., & Tung, F. (2012). *Paying Bank Examiners for Performance: Should Regulators Receive Bonuses for Effectively Guarding the Public Interest? Spring*, 32–38.

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic Markers of Grandiose Narcissism: A LIWC Analysis of 15 Samples. *Journal of Language and Social Psychology*, *38*(5–6), 773–786.

Johnson, S. (1779). *Lives of the Most Eminent English Poets*.

Judge, T. A., & Bretz, R. D. (1992). Effects of work values on job choice decisions. *Journal of Applied Psychology*, *77*(3), 261–271.

Kaustia, M., & Perttula, M. (2012). Overconfidence and debiasing in the financial industry. *Review of Behavioural Finance*, *4*(1), 46–62.

Kernberg, O. F. (1998). aranoid social developments as a consequence of ideological and bureaucratic regression. In *Even Paranoids Have Enemies: New Perspectives on Paranoia and Persecution* (pp. 87–99). Routledge.

Kim, Y. H. (Andy). (2013). Self attribution bias of the CEO: Evidence from CEO interviews on CNBC. *Journal of Banking and Finance*, *37*(7), 2472–2489.

Klein, K., & Boals, A. (2001). Expressive writing can increase working memory capacity. *Journal of Experimental Psychology: General*, *130*(3), 520–533.

Kosfeld, M., & von Siemens, F. A. (2011). Competition, cooperation, and corporate culture. *The RAND Journal of Economics*, *42*(1), 23–43.

Lagarde, C. (2014). *Economic Inclusion and Financial Integrity—an Address to the Conference on Inclusive Capitalism*. https://www.imf.org/en/News/Articles/2015/09/28/04/53/sp052714

Lagarde, M., & Blaauw, D. (2014). Pro-social preferences and self-selection into jobs: Evidence from South African nurses. *Journal of Economic Behavior & Organization*, *107*(PA), 136–152.

Lee, C. H., Kim, K., Seo, Y. S., & Chung, C. K. (2007). The Relations Between Personality and Language Use. *The Journal of General Psychology*, *134*(4), 405–413.

Leonardi, P. M., & Rodriguez-Lluesma, C. (2013). Occupational Stereotypes, Perceived Status Differences, and Intercultural Communication in Global Organizations. *Communication Monographs*, *80*(4), 478–502.

Libby, R., & Rennekamp, K. (2012). Self-serving attribution bias, overconfidence, and the issuance of management forecasts. *Journal of Accounting Research*, *50*(1), 197–231.

Lucca, D., Seru, A., & Trebbi, F. (2014). The revolving door and worker flows in banking regulation. *Journal of Monetary Economics*, *65*, 17–32.

Lyons, S. T., Duxbury, L. E., & Higgins, C. A. (2006). A Comparison of the Values and Commitment of Private Sector, Public Sector, and Parapublic Sector Employees. *Public Administration Review*, *66*(4), 605–618.

McCannon, B. C., Asaad, C. T., & Wilson, M. (2016). Financial competence, overconfidence, and trusting investments: Results from an experiment. *Journal of Economics and Finance*, *40*(3), 590–606.

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2012). How Taking a Word for a Word Can Be Problematic: Context-Dependent Linguistic Markers of Extraversion and Neuroticism. *Journal of Methods and Measurement in the Social Sciences*, *3*(2), 30–50.

Merkle, C. (2013). Financial Overconfidence Over Time - Foresight, Hindsight, and Insight of Investors. *Journal of Banking and Finance*, *84*(August), 1–46.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

Mulac, A., & Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. *Language & Communication*, *14*(3), 299–309.

Naidu, E. S., Patock-Peckham, J. A., Ruof, A., Bauman, D. C., Banovich, P., Frohe, T., & Leeman, R. F. (2019). Narcissism and devaluing others: An exploration of impaired control over drinking as a mediating mechanism of alcohol-related problems. *Personality and Individual Differences*, *139*(August 2018), 39–45.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, *45*(3), 211–236.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675.

O'Boyle, E. H., Humphrey, R. H., Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior*, *32*(5), 788–818.

Obschonka, M., & Fisch, C. (2018). Entrepreneurial personalities in political leadership. *Small Business Economics*, *50*, 851–869.

Obschonka, M., Fisch, C., & Boyd, R. (2017). Using digital footprints in entrepreneurship research: A Twitter-based personality analysis of superstar entrepreneurs and managers. *Journal of Business Venturing Insights*, *8*, 13–23.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *Journal of Finance*, *53*(6), 1887–1934.

Oliver, K. M., Houchins, J. K., Moore, R. L., & Wang, C. (2020). Informing Makerspace Outcomes Through a Linguistic Analysis of Written and Video-Recorded Project Assessments. *International Journal of Science and Mathematics Education (forthcoming)*.

Orth, U., Erol, R. Y., & Luciano, E. C. (2018). Development of self-esteem from age 4 to 94 years: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *144*(10), 1045–1080.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development And Psychometric Properties of LIWC2015*.

Pennebaker, J. W., & King, L. A. (1999). Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.

Pennebaker, J. W., Slatcher, R. B., & Chung, C. K. (2005). Linguistic Markers of Psychological State through Media Interviews: John Kerry and John Edwards in 2004, Al Gore in 2000. *Analyses of Social Issues and Public Policy*, *5*(1), 197–204.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*(2), 291–301.

Philippon, T., & Reshef, A. (2012). Wages and Human Capital in the U.S. Finance Industry: 1909–2006. *The Quarterly Journal of Economics*, *127*(4), 1551–1609.

Posner, B. Z., & Schmidt, W. H. (1982). Determining managerial strategies in the public sector: What kind of people enter the public and private sectors? an updated comparison of perceptions, stereotypes, and values. *Human Resource Management*, *21*(2–3), 35–43.

Rahwan, Z., Yoeli, E., & Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. *Nature*, *575*(7782), 345–349.

Roberts, M. E., Stewart, B. M., Dustin, P., & Harvard, T. (2014). stm: R Package for Structural Topic Models. *Journal of Statistical Software*, *10*(2), 1–40.

Rosen, R. J. (2003). Is Three a Crowd? Competition Among Regulators in Banking. *Journal of Money, Credit, and Banking*, *35*(6a), 967–998.

Rubinstein, A. (2006). A Sceptic's Comment on the Study of Economics. *The Economic Journal*, *116*(510), C1–C9.

Sadri, G. (2012). Emotional Intelligence and Leadership Development. *Public Personnel Management*, *41*(3), 535–548.

Scheuerlein, J., & Chládková, H. (2019). Leadership and Effectiveness: A Content Analysis of Letters to Shareholders During the Financial Crisis. *International Journal of Business Communication*, 1–21.

Serra, D., Serneels, P., & Barr, A. (2011). Intrinsic motivations and the non-profit health sector: Evidence from Ethiopia. *Personality and Individual Differences*, *51*(3), 309–314.

Shive, S. A., & Forster, M. M. (2017). The Revolving Door for Financial Regulators. *Review of Finance*, *21*(4), 1445–1484.

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality*, *41*(1), 63–75.

Smith-Keiling, B. L., & Hyun, H. I. F. (2019). Applying a Computer-Assisted Tool for Semantic Analysis of Writing: Uses for STEM and ELL. *Journal of Microbiology & Biology Education*, *20*(1), 1–6.

Suedfeld, P. (2009). Integrative complexity. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 354–366). The Guilford Press.

Suntheim, F. (2016). CEO Overconfidence in Banking. In *SSRN Electronic Journal*.

Taddy, M. A. (2012). On Estimation and Selection for Topic Models. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR 22*, 1184–1193.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Detecting Deceit Via Analyses of Verbal and Nonverbal Behavior in Children and Adults. *Human Communication Research*, *30*(1), 8–41.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Costs of secondary parasitism in the facultative hyperparasitoid Pachycrepoideus dubius: Does host size matter? *Journal of Nonverbal Behavior*, *24*(4), 239–248.

Weitzel, U., Urbig, D., Desai, S., Sanders, M., & Acs, Z. (2010). The good, the bad, and the talented: Entrepreneurial talent and selfish behavior. *Journal of Economic Behavior & Organization*, *76*(1), 64–81.

Zhang, M., Xin, Z., & Lin, C. (2012). Measures of Cognitive Complexity and Its Development in Chinese Adolescents. *Journal of Constructivist Psychology*, *25*(2), 91–111.

7. Appendix

Appendix IV-A: Definitions of variables

Table IV-A.1: Definition of variables

Variable [Measurement unit]	Explanation
A. *General Variables*	
Regulator [1/0]	Indicator variable equaling one if the interviewee is a regulator and zero otherwise.
Age of interviewee [Years]	Age of interviewee at the time of the interview.
Sex of interviewee [1/0]	Sex of interviewee.
Sex of interviewer [1/0]	Sex of interviewer.
Leading Position [1/0]	Indicator variable equaling one if the interviewee is in a leading position and zero otherwise. For bankers, we identify an interviewee being in a leading position if she is a member of the executive board or a member of the board of supervisors in the corresponding company. For regulators, we identify an interviewee being in a leading position if she is president or vice-president or member of the executive board in the corresponding institution.
B. *Demographics Bankers*	
Member board of supervisors [1/0]	Indicator variable equaling one if the interviewee is a member of the board of supervisors in the corresponding company and zero otherwise.
CEO [1/0]	Indicator variable equaling one if the interviewee is the CEO of the board of supervisors in the corresponding company and zero otherwise.
Board member (other) [1/0]	Indicator variable equaling one if the interviewee is a member of the executive board (not CEO) in the corresponding company and zero otherwise.
Division manager [1/0]	Indicator variable equaling one if the interviewee is a division manager in the corresponding company and zero otherwise.
Operational employee [1/0]	Indicator variable equaling one if the interviewee is neither a member of the executive board nor a member of the board of supervisors or a division manager in the corresponding company and zero otherwise.
C. *Demographics Regulators*	
Member of executive board [1/0]	Indicator variable equaling one if the interviewee is a member of the executive board in the corresponding institution and zero otherwise.
Member of presiding board [1/0]	Indicator variable equaling one if the interviewee is a member of the presiding board in the corresponding institution and zero otherwise.
Other position [1/0]	Indicator variable equaling one if the interviewee is neither a member of the executive board nor a member of the presiding board in the corresponding institution and zero otherwise.
Financial regulation [1/0]	Indicator variable equaling one if the interviewee works for a financial regulation authority and zero otherwise.
Central bank [1/0]	Indicator variable equaling one if the interviewee works for a central bank and zero otherwise.
D. *Textual variables*	
Wordcount [#]	Total count of words an interviewee uses in an interview.
Interviewee's speech share [%]	The proportion of interviewee's total word count in relation to the length of the interview.
SelfRef [%]	Sum of the share of first person singular personal pronouns (I) (relative to interviewee's total word count) and first person plural personal pronouns (*we*) minus the share of third person personal pronouns (*other*): *SelfRef = I + we – other*.
Narc [-]	Composition of the following LIWC categories converted to z scores (across speakers): achievement words (*achiev*), inhibition words (*inhib*), optimism and energy words (*optim*), 2nd person personal pronouns (*youtotal*), references to sports (*sports*), death (*death*), third person personal pronouns (*other*), anxiety words (*anx*), negative emotions words (*negemo*), insight words (*insight*), tentative words (*tent*), perception words (*perc*), feeling words (*feel*) and references to home (*home*) and music (*music*). The score is calculated as follows: *Narc = zachiev + zinhib + zoptim + zyoutotal + zsports + zdeath – zother – zanx – znegemo – zinsight – ztent – zperc – zfeel – zhome – zmusic.*

CogScore [-]	Composition of the following LIWC categories converted to z scores (across speakers): exclusive words (excl), tentative words (tent), negations (negate), discrepancies (disc) and inclusive words (incl). The score is calculated as follows: $CogScore = \underline{z}excl + \underline{z}tent + \underline{z}negate + \underline{z}disc - \underline{z}incl$
FRE [-]	Flesch-Reading-Ease score defined as $FRE = 180 - \frac{total\ words}{total\ sentences} - \left(58.5 * \frac{total\ syllabes}{total\ words}\right)$.
CogPro [%]	Total count of words from the LIWC category *cognitive processes* relative to interviewee's total word count.
DScore [-]	Composition of the following LIWC categories converted to z scores (across speakers): negative emotion words (*negemo*), motion words (*motion*), first-person singular pronouns (*i*), references to others (*other*) and exclusive words (*excl*). The score is calculated as follows: $DScore = \underline{z}negemo + \underline{z}motion - \underline{z}I - \underline{z}other - \underline{z}excl$
SAB [-]	Score that equals *SelfRef* in an overall positive interview (*Tone*>0) and that equals the absolute value of *SelfRef* (if negative) in an overall negative interview (*Tone*<0). Otherwise, the score is zero. $SAB = \begin{cases} \lvert SelfRef \rvert, if\ (Tone > 0\ \wedge SelfRef > 0) \vee (Tone < 0\ \wedge SelfRef < 0) \\ 0, \quad otherwise \end{cases}$
Clout [-]	Summary variable provided by the LIWC software indicating interviewee's confidence on a scale from 0 to 100.

Appendix IV-B: Summary statistics for the used linguistic measures

Table IV-B.1: Summary statistics for the used linguistic measures

This table presents the summary statistics for the used linguistic measures in our study. Please see section 3.2 and Table IV-A.1 in Appendix IV-A for detailed description of all variables.

VARIABLES	N	Mean	SD	Min	p25	Median	p75	Max
SelfRef	532	2.776	2.129	-3.070	1.210	2.705	4.355	9.030
Narc	532	0.000	4.087	-12.766	-2.656	-0.029	2.468	16.796
CogScore	532	-0.000	3.358	-12.435	-2.266	0.257	2.241	9.657
FRE	532	53.833	6.734	22.785	50.309	54.588	58.236	69.085
CogPro	532	15.105	2.351	7.640	13.475	15.195	16.785	22.120
DScore	532	-0.000	2.378	-8.101	-1.481	-0.046	1.471	7.336
SAB	532	1.901	2.160	0.000	0.000	1.035	3.490	9.030
Clout	532	71.798	11.628	29.740	63.705	71.960	80.685	95.240

Appendix IV-C: Regression results with standardized dependent variables

Table IV-C.1: Regression results with standardized dependent variables

This table presents our main regression results. We estimate the following model $LinguisticScore_{i,j,k,t} = \alpha_{i,j} + \beta Regulator_j + \delta' m_j + \varepsilon_{i,j,k,t}$ using ordinary least square regression analysis, where $LinguisticScore_{i,j,k,t}$ denotes the corresponding linguistic variable i for interviewee j in an interview conducted by interviewer k in year t. We add year fixed effects to address potential time effects as well as interviewer fixed effects to control for a potential interviewer effect. Our main explanatory variable is $Regulator_j$ that denotes an indicator variable equaling one if the interviewee is a regulator and zero otherwise. Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables. Double-clustered standard errors for interviewee and year are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Selfishness		Cognitive complexity			Dishonesty	Overconfidence	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	*SelfRef*	*Narc*	*CogScore*	*FRE*	*CogPro*	*DScore*	*SAB*	*Clout*
Regulator	-0.916***	-0.284	0.358*	-0.885***	0.844***	-0.106	-0.797***	-1.054***
	(-5.371)	(-1.308)	(1.827)	(-5.487)	(4.599)	(-2.674)	(-5.381)	(-6.724)
Leading Position	0.190	-0.262	0.020	-0.201	0.065	-0.269	-0.147	-0.121
	(1.262)	(-1.261)	(0.115)	(-1.270)	(0.404)	(-1.459)	(-0.949)	(-0.802)
Sex of interviewee	0.206	0.218	-0.253	-0.007	-0.098	-0.159	0.145	0.231
	(0.955)	(0.848)	(-1.063)	(-0.035)	(-0.411)	(-0.428)	(0.764)	(1.375)
Interviewee's speech share	-0.025***	0.010	-0.001	-0.015	-0.012	-0.003	-0.027***	-0.014
	(-2.710)	(0.939)	(-0.120)	(-1.536)	(-1.293)	(-1.085)	(-2.915)	(-1.488)
Log(Wordcount)	0.019	-0.103	0.384***	0.460***	0.395***	0.438***	0.144	-0.108
	(0.214)	(-0.905)	(3.952)	(4.526)	(4.557)	(3.351)	(1.491)	(-1.066)
Age group = 1, 50-64	0.066	-0.096	0.020	-0.150	0.041	0.149	-0.053	-0.120
	(0.656)	(-0.847)	(0.211)	(-1.607)	(0.458)	(0.994)	(-0.497)	(-1.296)
Age group = 2, >64	-0.353*	0.002	0.560***	0.028	0.529***	0.397	-0.320	-1.022***
	(-1.748)	(0.009)	(2.782)	(0.124)	(2.991)	(-0.742)	(-1.445)	(-4.448)
Constant	0.645	-0.302	-2.234***	-1.843**	-1.874**	-2.705***	0.747	0.937
	(2.288)	(-0.303)	(-2.669)	(7.029)	(5.534)	(-2.318)	(1.789)	(8.161)
Observations	493	493	493	493	493	493	493	493
Adjusted R^2	0.309	0.054	0.249	0.363	0.348	0.221	0.230	0.316
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Interviewer FE	YES	YES	YES	YES	YES	YES	YES	YES

Table IV-C.2: Regression results with standardized dependent variables and with inclusion of interviews' topic proportions

This table presents a rerun of our main regression with standardized dependent variables extended by the 29 variables indicating the proportions of the 30 identified topic within an interview as derived from structural topic modelling (we exclude one variable to avoid perfect multicollinearity). We standardized our linguistic measures across the entire sample by subtracting the mean and dividing by the standard deviation. Our main explanatory variable is $Regulator_j$ that denotes an indicator variable equaling one if the interviewee is a regulator and zero otherwise. Control variables are the same as presented in Table IV-C.1. Please see Table IV-A.1 in Appendix IV-A for detailed description of all variables. Double-clustered standard errors for interviewee and year are reported in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	Selfishness		**Cognitive complexity**			**Dishonesty**	**Overconfidence**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	*SelfRef*	*Narc*	*CogScore*	*FRE*	*CogPro*	*DScore*	*SAB*	*Clout*
Regulator	-0.481*	-1.020***	0.490**	-0.736***	0.587**	0.073	-0.414**	-0.546**
	(-1.669)	(-3.238)	(2.019)	(-2.714)	(2.479)	(0.313)	(-2.143)	(-2.084)
Observations	493	493	493	493	493	493	493	493
Adjusted R^2	0.403	0.191	0.303	0.399	0.438	0.224	0.318	0.429
Further Controls	YES	YES	YES	YES	YES	YES	YES	YES
Topic Proportions	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Interviewer FE	YES	YES	YES	YES	YES	YES	YES	YES

Appendix IV-D: Structural topic modelling

Figure IV-D.1: Model diagnostics by number of topics

This figure illustrates the model diagnostics by number of topics K derived from structural topic modelling (Roberts et al., 2014). *Held-out likelihood* captures how surprised a trained model, i.e. a trained topic-deriving model with a subset of the total data, is of new data it has not seen before. The lower this value, the better the trained model captures patterns of natural language. *Lower bound* is an approximation of the lower bound on the marginal likelihood representing model's internal measure of fit. The lower this value, the better the trained model. *Residuals* illustrates the metric based on the residual-based diagnostic method of Taddy (2012). The lower this value, the closer the specified number of topics is to the optimal number of topics. *Semantic coherence* is a metric related to the degree of semantic similarity between high scoring words in a topic based on Mimno et al. (2011). The higher this value, the more likely words, which are most probable under a topic, co-occur within the same document.

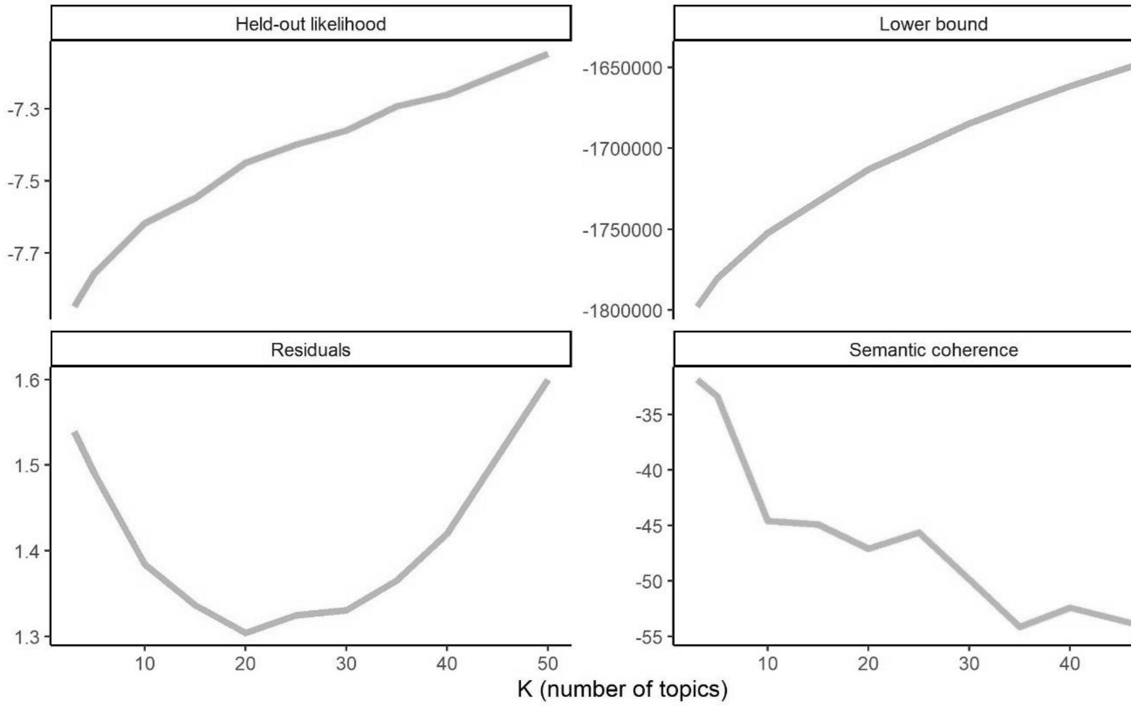


Figure IV-D.2: Comparison of exclusivity and semantic coherence by number of topics

This figure illustrates the comparison of exclusivity and semantic coherence by number of topics K. *Semantic coherence* is a metric related to the degree of semantic similarity between high scoring words in a topic based on Mimno et al. (2011). The higher this value, the more likely words, which are most probable under a topic, co-occur within the same document. *Exclusivity* measures the extent to which the top words for a topic do not appear as top words in other topics. The higher this value, the higher the exclusivity for each topic.

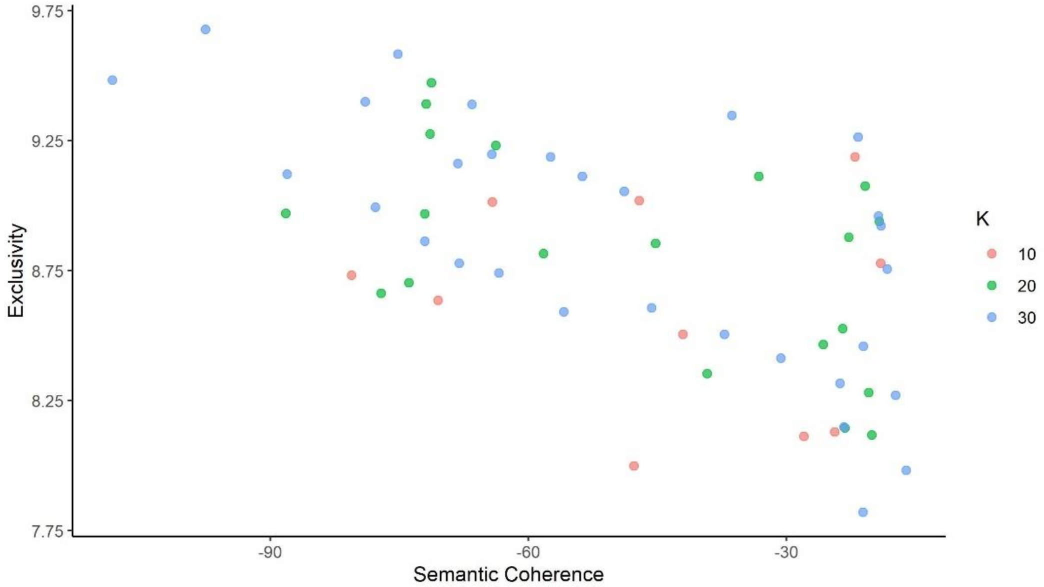


Figure IV-D.3: Word clouds for top 10 topics (by prevalence)

This figure presents a word cloud of the most probable words under a topic for each of the top ten topics (by prevalence).

Topic 3

Topic 29

Topic 14

Topic 5

Topic 15

Topic 6

Topic 20

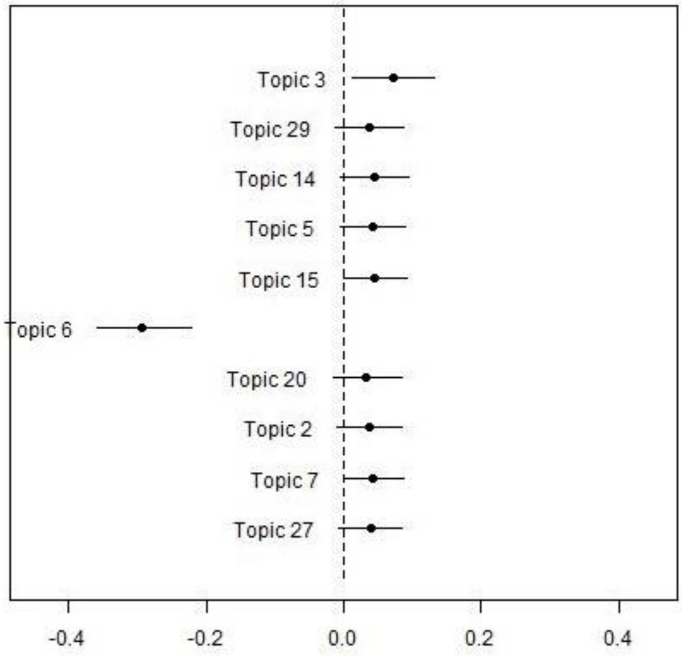
Topic 2

Topic 7

Topic 27

Figure IV-D.4: Distribution of top 10 topics (by prevalence) by interviewee group

This figure illustrates the distribution of the top ten topics (by prevalence) by interviewee group, i.e. bankers and regulators. On the horizontal axis, the differences between the two groups in the topic prevalence of each topic are plotted.



Regulators… Bankers

Affidavit

Ich erkläre hiermit, dass ich die vorgelegten und nachfolgend aufgelisteten Aufsätze selbstständig und nur mit den Hilfen angefertigt habe, die im jeweiligen Aufsatz angegeben oder zusätzlich in der nachfolgenden Liste aufgeführt sind. In der Zusammenarbeit mit den angeführten Koautoren war ich mindestens anteilig beteiligt. Bei den von mir durchgeführten und in den Aufsätzen erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis niedergelegt sind, eingehalten.

Daniel Czaja
Gießen, den 27.08.2020

Submitted Papers

I. Czaja, D. and Röder, F. (2020): Self-attribution bias and overconfidence among nonprofessional traders, *The Quarterly Review of Economics and Finance (forthcoming)*. https://doi.org/10.1016/j.qref.2020.02.003

II. Czaja, D. and Röder, F. (2020): Signaling in initial coin offerings - the key role of entrepreneurs' self-efficacy and media presence (*working paper*).

III. Czaja, D., Ritter, P., and Stolper, O. (2020): Among peers: the impact of homophily in online investment (*working paper*).

IV. Czaja, D., Tillmann, P., and Walter, A. (2020): Occupational self-selection among bankers and financial regulators: evidence from content analysis (*working paper*).